Enhancing Privacy and Fairness in Search Systems

A dissertation submitted towards the degree
Doctor of Engineering (Dr.-Ing.)
of the Faculty of Mathematics and Computer Science
of Saarland University

by
JOANNA BIEGA

Saarbrücken
December 2018

Defense Colloquium

Date: 1 April 2019

Dean of the Faculty: Prof. Dr. Sebastian Hack

Examination Committee

Chair: Prof. Dr. Bernt Schiele

Reviewer, Advisor: Prof. Dr. Gerhard Weikum

Reviewer, Co-Advisor: Prof. Dr. Krishna P. Gummadi

Reviewer: Prof. Dr. Carlos Castillo

Reviewer: Prof. Dr. Wolfgang Nejdl

Academic Assistant: Dr. Erisa Terolli

Acknowledgments

I am greatly indebted to my PhD advisors, Gerhard Weikum and Krishna Gummadi, for their mentorship. Thank you Gerhard for your excellent technical guidance, for pushing me to work to the best of my ability, and for allowing me so much freedom to pursue my research interests. Thank you Krishna for teaching me to think deeply about the problems I work on, for showing how to be playful in coming up with new ideas, and for the many hours of non-technical discussions that helped me to grow as a researcher.

Tremendous thanks to Prof. Carlos Castillo and Prof. Wolfgang Nejdl for reviewing this thesis and for all the helpful feedback. Thanks also to Prof. Bernt Schiele and Dr. Erisa Terolli for agreeing to devote their time to serve on my examination committee.

Special thanks to Fabian Suchanek for being a fantastic mentor, for introducing me to the research world and later convincing me to do a PhD, and for being an inspiring example of how to reconcile professional excellence with various hobbies.

Thanks to the Privacy Infrastructure Team at Google Zurich for hosting me during my internship and teaching me so many valuable lessons about engineering excellence.

I want to thank all the friends and colleagues from D5@MPI-INF, SocSys@MPI-SWS, and the institute administration, for creating such an amazing working environment.

Thanks to my old friends from the 'Sanok' group for being a constant in my life despite the fact that we are so geographically distributed now, and for reminding me that I am not my work. Thanks to Dominika as well as all the friends from our Polish geek group in Sb for all the good times and great discussions. Thanks to all the 'Balnica' friends for all the artistically creative times that powered me through the work times.

Heartfelt thanks to the family Roth for making Germany and Switzerland feel a bit more like home.

Thanks to my sister Ela for all the emotional support - during the good times and the bad times.

I want to thank my parents, Lilianna and Marek, for making our education a priority. Particular thanks go to my mom who, in a small Polish town in the 80s and 90s, had a visionary dream that her daughters become engineers and grow up to be independent women. It is only years later that I realize how privileged I was to have been raised this way as a girl.

Last but not least, thank you Ben for having supported me in countless ways through all these years, for all the amazing experiences we've had together, and for all the value you bring to my life.

Asia

Abstract

FOLLOWING a period of expedited progress in the capabilities of digital systems, the society begins to realize that systems designed to assist people in various tasks can also harm individuals and society. Mediating access to information and explicitly or implicitly ranking people in increasingly many applications, search systems have a substantial potential to contribute to such unwanted outcomes. Since they collect vast amounts of data about both searchers and search subjects, they have the potential to violate the privacy of both of these groups of users. Moreover, in applications where rankings influence people's economic livelihood outside of the platform, such as sharing economy or hiring support websites, search engines have an immense economic power over their users in that they control user exposure in ranked results.

This thesis develops new models and methods broadly covering different aspects of privacy and fairness in search systems for both searchers and search subjects. Specifically, it makes the following contributions:

- We propose a model for computing *individually fair rankings* where search subjects get exposure proportional to their relevance. The exposure is amortized over time using constrained optimization to overcome searcher attention biases while preserving ranking utility.
- We propose a model for computing *sensitive search exposure* where each subject gets to know the sensitive queries that lead to her profile in the top-k search results. The problem of finding exposing queries is technically modeled as reverse nearest neighbor search, followed by a weekly-supervised learning to rank model ordering the queries by privacy-sensitivity.
- We propose a model for quantifying *privacy risks from textual data* in online communities. The method builds on a topic model where each topic is annotated by a crowdsourced sensitivity score, and privacy risks are associated with a user's relevance to sensitive topics. We propose relevance measures capturing different dimensions of user interest in a topic and show how they correlate with human risk perceptions.
- We propose a model for *privacy-preserving personalized search* where search queries of different users are split and merged into synthetic profiles. The model mediates the privacy-utility trade-off by keeping semantically coherent fragments of search histories within individual profiles, while trying to minimize the similarity of any of the synthetic profiles to the original user profiles.

The models are evaluated using information retrieval techniques and user studies over a variety of datasets, ranging from query logs, through social media and community question answering postings, to item listings from sharing economy platforms.

Kurzfassung

Nach einer Zeit schneller Fortschritte in den Fähigkeiten digitaler Systeme beginnt die Gesellschaft zu erkennen, dass Systeme, die Menschen bei verschiedenen Aufgaben unterstützen sollen, den Einzelnen und die Gesellschaft auch schädigen können. Suchsysteme haben ein erhebliches Potenzial, um zu solchen unerwünschten Ergebnissen beizutragen, weil sie den Zugang zu Informationen vermitteln und explizit oder implizit Menschen in immer mehr Anwendungen in Ranglisten anordnen. Da sie riesige Datenmengen sowohl über Suchende als auch über Gesuchte sammeln, können sie die Privatsphäre dieser beiden Benutzergruppen verletzen. In Anwendungen, in denen Ranglisten einen Einfluss auf den finanziellen Lebensunterhalt der Menschen außerhalb der Plattform haben, z. B. auf Sharing-Economy-Plattformen oder Jobbörsen, haben Suchmaschinen eine immense wirtschaftliche Macht über ihre Nutzer, indem sie die Sichtbarkeit von Personen in Suchergebnissen kontrollieren.

In dieser Dissertation werden neue Modelle und Methoden entwickelt, die verschiedene Aspekte der Privatsphäre und der Fairness in Suchsystemen, sowohl für Suchende als auch für Gesuchte, abdecken. Insbesondere leistet die Arbeit folgende Beiträge:

- Wir schlagen ein Modell für die Berechnung von fairen Rankings vor, bei denen Suchsubjekte entsprechend ihrer Relevanz angezeigt werden. Die Sichtbarkeit wird im Laufe der Zeit durch ein Optimierungsmodell adjustiert, um die Verzerrungen der Sichtbarkeit für Sucher zu kompensieren, während die Nützlichkeit des Rankings beibehalten bleibt.

- Wir schlagen ein Modell für die Bestimmung kritischer Suchanfragen vor, in dem für jeden Nutzer Aanfragen, die zu seinem Nutzerprofil in den Top-k-Suchergebnissen führen, herausgefunden werden. Das Problem der Berechnung von exponierenden Suchanfragen wird als Reverse-Nearest-Neighbor-Suche modelliert. Solche kritischen Suchanfragen werden dann von einem Learning-to-Rank-Modell geordnet, um die sensitiven Suchanfragen herauszufinden.

- Wir schlagen ein Modell zur Quantifizierung von Risiken für die Privatsphäre aus Textdaten in Online-Communities vor. Die Methode baut auf einem Themenmodell auf, bei dem jedes Thema durch einen Crowdsourcing-Sensitivitätswert annotiert wird. Die Risiko-Scores sind mit der Relevanz eines Benutzers mit kritischen Themen verbunden. Wir schlagen Relevanzmaße vor, die unterschiedliche Dimensionen des Benutzerinteresses an einem Thema erfassen, und wir zeigen, wie diese Maße mit der Risikowahrnehmung von Menschen korrelieren.

- Wir schlagen ein Modell für personalisierte Suche vor, in dem die Privatsphäre geschützt wird. In dem Modell werden Suchanfragen von Nutzer partitioniert und in synthetische Profile eingefügt. Das Modell erreicht einen guten Kompromiss zwischen der Suchsystemnützlichkeit und der Privatsphäre, indem semantisch kohärente Fragmente der Suchhistorie innerhalb einzelner Profile beibehalten werden, wobei gleichzeitig angestrebt wird, die Ähnlichkeit der synthetischen Profile mit den ursprünglichen Nutzerprofilen zu minimieren.

Die Modelle werden mithilfe von Informationssuchtechniken und Nutzerstudien ausgewertet. Wir benutzen eine Vielzahl von Datensätzen, die von Abfrageprotokollen über soziale Medien Postings und die Fragen vom Q&A Forums bis hin zu Artikellistungen von Sharing-Economy-Plattformen reichen.

Contents

Chapter 1

Introduction

Contents

1.1 Motivation

Following a period of expedited progress in the capabilities of digital systems, the society begins to realize that systems designed to assist people in various tasks can also harm individuals and society. The harm may occur across a number of dimensions, ranging from privacy intrusion caused by massive collection of personal data, discrimination caused by algorithms trained on biased data, marginalization of certain groups in online communities, polarization of society caused by massive personalization, disinformation caused by viral spread of false information, all the way to addiction caused by systems aiming to aggressively monetize people's attention. These problems have gained the attention of an interdisciplinary community of researchers, including computer scientists, social scients, and legal scholars[1].

The information retrieval (IR) community has also recently recognized FATE (standing for Fairness, Accountability, Transparency, and Ethics) and the societal impact of IR technology as one of the crucial directions for the field [Culpepper et al. 2018]. In line with that direction, *user rights* in search systems motivate the work carried out in this thesis. In particular, we focus on the issues of privacy and fairness.

1.1.1 Search systems

Search systems mediate access to information. Figure 1.1 schematically describes a search environment. People may participate in this environment in two different roles – as *searchers*

[1] See, for instance, `fatconference.org`, `ainow.com`.

or *search subjects*. Searchers are the users who turn to search engines to find information. Typically, they phrase their information needs as keyword *queries*, which are issued to the search system. The *ranking mechanism* computes the relevance of each document in the underlying collection to the issued search query, and returns a ranked list of documents likely to be relevant to the searcher's information need.

If a search system observes a searcher over time and collects all the queries she issues, the relevance and the document ranking mechanism can be personalized. For instance, it is possible to determine that a user querying for *python* is interested in the programming language rather than the animal if she has issued other queries related to programming in the past.

While traditionally documents are thought of as text, without necessarily any person associated with them, many search systems nowadays implicitly or explicitly rank people. Those people might be job seekers on human resource support platforms, such as LinkedIn, or content creators on platforms like Spotify (music rankings), Amazon (product rankings), Airbnb (apartment rankings), or Twitter (social media posting rankings). We call these ranked users search subjects. Search subjects are *exposed* to searchers in ranking results, and the currency in which they are being paid on the platform is the searcher's *attention*. Attention can be measured in terms of click rates, gaze fixation times measured in eye-tracking studies, or more directly by the total amount of income earned by search subjects from successful transactions.

Because search subjects are exposed to searchers in response to queries, search queries determine the *context* in which exposure happens. Such a context can be positive or negative, yielding exposure desirable in some scenarios and undesirable in others. Exposure might be desirable, for instance, on hiring support platforms where job candidates want to be exposed to recruiters searching for new employees. Exposure might be undesirable, however, in social media search engines, when searchers issue sensitive queries related to diseases or controversial political issues.

Since search systems collect vast amounts of data about both searchers and search subjects, they have the potential to violate the privacy of both of these groups of users. Moreover, by controlling the exposure of different search subjects and the quality of results for different searchers, they have an immense power to deliver unfairly disparate levels of service and experience to different people. The increasing dependence on search systems in various platforms and areas of life thus calls for investigation of the issues of privacy and fairness in search.

1.1.2 Privacy in search

Privacy of searchers. The potential to violate the privacy of searchers is a result of search systems collecting search queries and aggregating them into detailed user profiles. Because search engines are often the first source people refer to when seeking information necessary for their work or hobbies, when seeking information related to health issues or personal problems, or when planning travels, search histories often paint a very intimate picture of a searcher's life. Having all of this information aggregated per user profile leads to a number of privacy risks, including linking of sensitive queries to real-world individuals,

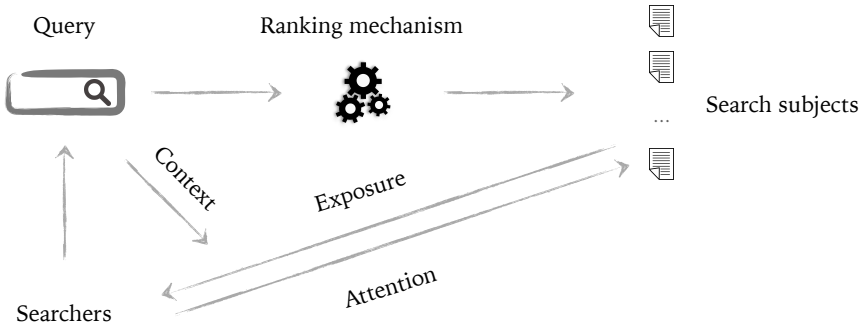

Figure 1.1: A schematic depiction of a search system. Users participate in the system either as searchers or search subjects.

inference of additional user attributes that were not disclosed to the search system directly, or profiling and targeting. We describe these risks in detail in the following paragraphs.

Linking of search histories to individuals is possible through queries containing pieces of individual-specific information. This harm was exemplified when AOL released a naively anonymized querylog with usernames replaced by random IDs in 2006[2]. Following the release, journalists were able to deanonymize some individuals by cross-referencing queries containing phone numbers (some users might query for their own phone numbers either deliberately or through copy-pasting mistakes) with phone book entries. Once deanonymized, search histories enable further linking of individuals to sensitive information, including queries related to topics like health or hygiene. Such attacks were demonstrated to be viable beyond just search data. For instance, it is possible to match anonymized Netflix movie recommendations with publicly available IMDB movie rating data, thus matching real-world identities to possibly sensitive ratings revealing political or sexual orientations [Narayanan and Shmatikov 2008].

A collection of large amounts of user data might also enable inference of information that is not present in the data and perhaps explicitly protected by the users. It is feasible, for instance, to infer demographic information using search logs [Bi et al. 2013]. Even beyond relatively rich and complex search histories, very private data, such as personality traits, can be inferred from items a user likes on a platform like Facebook[3] [Kosinski et al. 2013].

Detailed user profiles might also be made available to third-parties, often without the knowledge of the profile owner. Data can be passed on beyond the original intentions in case of company mergers, or if the search provider infrastructure gets compromised.

[2] https://en.wikipedia.org/wiki/AOL_search_data_leak
[3] https://facebook.com

Beyond unintentional leaks and breaches, user data might also be intentionally used against their interest. For instance, search providers use query histories for profiling and targeted advertising. Especially if the topic of an ad is sensitive, such advertisement being displayed in a user browser might lead to privacy breaches if seen by external observers.

Privacy of search subjects Privacy problems of search subjects are tied to exposure in the ranking results of sensitive queries. As depicted in Figure 1.1, the query of a searcher determines the context in which search subjects are exposed. Sensitive queries will as a result lead to sensitive exposure. Examples of such queries include those topically related to health, finance, or other personal issues, those which are unique to a user, such as phone numbers or e-mail addresses, or those which are rather uncommon for a given user profile.

Sensitive exposure enabled hackers to scrape the profile data of around 2 billion Facebook users in April 2018[4]. Scraping user data on such a massive scale faces a fundamental problem – how to enumerate all users in the system to get their profile URLs. The hackers had reportedly acquired a database of e-mail addresses and phone numbers on the Dark Web, and issued these as queries to Facebook's search engine. As a response to these queries, the engine returned links to the profiles of users to whom the e-mail or the phone number belonged. There exist other scenarios of this kind - with adversaries not targeting any particular individual but rather searching for targets matching their relevance criteria. Examples include bloggers looking for examples to their stories, or governments looking for politically controversial statements.

Adversarial situations like these could be potentially avoided if the search engines computed the exposure information. Computing search exposure means reversing search and, for each user profile, finding all the queries that yield the profile in the top-k results.

1.1.3 Fairness in search

Fairness for searchers. Delivering search results of disparate quality to different demographic groups might mean that historically disadvantaged groups receive worse access to information. Thus, fairness for searchers has been understood as a lack of disparity in the quality of results. In this spirit, Mehrotra et al. [2017] proposed a methodology to measure whether a search engine delivers less satisfactory results to searchers from different groups determined by attributes such as gender or age. It is worth noting that a search engine might underperform for searchers from minority groups not necessarily intentionally, but because it might have less observational data about these groups at its disposal to train the algorithms, or perhaps because the relevance feedback collected from the majority groups is biased against the minority.

Fairness for search subjects. Fairness for *search subjects* matters especially in scenarios where rankings influence people's lives outside of the platform. Such is the case for two-sided economy platforms, including Airbnb or Uber, or hiring support platforms such as LinkedIn. In each of these systems, subjects seek to be displayed high in the rankings as it increases

[4] https://www.independent.co.uk/news/world/americas/facebook-hackers-personal-data-collection-users-cambridge-analytica-trump-mark-zuckerberg-latest-a8289816.html

their chances of getting a real-world advantage, be it a higher income or being contacted by recruiters. With such a tangible influence over people's lives, search engines in these scenarios should make sure their results are fair, or more specifically, that subjects get a fair representation in the ranked results. Most papers thus far have proposed to quantify such fair representation using different forms of diversity and exposure. In practice, exposure can be determined in eye-tracking studies by measuring the time the searchers spend investigating a result, or by estimating click probabilities for different ranking results. To ensure fairness to individuals, a system should provide each subject the amount of exposure that is proportional to her relevance [Biega et al. 2018]. Unfairness, however, often falls along the lines of historical inequities. Ensuring fairness on a group level – for groups defined by legally protected attributes, such as gender or race – means granting equal exposure to different groups [Singh and Joachims 2018; Zehlike et al. 2017].

1.2 Challenges

In the context of the described privacy and fairness problems, this thesis tackles the following specific challenges:

- **Fair exposure for search subjects.** To be individually fair to each subject, a search system should grant subjects exposure that is proportional to their relevance. However, if many subjects have similar relevance in a given search task, it is impossible to grant everyone the attention they deserve in a single ranking. This problem arises because of a phenomenon called *position bias* where the searchers pay disproportionately more attention to subjects ranked higher, often irrespective of their relevance. As a result, it is impossible to be individually fair to subjects in a single ranking. We can instead look at sequences of rankings, and amortize exposure over time. This thesis tackles the challenge of *granting every subject in a ranking system the amortized exposure they deserve.*

- **Sensitive exposure of search subjects.** If a user's post is returned as a top-k answer to a sensitive search query, the user is exposed in a sensitive context. The richness and volume of the content we post online make it challenging to maintain awareness of the contexts in which our posts are returned as top-k results in search systems. Online users have very limited information about the queries that lead others to their profiles, yet - from the privacy perspective - such information is crucial if these exposing queries are of sensitive nature. This thesis tackles the problem of privacy-sensitive search exposure, that is, *finding the sensitive queries for which any post of a given user is returned as a top-k search result.*

- **Quantifying privacy risks from textual data.** Prior work on privacy has largely focused on structured data, such as databases or graphs. These solutions prove insufficient for users in online communities that allow for creation of textual contents. In particular, quantifying sensitive exposure requires a methodology for quantifying privacy-sensitivity in text. This thesis *tackles the problem of quantifying privacy risks from textual data.*

- **Privacy-preserving personalization for searchers.** Information accumulated within a single user account often draws an exact picture of a person's life. Such massive accumulation of personal data leads to significant privacy concerns. At the same time, while caring about privacy, many users feel compelled to give in all of this information in exchange for quality personalized results. This thesis tries to challenge the assumption that detailed user profiles are necessary to personalize the results and *tackles the problem of designing mechanisms for delivering personalized search results without the need for accurate user profiling.*

1.3 Thesis contributions

Equity of Attention in Rankings. This dissertation develops a mechanism for reordering rankings such that each subject in the system receives attention from the searchers that is proportional to their relevance. It is, however, impossible to achieve such a proportionality in a single ranking – searchers are susceptible to position bias, which makes them pay disproportionately more attention to subjects ranked at the top, irrespective of relevance. We thus propose to amortize attention over time by reordering consecutive rankings. While addressing fairness concerns, reordering subjects in the ranking will lead to accuracy loss when order is no longer determined by relevance. Trying to balance both of these dimensions, we formalize reordering as a constrained optimization problem, where we minimize unfairness (measured as a disparity between attention and relevance) subject to constraints on ranking accuracy loss. Choosing appropriate fairness and quality measures, the problem can be solved as an Integer Linear Program (ILP). We apply and analyze the behavior of the proposed mechanism on synthetic and real-world data of rental apartments from the Airbnb platform[5]. This work was published as a full paper at SIGIR 2018 [Biega et al. 2018].

Sensitive Search Exposure. This dissertation develops methodology for quantifying sensitive search exposure. We define search exposure as the problem of finding all the queries that expose any of a user's posts in the top-k results in a community's search engine. With this formulation, the problem can be seen as reverse search. Thus, if we think about search as the problem finding k-nearest-neighbors (i.e., k documents closest to the search query by a given similarity or relevance metric), one can cast search exposure as an instance of a well-defined problem of Reverse-k-Nearest-Neighbors. Generating such queries is not enough – our empirical analysis with user profiles from Twitter reveals that exposure sets for some users might be enormous and largely contain noisy and meaningless queries. To make the outputs useful to end users, we design a weakly-supervised learning-to-rank method, ordering the queries such that those at the top are most concerning. We show that the queries can be effectively ranked using only implicit signals which are readily available to service providers. This work was published as a full paper at CIKM 2017 [Biega et al. 2017a].

R-Susceptibility. This dissertation develops methodology for quantifying privacy risks from textual data in an online community. We propose to quantify the risks using a skeleton

[5] https://airbnb.com

of a topic model, which is a set of distributions over words. Each topic is annotated with a privacy sensitivity score determined in a crowdsourcing study. The model provides each user with the information on how relevant her postings are to each of the sensitive topics. Relevance is determined using a number of measures capturing how personal a user's interest in a given topic is. To this end, beyond pure lexical relevance, we model how broad a user's interest in a domain the topic comes from is (attempting to differentiate professional from personal interests), and how temporally spread a user's interest is (attempting to differentiate occasional and recurring interest). We moreover propose a notion of R-Susceptibility as a measure showing the user how high they rank in a given community with respect to a given sensitive topic. We evaluate the approach in a user study over profiles from three different online communities. This work was published as a full paper at SIGIR 2016 [Biega et al. 2016].

Privacy-preserving personalization. This dissertation proposes a framework of Mediator Accounts allowing for personalization of search results without the need to store exact user interaction histories. Mediator platform splits and merges queries of different users into synthetic user profiles, guided by a privacy-utility trade-off. Privacy is achieved by random assignments, and utility by keeping semantically coherent contexts intact (that is, topically similar queries of a user a kept together). The thesis moreover proposes a formalization of the notions of profiling privacy and individual user utility. Our experimental results using a querylog synthesized from the questions on the StackExchange platform[6] provided a detailed analysis of the trade-offs from the perspective of individual users, which should be contrasted to much of the previous works focusing on system utility. Our results showed that it is indeed possible to reconcile big profiling privacy gains with low personalization utility loss, particularly for users with rich profiles and diversified interests. This work was published as a full paper at SIGIR 2017 [Biega et al. 2017b].

In summary, the contributions of the thesis complement each other in a number of ways. First, we are investigating two different societal problems – fairness (Chapter 4) and privacy (Chapters 5, 6, 7). Second, we cover the problems of both search subjects (Chapters 4, 5, 6) and searchers (Chapter 7). Third, in the context of exposure specifically, we propose mechanisms for dealing with both wanted (Chapter 4) and unwanted exposure (Chapters 5, 6).

1.4 Other contributions of the author

The author of this thesis has co-authored a number of other papers and initiatives related to fairness and privacy, which are not included as contributions of this thesis.

- Our FATREC@RecSys 2017 paper [Chakraborty et al. 2017] focused on two-sided match-making platforms such as Uber. We argued that a single match cannot be fair as there are many relevant providers, yet only one provider receives the benefit of

[6] http://stackexchange.com

the match. To allow for a more uniform distribution of the benefits, we proposed to evaluate fairness over time by making sure the cumulative ratios of the deserved benefit and the actual benefit are the same for all the providers in a system. Interpreting the problem this way allowed us to draw a parallel to fair resource sharing algorithms, which have been well studied in the networking community.

- We have demonstrated the generalizability of the Mediator Accounts framework [Biega et al. 2017b] applying it to the problem of Privacy of Hidden Profiles, which we identified and defined in our CIKM 2017 paper [Eslami et al. 2017]. Hidden profiles are the profiles of users who decided to leave an online community, but whose data was retained by the service providers for analytic purposes. Such data still poses privacy risks for the users as it can easily be passed on beyond the original intentions, upon a governmental inquiry, a company merger, or when the infrastructure of the provider is compromised. Our results show that it is possible to protect the hidden profiles by scrambling user data, at the same time keeping the analytic utility of the data minimally affected.

- Our work on Mediator Accounts [Biega et al. 2017b] used a querylog synthesized from an online Community Question Answering community. The derivation methodology employed a simple heuristic for converting user questions to queries. To design better querylog derivation methods, we conducted a user study with the goal of understanding the query formulation process. We moreover proposed a methodology for deriving other characteristics of information retrieval collections, such as relevance judgments, from the structure of the CQA forums. This work is under submission.

- To facilitate further work on fairness in rankings, the author of this thesis has co-authored a successful proposal for a TREC track focusing on the problem of fairness[7]. TREC is an information retrieval conference whose goal is to design bechmarks (document collections with relevance judgments), as well as metrics and standardized experimentation protocols for the most important information retrieval tasks. This track will run for the first time in TREC 2019.

- Our PSBD@CIKM 2014 paper proposes a method to probabilisticaly predict whether users are personally afflicted by privacy sensitive states (such as depression or pregnancy), feeding the lexical information from their search histories into a probabilistic graphical model. Preliminary experimental analysis in this paper showed that the method can achieve a good accuracy in predicting privacy sensitive states. This work laid a foundation for the R-Susceptibility project [Biega et al. 2016].

[7] https://fair-trec.github.io/

1.5 Prior publications

The results of this thesis have been published in the following articles:

1. Asia J. Biega, Krishna P. Gummadi, and Gerhard Weikum. **Equity of attention: Amortizing individual fairness in rankings**. In *Proceedings of the 41st International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR 2018, Ann Arbor, MI, USA, July 08-12, 2018, pages 405–414.

2. Asia J. Biega, Azin Ghazimatin, Hakan Ferhatosmanoglu, Krishna P. Gummadi, and Gerhard Weikum. **Learning to un-rank: Quantifying search exposure for users in online communities**. In *Proceedings of the 2017 ACM Conference on Information and Knowledge Management*, CIKM 2017, Singapore, November 06-10, 2017, pages 267–276.

   The efficient algorithm for generating exposure sets and the corresponding experiments (Sections 3, 5.2 and 5.3 in this publication) are not contributions of this thesis.

3. Joanna Asia Biega, Krishna P. Gummadi, Ida Mele, Dragan Milchevski, Christos Tryfonopoulos, and Gerhard Weikum. **R-Susceptibility: An IR-centric approach to assessing privacy risks for users in online communities**. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, SIGIR 2016, Pisa, Italy, July 17-21, 2016, pages 365–374.

4. 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*, Shinjuku, Tokyo, Japan, August 7-11, 2017, pages 675–684.

   The experiments on applying the Mediator Account framework to the scenario of recommender systems (Sections 5 and 7 in this publication) are excluded from this dissertation.

1.6 Organization

The remainder of the thesis is organized as follows. Chapters 2 and 3 provide a background on user privacy and algorithmic fairness. Chapter 4 describes our contributions related to fairness in rankings. Chapter 5 describes our contributions related to sensitive search exposure. Chapter 6 describes our contributions related to quantifying privacy risks from textual data. Chapter 7 describes our contributions related to privacy-preserving personalization for searchers. Finally, Appendices B and A provide additional details on the users studies conducted in this thesis, which were omitted in the corresponding publications due to space constraints.

Chapter 2

Background: User Privacy

Contents

2.1 Preliminaries

This background chapter provides an overview of privacy notions and methods for achieving them in various systems. We are concerned with the *privacy of an individual user* whose data is collected by the system. The data may be created or generated by the user (as is the case for social network postings or web browsing histories), or by others (as is the case for medical databases with patient information collected by hospitals).

Types of user data. User data may consist of *structured attributes* as well as *unstructured text*. Structured attributes may be *binary* (for instance, whether a user liked a certain item), *categorical* (for instance, marital status), or *numerical* (for instance, user age). Textual data may occur in the form of *search queries*, or social media *postings*. Certain attributes may also encode connections between multiple users - for instance in the form of *friendship edges* in social networks.

Privacy-sensitivity of data. Certain items in user data might be considered *personally identifiable information* (PII). While determining what constitutes PII is generally context specific, one might assume attributes such as a username or a social security number are personally identifying. These attributes are often referred to as *identifiers*. Certain attributes

constitute *quasi-identifiers*. These attributes are not PII on their own, but may uniquely identify a user when combined with other quasi-identifiers. For instance, a given combination of a surname, a birth date, and a zip code, might describe a unique individual.

We call certain attributes *sensitive*. This is the information a user would like to protect in a given context – for instance, it might be undesirable to link diseases to patients, in which case the attribute *disease* would be considered sensitive. Sensitive attributes do not necessarily need to be explicitly present in the data - *inferring* the value of a sensitive attribute using the available data would also be considered a privacy breach.

Data usage scenarios. User data might be sanitized by a service provider if they want to make a public release (*data publishing*) or if they provide an *interface for querying* the data by analysts. Users might want to *protect their data from the service providers*, in which case sanitization is performed either by the user or a trusted third-party before the data reaches the service provider.

2.2 Privacy risks and notions

2.2.1 Information leakage

Linkability. Linkability (or identity disclosure) is a risk of matching data to a real-world individual who is the owner of the data or whom the data describes. While the most straightforward way to prevent this breach is to remove any personally identifying information, such protection is often not enough[1] because of the possibility of matching quasi-identifiers with external data sources. The protected data might contain sensitive information with quasi-identifiers but without personally identifiable information. External public data sources might contain personally identifiable information together with quasi-identifiers, making it possible to mach anonymized records by quasi-identifier values. Such a strategy was used to deanonymize medical records of the Governor of Massachusetts using public employee records [Sweeney 2002b]. Moreover, research shows it is possible to de-anonymize individuals from social network graph data by correlating anonymized graphs with publicly available graphs [Narayanan and Shmatikov 2009], from browser fingerprints constructed from browser settings and plugins, which are available to any website a user visits [Eckersley 2010], or from movie rating data by correlating anonymized ratings with ratings publicly available on the IMDB[2] film website[Narayanan and Shmatikov 2008].

Attribute disclosure. Even if matching of an individual to a specific data record is not possible, it might still be possible to learn the value of a sensitive attribute of a given individual [Machanavajjhala et al. 2007]. Such a breach is called *attribute disclosure*. For instance, imagine we know an individual with the quasi-identifier values of (66123, male, 07.09.1982) who has been hospitalized at an institution releasing patient records. Even when we cannot uniquely link the person to a single row in the dataset, attribute disclosure

[1] https://en.wikipedia.org/wiki/AOL_search_data_leak
[2] htpps://imdb.com

will occur if each row the individual possibly maps to has the same value of the sensitive attribute.

Attribute disclosure can also be understood as gaining additional information about the value of a sensitive attribute from a dataset when compared to the prior knowledge [Li et al. 2007]. For instance, if an individual has a heart disease with the probability 0.6 according to the dataset, while the prior probability in a given population for having a heart disease is 0.1, the adversary increases her certainty about the sensitive value even though she does not learn the exact value.

Attribute inference. Privacy may also be violated when attributes and information not directly observable or recorded in the data are inferred from the observable data of multiple users using machine learning methods. While the two concepts are related, the main conceptual difference between attribute inference and disclosure is that disclosure typically refers to recovering the value of a sensitive attribute that is present in an anonymized dataset, while inference refers to learning the attributes not present in the data based on observations of attribute patterns in user populations.

Various types of attribute inference have been demonstrated in the literature. For instance, it is possible to predict psychological and demographic traits from the items people *like* in a social network such as Facebook [Kosinski et al. 2013], predict whether two accounts in two different online communities belong to the same individual [Goga et al. 2015], predict a user's location from the tags they add to their online postings [Zhang et al. 2018], or predict sensitive information (such as whether a person is on a vacation, driving drunk, or having a certain disease) from the textual contents of their social media postings [Mao et al. 2011].

Classifiers based on stylometric techniques for text, such as the analysis of usage of different words or syntactic and linguistic patterns, have been shown to enable authorship inference [Abbasi and Chen 2008; Narayanan et al. 2012].

2.2.2 Profiling

Loss of privacy can also occur when a large amount of data is collected about a user, as the entirety of such data (for instance, a search history spanning multiple years) may paint a very intimate picture of a person's life. The risk might be defined, for instance, as the total amount of data collected [Singla et al. 2014; Biega et al. 2017b], or the total number of topics present in a user profile [Biega et al. 2017b; Meng et al. 2016b]. Such detailed data is usually collected to enable personalization [Biega et al. 2017b; Singla et al. 2014], or targeted advertising [Yu et al. 2016; Meng et al. 2016b].

2.2.3 Exposure

Privacy breaches include inappropriate exposure of user data. Inappropriate exposure might mean that data is accessible by unintended audience [Sandhu et al. 1996], either shortly after data creation or long thereafter, when a user does not necessarily remember about the data's existence and exposure possibility [Mondal et al. 2016]. Moreover, access to data might be made easier through various platform features. Examples include Facebook's News

Feed where updates to the content of user profiles are actively broadcast to other users, or search engines, where access to user profiles by is enabled by matching of keyword queries (see Chapter 5).

2.3 Achieving privacy

2.3.1 Limiting information leakage

K-anonymity. The success ratio of linkability based on quasi-identifiers depends on how many individuals share unique combinations of values. This insight underlies the anonymization technique called *k-anonymity* [Sweeney 2002b]. A dataset is k-anonymous if each combination of quasi-identifier values appears at in the dataset at least k times. The protection offered by this mechanism is that the probability of correctly mapping an individual to a row in the database is lower than $\frac{1}{k}$.

Means of preventing linkability include removal of any personally identifiable information, or perturbation of data so as to satisfy the k-anonymity requirement [Sweeney 2002a; Jr. and Agrawal 2005; LeFevre et al. 2005, 2006]. Such perturbations include generalization, where values are replaced by sets of values (for instance, zip code 66123 might become $66 * **$), and suppression, where individual user records are fully removed from the data.

Alternative approaches include generation of synthetic datasets that satisfy privacy criteria at the same time preserving patterns from the original data. This idea has been explored in the context of preserving frequent itemsets while ensuring k-anonymity [Vreeken et al. 2007], or preserving the distribution of attribute values, while adding controlled noise to the values generated from these distributions [Howe et al. 2017].

L-diversity. While mitigating the threat of linkability, k-anonymity does not offer full protection against sensitive *attribute disclosure*. To prevent such attribute disclosure, Machanavajjhala et al. [2007] proposed the notion of l-diversity [Machanavajjhala et al. 2007], which requires that within each equivalence class of rows defined by a given combination of quasi-identifier values (that is, within each k-anonymous block), there exist at least l different values of the sensitive attribute. As a result, even if an adversary is able to map an individual to a given k-anonymous block unambiguously, she still faces uncertainty about the individual's sensitive data.

T-closeness. L-diversity does not protect against attribute disclosure in a probabilistic sense. Aggregate statistics over the whole dataset provide the prior probability over the values of sensitive attributes in the population. If the distribution within the equivalence class an individual is mapped to is different from the prior distribution, the certainty about the value of the sensitive attribute of the individual changes. To prevent this kind of disclosure, the notion of t-closeness [Li et al. 2007] requires that the distributions of sensitive values within anonymous blocks are within a distance of at most t from the global distribution in the whole dataset. Distribution distance can be captured by metrics like the KL-divergence or the Earth Mover's distance.

Differential privacy. The notion of differential privacy [Dwork 2008] requires that the presence or absence of an individual's data in a dataset does not significantly change the output of a mechanism applied to the data. More precisely, a mechanism A satisfies ε-differential privacy if the following inequality holds for any two neighboring datasets D_1 and D_2 that differ by at most one row: $P(A(D_1) \in S) \leq e^{\varepsilon} \cdot P(A(D_2) \in S)$, for any set S that the mechanism A may compute. Note that this requirement is imposed over a mechanism applied to the data, and not over the data itself, and is studied mostly in scenarios where an analyst uses a data querying interface.

For example, assume an analyst wants to learn from census data D how many people in a given town have an income over \$100k. Let us denote this query function as f, and the mechanism $A(D) = f(D)$. The presence of an individual u with an income over \$100k in the dataset will change this value by 1. If the answer of the system for the dataset without the individual is n, then $P(A(D) \in \{n\}) = 1$ and $P(A(D \cup x) \in \{n\}) = 0$, thus violating the differential privacy requirement for any ε.

To satisfy the requirement of differential privacy, a system needs to add random noise to a mechanism results. For count queries, is has been shown that the privacy requirement can be satisfied by adding noise from the Laplace distribution with the scale parameter $\frac{\Delta f}{\varepsilon}$, where $\Delta f = max_{D_1,D_2}|f(D_1) - f(D_2)|$, also called *sensitivity*, determines the maximum difference in the value of f when applied to neighboring datasets D_1, D_2 [Dwork 2008]. In the aforementioned example, instead of returning $A(D) = f(D)$, the system returns $A(D) = f(D) + Y$, where $Y \sim Laplace(\frac{\Delta f}{\varepsilon})$. The scale of the noise that needs to be added depends on the sensitivity of the function f (the higher the sensitivity, the bigger the scale of the noise), as well as the privacy parameter ε (the lower the ε and thus the stricter the privacy requirement, the bigger the scale of noise).

Because differential privacy prevents inference of an individual's presence in the data, it offers protection from both linkability and attribute disclosure.

Decreasing inference accuracy. In the context of sensitive attribute inference using machine learning models, privacy loss is usually quantified as the accuracy of the applied model – the more accurately a sensitive attribute can be predicted, the more user privacy is at stake. Protections against inference attacks focus on perturbing the data to decrease the accuracy of the predictions. For instance, Zheleva and Getoor [2009] studied how friendship and group membership information influences the accuracy of sensitive attribute prediction in social networks. Zhang et al. [2018] propose a method to select tags appended to a social posting to prevent the inference of the posting user's location. Decreasing inference accuracy is also used to prevent authorship attribution. Solutions along these lines include stylistic suggestions for the authors to decrease the uniqueness of their style [Kacmarcik and Gamon 2006; McDonald et al. 2012], or crowdsourcing text reformulations [Almishari et al. 2014].

2.3.2 Limiting profiling

Limiting data collection. To limit the total amount of data collected from users, Singla et al. [2014] propose the notion of *stochastic privacy*. Stochastic privacy limits the probability

that different pieces of user data will be stored to enable further personalized service, thus effectively limiting the size of user profiles.

Obfuscating data. Profiling privacy might be protected by hiding real user interactions within fake data. Such an approach is taken, for instance, in the TackMeNot browser extension, which issues synthetic search queries on a user's behalf [Howe and Nissenbaum 2009], Similarly, a number of approaches have been proposed to obfuscate the topical interest of search users by issuing fake queries on topics not covered in a user's real search history [Pang et al. 2012; Wang and Ravishankar 2014]. Masood et al. [2018] propose a profile obfuscation method, where sensitive user data is replaced by semantically similar non-sensitive data.

Data grouping and splitting. An early idea of grouping user interactions was implemented in the Crowds system, in which requests were passed on in a random walk over a network of users before being passed on to a server [Reiter and Rubin 1998]. A procedure like that results in requests of individuals being split across multiple identities. This thesis pursues a related idea by proposing a Mediator Accounts framework (Chapter 7), where the goal is to to split and merge search queries of different users such that the personalization quality is minimally reduced. At the same time, we want to create synthetic profiles, which do not resemble original user profiles. Instead of merging user profiles or interactions, it is also possible to split them into smaller chunks. Such ideas have been investigated in the context of personalized search, where search histories were divided into chunks with topically related queries [Chen et al. 2011; Xu et al. 2007]. As a result, for instance, instead of seeing an individual interested in programming, cooking, and sports, a search engine would see three individuals, each interested in one of the three aforementioned topics.

Anti-tracking. Beyond service providers directly collecting data about their users, there are third parties who *track* users when they browse the web. Such trackers collect information about the websites users visit, thus learning about their topical interests over time. Such data is then used to deliver targeted advertising to the users, a practice referred to as Online Behavioral Advertising. Several protection mechanisms have been proposed to protect user privacy in this context, either by requiring the website publishers to mediate between users and trackers by adding noise to user data [Akkus et al. 2012], by designing targeting architectures where user data is stored on a local device [Toubiana et al. 2010], by preventing collection of unique user attributes by the trackers [Yu et al. 2016], or by allowing users to select the information they share with the trackers, acknowledging that users might want to receive personalized ads for certain topics [Meng et al. 2016b].

2.3.3 Limiting exposure

Limiting data audience. Early approaches to privacy revolved around limiting access to data to pre-specified audience. Such access control lists (ACLs) were defined either by specifying individuals or groups of individuals (role-based ACLs) [Sandhu et al. 1996]. Such approaches might be too cumbersome and give little flexibility to users with hundreds of

connections in online social networks. To overcome this limitation, a number of solutions have been proposed to help users statically predefine ACLs based on the network structure using community detection approaches [Mazzia et al. 2012], or by suggesting friends to share the content with on the fly when adding new posts using machine learning approaches [Fang and LeFevre 2010].

Limiting data lifetime. Even when exposure is desirable at the time uploading the content, it might become undesirable when the content is re-discovered some time after publication. During this time, a user might forget such content is present in her digital traces, while her preferences and views change. An evolution like this is especially prominent when a young adult enters the job market with a long history of teenage social media postings. Motivated by such scenarios among others, Mondal et al. [2016] proposes an inactivity-based content withdrawal mechanism where postings are automatically withdrawn after the initial audience interest fades.

Exposure awareness. Apart from strict exposure control, it is important to help users be aware of the exposure of their content, especially when it happens through complex and often non-transparent mechanisms. A number of studies have motivated the need for exposure support showing that users consistently underestimate the size of their content's audience [Bernstein et al. 2013], or that people have strong feelings regarding exposure (thousands of Facebook users protested when the platform introduced the News Feed thinking their privacy was breached when the updates to the content of their profiles were actively broadcast to other users, even though the content was accessible to those same users upon a visit to individual profiles) [Boyd 2008]. Service providers do realize that such support is a crucial privacy awareness feature. For instance, Facebook allows its users to preview how their profile looks like to other people via a functionality called *View As*. Several interface designs were proposed to make users aware of the size of their content's audience, including showing a pair of eyes whose size is proportional to the size of the audience [Schlegel et al. 2011].

Arguably, in the context of search, the information on which *keyword queries* return a user's posts in a platform's search engine results is equally important as the information on who can see user personal content. The work presented in Chapter 5 contributes to the line of work on exposure awareness in search.

2.4 Cost of privacy

Achieving privacy comes at the cost of *utility loss*. For instance, the goal of releasing structured data is for people to be able to compute certain statistics and gain insights from the data. When trying to achieve the requirements of k-anonymity, l-diversity, t-closeness, or differential privacy, these computations become inaccurate when the attribute values are generalized, when rows get suppressed, or when noise is added to the results of queries. Generally, the higher the privacy, the lower the resulting utility of the data. Because there can be a lot of different anonymizations satisfying a chosen privacy criterion, one usually chooses the one which offers the highest utility. For models such as k-anonymity, l-diversity,

or t-closeness, utility is measured using proxy measures such as the number of resulting abstraction classes, the average number of rows in abstraction classes [LeFevre et al. 2005; Machanavajjhala et al. 2007; Li et al. 2007], the accuracy of classifiers predicting a sensitive attribute [Brickell and Shmatikov 2008], or the difference between the distributions of sensitive attributes in the original and the sanitized datasets [Li and Li 2009].

Techniques for achieving differential privacy control the amount of noise added to the data through the parameter ε. The lower the value of ε, the bigger the amount of noise one needs to add to achieve ε-differential privacy. Thus, the higher the privacy, the lower the utility of the data (that is, the accuracy of the mechanism results).

Utility loss caused by obfuscation of user profiles is often measured using various personalization quality measures [Singla et al. 2014; Wang and Ravishankar 2014; Chen et al. 2011; Zhu et al. 2010; Biega et al. 2017b].

2.5 Privacy in search systems

In the context of the main theme of this thesis, it is worth reiterating over the privacy problems specific to search systems.

Obfuscating searcher profiles. One of the goals in privacy-preserving IR is that of privacy-preserving search personalization. To this end, various approaches for obfuscating user profiles with constraints on the personalization utility were proposed. Different types of such obfuscation approaches include: removing parts of the logs [Singla et al. 2014], generating fake search queries and mixing them with the user-issued search queries [Howe and Nissenbaum 2009; Pang et al. 2012; Wang and Ravishankar 2014], splitting logs into multiple logs [Chen et al. 2011; Xu et al. 2007], or grouping the logs of different users [Biega et al. 2017b; Zhu et al. 2010].

Anonymizing search logs. If a service provider wants to release a query log, the log needs to be sanitized to ensure anonymity to the searchers whose data is being made public. A crucial step is the removal of any personally identifiable information. Assuming any query which appears in the log infrequently can be personally identifying, [Adar 2007] proposed a scheme where queries in the logs are masked until they appear in a certain number of user profiles.

Stronger anonymization techniques apply differential privacy mechanisms over raw query logs and publish the results returned by these mechanisms instead of the original log. For example, Zhang and Yang [2017] proposes publishing of query sessions (small subsets of user profiles encompassing sequences of queries issued within a short time frame to satisfy a single information need) with differentially private session counts, while Zhang et al. [2016b] propose a differentially private mechanism that results in publishing query counts without user profiles. Götz et al. [2012] provides a summary of guarantees offered by query log anonymization mechanisms whose goal is to publish frequent query log elements.

Adversarial inference using search logs. Another line of work investigates various adversarial attacks using search logs. For instance, Peddinti and Saxena [2010] and Gervais et al. [2014] have demonstrated machine learning approaches that enable distinguishing real and fake queries in obfuscated user search logs, while Jones et al. [2007] showed that it is possible to infer demographic and location information from query histories.

Private information retrieval. A related theme in security research is that of private information retrieval, where a user wants to retrieve an item from a database stored on a server, without revealing to the server which item is being requested [Chor et al. 1995]. The goal in this field is to design protocols better than the baseline approach where the server sends a full copy of the database to the requester.

Sensitive search exposure. This thesis tackles a novel problem of sensitive search exposure for search subjects. In Chapter 5 we propose a methodology for finding all the sensitive queries that expose a subject in the top-k results of a given search system. In Chapter 6 we show how sensitive exposure can serve as a tool for quantifying privacy risks in textual data.

2.6 Selected other dimensions in privacy research

Sensitivity analysis. For textual data, automated privacy mechanisms should be able to determine which content is sensitive. To this end, several approaches proposed the idea that content is sensitive if people tend to create it anonymously. In this spirit, Correa et al. [2015] showed that linguistic factors such as usage of the first person singular pronoun or semantic factors such as the topics of money, work, emotions, and sexuality can be used to train machine learning methods to distinguish anonymous and non-anonymous social media posts. Peddinti et al. [2014] studied the differences between anonymous and non-anonymous posts in the online question and answer community Quora, finding that beyond conventional sensitive topics such as sex, health or religion, people often choose to post anonymously on topics such as education and educational institutions, or the patent law.

Longitudinal privacy. Beyond the privacy threats stemming from occasional data release, it is important to understand long-term effects of information disclosure. Mondal et al. [2016] studied the content deletion behavior of Twitter users and proposed automatic methods to control the longitudinal privacy by automatically hiding inactive content. Rizoiu et al. [2016] studied the longitudinal privacy of Wikipedia editors measured as the prediction accuracy of features such as gender, education or religion. The paper demonstrates that the privacy of editors who become inactive and do not contribute any new data still decreases over time, as the accuracy of predictors increases thanks to data contributed by other users. Eslami et al. [2017] proposed a mechanism for perturbing information that stays in the system after a user decides to close her account, so as to minimize the effect of privacy breaches over data she has no control over.

Human perceptions of privacy. Research in privacy is often guided by user perceptions and privacy needs. To that end, researchers in the field of usable privacy perform user studies and interviews to understand these preferences as well as misconceptions people have about the way systems deal with user data. In the context of online identity management, for instance, Leavitt [2015] has found that people who feel non-anonymous using their primary identity on Reddit are more likely to create temporary throwaway accounts to post content. In the context of targeted advertising, Ur et al. [2012] have found that users consider such advertising techniques both useful and creepy, incorrectly believing that personally identifiable information is collected as a part of the process. Following up on this study, Agarwal et al. [2013] found that user concerns are context-dependent, and that users are generally only concerned about topically embarrassing ads. Moreover, users do not generally want to opt-out of targeted advertising, considering some of the personalized ads as useful. These results highlight that users do not want to completely lose the utility of online services to preserve their privacy.

Economics of privacy. One of the interesting lines of thought in privacy is that of economics of privacy. For instance, data can be seen as a product users need to be remunerated for. Li et al. [2014] have proposed a pricing mechanism where analysts pay for making queries over a dataset, and the payments are distributed to the dataset users in proportion to the contribution of their data to the query answer. Behavioral economics techniques have been applied to study how much users value privacy or how they make data sharing decisions [Acquisti 2009]. Acquisti et al. [2016] provide a broad literature survey in this area.

Chapter 3

Background: Algorithmic Fairness

Contents

Anti-discrimination laws have been introduced to account for the fact that certain groups of people have been historically subordinated. Regulations specify attributes, such as gender, race, or religion, along which it is illegal to discriminate in domains including education, employment, credit, or housing. The society begins to realize that digital systems can unfairly discriminate as well. The realization that seemingly objective computer systems can be biased has led to a number of questions the research community has sought to answer. How to define fairness mathematically? How to design mechanisms that are fair? How to audit black-box systems to hold them accountable? What constitutes undesired bias and how can it be measured?

3.1 Preliminaries

We investigate algorithmic fairness for *individuals* or *groups* of individuals defined by one of the aforementioned *protected attributes*. Issues of fairness matter in tasks such as classification or search, and applications that involve people as subjects or system users.

Classification and regression. Classification and regression methods are increasingly used in finance to predict customer creditworthiness, in hiring to predict whether a candidate is going to make a good employee, or in justice systems to predict whether a convict will

reoffend on a parole. Because in each of such scenarios algorithmic decisions influence the lives and opportunities of the subjects, fairness of these decisions has become a concern.

To illustrate most important fairness notions, we focus on binary classification. In this setup, one of the classification outcomes is usually considered positive. Assume a classifier is to make a hiring decision - whether to hire an individual or not. Each individual X is represented using a feature vector $\vec{x}$ describing various characteristics the designer of the classifier thought were important for the task, for instance, age, gender, highest degree received, GPA score, etc. Values can be categorical or numerical. The goal is to train a classifier to make binary hiring decisions $D \in \{0, 1\}$ (hire or no hire). The classifier is trained using data in the form of pairs $\vec{x}, y$ of feature vectors of past hires annotated with binary ground-truth decisions y specifying whether individual X had good performance reviews from her manager two years after being hired. The performance score is a proxy definition for being a good employee and a ground for a positive hiring decision. A classifier learns patterns from a collection of vectors $\vec{x}$ distinguishing people with positive and negative ground-truth values. We further choose gender as the protected attribute and denote x_G as the value of the gender feature of individual X.

The setup for a regression task is similar to that of classification, the difference being that we predict a real-valued target attribute. In the hiring context, we might want to predict, for example, how many years a person is likely to stay at the company.

Search and recommendation. In many domains, search engines rank people (explicitly or implicitly by ranking the content and products people produce). Since ranking positions in scenarios like this influence people's real-world economic livelihood, issues of fair representation in ranking have become a major concern.

In the hiring context, for example, employers might screen for potential employees using search engines on hiring support platforms. As a response to a keyword query q issued by an employer, such as *machine learning engineer*, the platform returns a ranked list of candidates $e_1, ..., e_k$ ordered by a relevance score $r(q, e_i)$ computed by a ranking algorithm. The ranking method can be based on data statistics or employ machine learning techniques. Machine learning approaches use training data with labels provided by expert annotators (an annotator determines which user profiles are relevant to which queries), or implicitly inferred from user click patterns. For example, a user is likely to click on profiles she deems relevant to her query. Chuklin et al. [2015] provide a detailed overview of various click models used to infer relevance.

A hiring platform might also proactively recommend potential employees to recruiters. In such a context, a recommendation algorithm can be thought of as a search system where the employer is a query. Recommendation strategies can recommend candidates to employers based on candidate's similarity to previous candidates the employer interacted with (item-item recommendation), because the candidate interacted with other employers similar to the given employer (user-user recommendation), or using a mix of both approaches (collaborative filtering).

3.2 Algorithmic fairness notions

3.2.1 Group fairness.

Notions of group fairness broadly aim at making sure that algorithms do not disproportionately adversely impact members of the protected groups.

Demographic parity. The notion of *demographic parity* requires that $P(D = 1|x_G = 1) = P(D = 1|x_G = 0)$. For example, in the context of job application classification with the protected attribute gender, this requirement means that groups of people of different genders have an equal chance of being accepted. Beyond demographic or statistical parity, this notion has appeared in the literature under a variety of different names, including *avoiding disparate impact* [Feldman et al. 2015], *independence* [Barocas et al. 2018], and *anti-classification* [Corbett-Davies and Goel 2018].

For tasks other than classification, demographic parity is often understood as equal representation in the results. For instance, clustering algorithms should make sure that different groups are similarly represented in all the clusters [Chierichetti et al. 2017], recommendation algorithms should make sure different groups are similarly represented in recommendation sets [Mehrotra et al. 2018] or that group proportions in recommendation sets should be similar to group proportions in input ratings [Ekstrand et al. 2018b], while ranking algorithms should make sure groups are similarly represented in ranking prefixes [Yang and Stoyanovich 2017; Celis et al. 2018; Zehlike et al. 2017; Singh and Joachims 2018].

Performance parity. Another category of group fairness definitions revolves around the idea of equal error rates, thus requiring equal performance of an algorithm for different groups of people. Since error can be captured using a variety of different metrics, various papers have focused on satisfying different metric equalities. Notable examples include equality of true positive rates also known as *equality of opportunity* [Hardt et al. 2016]: $P(D = 1|y = 1, x_G = 1) = P(D = 1|y = 1, x_G = 0)$; equality of both true positive and false positive rates also known as *equalized odds* [Hardt et al. 2016]: $P(D = 1|y = 1, x_G = 1) = P(D = 1|y = 1, x_G = 0)$ and $P(D = 1|y = 0, x_G = 1) = P(D = 1|y = 0, x_G = 0)$; equality of missclassification rates – including equality of false negative rates – also know as *lack of disparate mistreatment* [Zafar et al. 2017]: $P(D = 0|y = 1, x_G = 1) = P(D = 0|y = 1, x_G = 0)$; or equality of positive predictive values also known as *calibration*: $P(y = 1|D = 1, x_G = 1) = P(y = 1|D = 1, x_G = 0)$.

Fairness has also been studied as error parity for different groups in recommendations [Ekstrand et al. 2018a; Yao and Huang 2017], and search (using measures of satisfaction for searchers) [Mehrotra et al. 2017].

It has been shown that satisfying different mathematical notions of fairness simultaneously is generally not feasible [Kleinberg et al. 2017b; Chouldechova 2017]. These results highlight the importance of analyzing the context of a given application and choosing a fairness definition best serving the cause.

3.2.2 Individual fairness

Dwork et al. [2012] observed that satisfying the requirement of demographic parity might be achieved by accepting qualified individuals from one group and random individuals from another. Thus, satisfying certain notions of group fairness might mean degrading fairness on an individual level. This observation has led to the notion of *individual fairness* which posits that individuals similar with respect to the task at hand should have similar probabilities of positive classification outcomes.

Definitions along these lines have also been investigated in other algorithmic scenarios. For instance, Kearns et al. [2017] proposed a notion of individual fairness for the problem of candidate set selection from diverse incomparable source sets. An example of such a problem is choosing faculty interview candidates from a number of diverse research communities which are not directly comparable to each other in terms of research metrics (for instance, citation rates are different in different research communities). The proposed notion of *meritocratic fairness* requires that less qualified candidates are probabilistically almost never preferred over more qualified candidates when selecting the candidate subsets.

This thesis contributes to the individual fairness line of research by developing methods for making rankings individually fair. Assuming search relevance can be used to determine similarity of individuals with respect to the task, we propose a fairness notion where each ranked individual receives the attention from searchers that is proportional to their relevance. This contribution is presented in Chapter 4.

3.3 Achieving algorithmic fairness

To achieve algorithmic fairness, interventions can be made at different steps of the processing pipeline. A broader overview of various approaches along these lines is provided by Friedler et al. [2018].

Pre-processing methods. Pre-processing methods aim at compensating for biases in the data which might contribute to algorithmic unfairness. Some of the approaches focus on balancing the datasets. For instance, Feldman et al. [2015] modify the numerical attributes in the data to equalize marginal distributions of these attributes conditioned on the sensitive attributes. Hajian and Domingo-Ferrer [2013] propose modifying the values of attributes and labels in a dataset to prevent mining of unfair association rules from the datasets [Pedreschi et al. 2008]. Other approaches construct intermediary (lower-dimensional) representations of the data so as to strip the information about sensitive attributes, while keeping the utility of the modified data for the task at hand [Zemel et al. 2013; Lahoti et al. 2018].

In-processing methods. In-processing approaches try to prevent unfair outcomes by modifying the algorithms. Interventions along these lines most commonly take the form of regularizers reflecting certain soft constraints. When defining an optimization objective for the algorithm training, beyond a component controlling the error, regularizers are introduced to control certain structural properties of models. While primarily used to reduce model complexity and prevent overfitting, regularizers can also capture unfairness of the model.

Fairness regularization has been, for example, considered for classification and regression [Berk et al. 2017; Kamishima et al. 2012], and recommendation [Yao and Huang 2017]. Zafar et al. [2017] propose encoding fairness notions as additional constrains added on top of accuracy optimization objectives.

Post-processing methods. Post-processing approaches modify the outputs of algorithms to satisfy fairness criteria. For example, Fish et al. [2016] propose a method which shifts decision boundaries of trained classifiers to achieve statistical parity, while ensuring a minimal decrease in accuracy. Hardt et al. [2016] modify the decision score thresholds of a trained model to balance the true positive rates of different groups. Kamiran et al. [2010] propose a method for relabeling the nodes of decision tree classifiers to ensure demographic parity.

3.4 Accountability

Once fairness criteria for algorithms are specified, a question remains whether systems actually adhere to such standards and how one can audit systems externally to hold them accountable. Audit mechanisms have been proposed to examine whether protected features influence the outcomes of algorithmic decisions in black-box systems [Adler et al. 2018]. Kilbertus et al. [2018] proposed a method based on encryption of sensitive attributes that enables auditing machine learning models for absence of disparate impact without having the users disclose the values of their sensitive attributes. Kroll et al. [2016] provide an overview of computational techniques that could be applied for assuring compliance of algorithmic outcomes with legal requirements, taking into account that transparency should be limited by the business incentives of service providers.

3.5 Cost of fairness

Satisfying different algorithmic fairness requirements might lead to a decrease in the quality and utility of the algorithmic outputs. This trade-off has been explored for all common algorithmic tasks including classification [Hardt et al. 2016; Zafar et al. 2017]), regression [Berk et al. 2017], and ranking [Zehlike et al. 2017; Singh and Joachims 2018; Biega et al. 2018]. Leonhardt et al. [2018] and Mehrotra et al. [2018] show how increasing the diversity of groups represented in recommendation sets might lead to a decrease in the satisfaction for recommendation consumers.

3.6 Sources of algorithmic unfairness

In search systems, implicit relevance information is often collected from click data. If searchers exhibit bias towards certain groups or individuals, algorithms will learn to imitate these biases in the displayed results. Researchers have uncovered that advertisements of high-paying jobs are shown more often to men than women [Datta et al. 2015], or that advertisements for criminal record checks are more often shown as a response to queries with names commonly associated with African Americans [Sweeney 2013]. The roots of both of

these phenomena can be traced to biased click behaviors of search engine users. Algorithmic unfairness is not necessarily a result of unfair mechanisms, but also of various forms of data and human biases.

Model biases. When translating complex real-world problems into computational tasks, we necessarily need to make certain assumptions and simplifications. For instance, as pointed out by Barocas and Selbst [2016], it is not straightforward to determine what creditworthiness or a good employee exactly mean in terms of machine learning prediction variables. Moreover, very often humans need to be described using simplistic low-dimensional representations, with only selected features present. Some of the features, for instance, even when unproblematic on the surface, might be strongly correlated with the protected attributes. Evaluation metrics will furthermore determine which aspects models optimize for and which will be ignored. Biases might also emerge from the mismatch between the modeling assumptions and the reality of system use contexts. For example, a platform creator might assume that when a recruiter likes a candidate profile on a hiring platform, she thinks the profile is relevant to her search requirement. In reality, however, recruiters might use the feature to simply mark profiles for further investigation. The development choices regarding the task abstraction, features, target variables, and metrics, will influence the bias of the resulting models.

Data biases. Unfairness may stem from underrepresentation of certain populations in the data. For instance, when ranking programming job candidates, more data may be available about male programmers, leading to better system performance for male applications. Note that such underrepresentation might be a result of biased sampling processes, or activity and self-selection biases of platform users (for instance, the platform might be unpopular with female programmers, or women might share less data with the platform on average).

Data might also contribute to algorithmic unfairness when it is re-purposed for a new task. System developers working with such data might not understand the methods and metrics used when collecting the data, or the technical and normative limitations of the platform on which the data was generated. Gebru et al. [2018] propose a standardized dataset description template which could help foster more conscious data reuse practices. Olteanu et al. [2016] provide a detailed discussion of various data biases and limitations.

Human biases. Human biases may enter digital systems in various ways. Machine learning algorithms might be trained on data encoding certain stereotypes – for instance, training labels for the hiring decision task might be provided by annotators with strong gender bias, or generated from historical training data of hiring decisions made when such gender bias was a reflection of the societal reality.

Various cognitive biases influence the way people interact with information and interfaces. For instance, the fact that people tend to scan information from the top down when investigating ranked results leads to position bias, where users pay most of their attention to items ranked high [Joachims et al. 2005]. Eickhoff [2018] has studied how human cognitive biases might influence the results of crowdsourcing studies. Baeza-Yates [2018] discusses further forms of user interaction bias.

Caliskan et al. [2017] has demonstrated that human association biases (perceiving certain semantic concepts as more related than others) are replicated in embeddings trained on text corpora. While the goal of word embeddings is to capture such similarities, certain associations are socially undesirable. For instance, Bolukbasi et al. [2016] show that word embeddings associate men more than women with words related to programming. An association of this kind would be problematic if the embeddings were used as a plug-in component in downstream applications such as resume ranking.

Quantifying bias. A number of efforts have been directed towards measuring and auditing systems for presence of undesired bias. These includes, for instance, empirical studies into gender influencing ranks in ranked outputs in various human resource platforms [Chen et al. 2018], auditing biased practices of surge pricing in ride-hailing platforms [Chen et al. 2015], frameworks for deconstructing input and output biases in search over political social media postings [Kulshrestha et al. 2017], or auditing bias in political personalized search results [Robertson et al. 2018].

Note that the biases discussed here differ from the notion of bias known in statistics, whereby the expected value of a statistical parameter estimator differs from the true parameter value. In particular, even if an estimator is unbiased in the statistical sense, the estimated value might represent an undesirable social phenomenon. For instance, even if a certain minority population underperforms in a school admission test, we might want to modify a statistically unbiased predictor of the performance knowing that the worse performance of the minority is a result of worse access to educational resources. Reversely, a statistically biased performance predictor might not violate any societal fairness notions if, for instance, it underestimates performance equally for everyone.

Friedman and Nissenbaum [1996]; Barocas and Selbst [2016]; Olteanu et al. [2016]; Baeza-Yates [2018] provide comprehensive overviews of the sources of technical, human, and data biases.

3.7 Fairness in search systems

In the context of the main theme of this thesis, it is worth reiterating over the fairness problems specific to search systems. Work in this area has primarily focused on fairness for the search subjects.

Fair representation through diversity. Zehlike et al. [2017] focused on fair representation of protected groups in ranking prefixes, and proposed a statistical fairness test for determining whether a given ranking was generated according to a Bernoulli trial, as well as a post-processing algorithm for reshuffling rankings to pass the fairness test. Celis et al. [2018] study the complexity of the problem of fair representation of groups in rankings.

Fair representation through exposure. Apart from the notions of diversity, fairness based on equal exposure has been proposed in parallel in our work [Biega et al. 2018], discussed in Chapter 4, and by Singh and Joachims [2018]. Singh and Joachims [2018] focus on the notions of group fairness and develop a probabilistic mechanism guaranteeing ex-ante group exposure fairness in expectation. Zehlike and Castillo [2018] follow up on this work by incorporating group-fair exposure regularizers in learning to rank algorithms. The work conducted in this thesis proposes a notion of individual fairness – where each ranked subject should get the attention from searchers that is proportional to her relevance – is explicitly amortized across a sequence of rankings for ex-post fairness.

Quantifying and detecting unfairness. Yang and Stoyanovich [2017] proposed measures to quantify bias in ranked outputs inspired by standard IR evaluation measures, where instead of the relevance information one uses the protected category membership information.

Wu et al. [2018] propose a methodology for analyzing the causality of error in ranked outputs. To this end, the authors construct a directed graph where discrete user profile attributes influence a synthetic score derived from the ranking position, and perform causality analysis on the resulting graph.

A notion of *nutritional label*, wherein different quantitative statistics are presented to the users, has been proposed both for Web documents returned as search results [Fuhr et al. 2017], as well as the rankings themselves [Yang et al. 2018].

Fair ranking quality. On the searchers' side, fairness has been understood as error parity. Mehrotra et al. [2017] proposed a measurement methodology to quantify different levels of satisfaction from the search results for different demographic groups.

3.8 Selected other dimensions in algorithmic fairness research

Human decision making. While the majority of work in the area of algorithmic fairness focuses on machine learning algorithms that replace humans in decision making, some authors have investigated how decision making can be enhanced with humans and automated predictions working in concord [Kleinberg et al. 2017a; Valera et al. 2018].

Human perceptions. Grgic-Hlaca et al. [2018a] have studied human perceptions of fairness with regard to the usage of certain features in machine learning algorithms. Beyond fairness, researchers have sought to understand whether increased interpretability of machine learning models increases the trust people have of these models [Poursabzi-Sangdeh et al. 2018].

Procedural fairness. While the majority of literature focuses on the fairness of outcomes, a separate question is that of procedural fairness, that is, whether the model itself operates in a fair way. Along these lines, Grgic-Hlaca et al. [2018b] have proposed to crowdsource

opinions on whether the usage of certain features is fair in the context of criminal risk prediction and designed a submodular optimization problem to minimize the unfairness of feature use while preserving classifier accuracy.

Fair matching and resource division. Work on fairness in computational economics includes designing incentives for two-sided economy producers to prevent discriminating treatment of consumers [Kannan et al. 2017], or procedures for fair division of resources [Abebe et al. 2017; Chakraborty et al. 2017]. Social and legal scientists have also investigated problems arising from power asymmetries in two-sided economy platforms [Rosenblat and Stark 2016; Calo and Rosenblat 2017].

Ethics of experimentation. Large scale experimentation involving humans, such as A\B testing, raises a lot of ethical concerns. Bird et al. [2016] propose a number principles the design of such experiments should follow, including informed consent from the users, minimizing the potential harm done to the users while maximizing research benefits, and fairly distributing the potential harm risks among the users.

Predictive policing. A number of articles point out the problems predictive policing might lead to. For example, it has been shown that there might exist feedback loops leading to increased police presence in historically over-policed neighborhoods [Lum and Isaac 2016; Ensign et al. 2018].

Long-term effects of fair machine learning. Recent efforts have begun to focus on the long-term impact of fairness constrains, investigating whether fairness interventions proposed in the literature thus far might have undesired effects. For instance, enforcing demographic parity in credit risk prediction might lead to a situation where members of protected groups default more often [Liu et al. 2018].

Chapter 4

Equity of Attention

Contents

Rankings of people and items are at the heart of selection-making, match-making, and recommender systems, ranging from employment sites to sharing economy platforms. As ranking positions influence the amount of attention the ranked subjects receive, biases in rankings can lead to unfair distribution of opportunities and resources such as jobs or income.

This chapter proposes new measures and mechanisms to quantify and mitigate unfairness from a bias inherent to all rankings, namely, the *position bias* which leads to disproportionately less attention being paid to low-ranked subjects. Our approach differs from recent fair ranking approaches in two important ways. First, existing works measure unfairness at

the level of subject *groups* while our measures capture unfairness at the level of *individual* subjects, and as such subsume group unfairness. Second, as no single ranking can achieve individual attention fairness, we propose a novel mechanism that achieves *amortized fairness*, where attention accumulated across a series of rankings is proportional to accumulated relevance.

We formulate the challenge of achieving amortized individual fairness subject to constraints on ranking quality as an online optimization problem and show that it can be solved as an integer linear program. Our experimental evaluation reveals that unfair attention distribution in rankings can be substantial, and demonstrates that our method can improve individual fairness while retaining high ranking quality.

4.1 Introduction

Motivation and Problem. Rankings of subjects like people, hotels, or songs are at the heart of selection, matchmaking and recommender systems. Such systems are in use on a variety of platforms that affect different aspects of life – from entertainment and dating all the way to employment and income. Notable examples of platforms with a tangible impact on people's livelihood include two-sided sharing economy websites, such as Airbnb or Uber, or human-resource matchmaking platforms, such as LinkedIn or TaskRabbit. The ongoing migration to online markets and the growing dependence of many users on these platforms in securing an income have spurred investigations into the issues of bias, discrimination and fairness in the platforms' mechanisms [Calo and Rosenblat 2017; Levy and Barocas 2017].

One aspect in particular has evaded scrutiny thus far – to be successful on these platforms, ranked subjects need to gain the *attention* of searchers. Since exposure on the platform is a prerequisite for attention, subjects have a strong desire to be highly ranked. However, when inspecting ranked results, searchers are susceptible to *position bias*, which makes them pay most of their attention to the top-ranked subjects. As a result, lower-ranked subjects often receive disproportionately less attention than they deserve according to the ranking relevance. Position bias has been studied in information retrieval in scenarios where subjects are documents such as web pages [Craswell et al. 2008; Chuklin et al. 2015]. It has been shown that top-ranked documents receive most clicks often irrespective of their actual relevance [Joachims and Radlinski 2007].

Systemic correction for the bias becomes important when ranking positions potentially translate to financial gains or losses. This is the case when ranking people on platforms like LinkedIn or Uber, products on platforms like Amazon, or creative works on platforms like Spotify. For example, cumulating the exposure on a subset of drivers in ride-hailing platforms might lead to economic starvation of others, while low-ranked artists on music platforms might not get their deserved chance of earning royalties.

Observing that attention is influenced by a human perception bias, while relevance is not, uncovers a fundamental problem: there necessarily exists a discrepancy between the attention that subjects receive at their respective ranks and their relevance in a given search task. For example, attention could decrease geometrically, whereas relevance scores may decrease linearly as the rank decreases. If a ranking is displayed unchanged to many searchers

over time, the lower-ranked subjects might be systematically and repeatedly disadvantaged in terms of the attention they receive.

Problem statement. A vast body of ranking models literature has focused on aligning system relevance scores with the true relevance of ranked subjects, and in this work we assume the two are proportional. What we focus on instead is the relation between relevance and attention. Since relevance can be thought of as a proxy for worthiness in the context of a given search task, the attention a subject receives from searchers should ideally be proportional to her relevance. In economics and psychology, a similar idea of proportionality exists under the name of equity [Walster et al. 1973] and is employed as a fairness principle in the context of distributive justice [Greenberg 1987]. Thus, in this thesis, we make a translational normative claim and argue for *equity of attention* in rankings.

Operationally, the problem we address in this thesis is to devise measures and mechanism which ensure that, for all subjects in the system, the received attention approximately equals the deserved attention, while preserving ranking quality. For a single ranking this goal is infeasible, since attention is influenced by the position bias, while relevance is not. Therefore, our approach looks at a series of rankings and aims at measures of *amortized fairness*.

State of the art and limitations. Fairness has become a major concern for decision-making systems based on machine learning methods. Various notions of *group fairness* have been investigated [Kamishima et al. 2012; Pedreschi et al. 2008; Feldman et al. 2015; Hardt et al. 2016; Zafar et al. 2017], with the goal of making sure that protected attributes such as gender or race do not influence algorithmic decisions. Fair classifiers are then trained to maximize accuracy subject to group fairness constraints. These approaches, however, do not distinguish between different subjects from within a group. The notion of *individual fairness* [Dwork et al. 2012; Zemel et al. 2013; Kearns et al. 2017] aims at treating each individual fairly by requiring that subjects who are similar to each other receive similar decision outcomes. For instance, the concept of *meritocratic fairness* requires that less qualified candidates are almost never preferred over more qualified ones when selecting candidates from a set of diverse populations. Relevance-based rankings, where more relevant subjects are ranked higher than less relevant ones, also satisfy meritocratic fairness. A stronger fairness concept, however, is needed for rankings to be a means of distributive justice.

Prior work on *fair rankings* is scarce and includes approaches that perturb results to guarantee various types of group fairness. This goal is achieved by techniques similar to those for ranking result diversification [Celis et al. 2018; Yang and Stoyanovich 2017; Zehlike et al. 2017], or by granting equal ranking exposure to groups [Singh and Joachims 2018]. Individual fairness is inherently beyond the scope of group-based perturbation.

Approach and contribution. Our approach in this thesis differs from the prior work in two major ways. First, the measures introduced here capture fairness at the level of *individual* subjects, and subsume group fairness as a special case. Second, as no single ranking can guarantee fair attention to every subject, we devise a novel mechanism that ensures *amortized fairness*, where attention is fairly distributed across a series of rankings.

For an intuitive example, consider a ranking where all the relevance scores are almost the same. Such tiny differences in relevance will push subjects apart in the display of the results, leading to a considerable difference in the attention received from searchers. To compensate for the position bias, we can reorder the subjects in consecutive rankings so that everyone who is highly relevant is displayed at the top every now and then.

Our goal is not just to balance attention, but to keep it proportional to relevance for all subjects while preserving ranking quality. To this end, we permute subjects in each ranking so as to improve fairness subject to constraints on quality loss. We cast this approach to an online optimization problem, formalizing it as an integer linear program (ILP). We moreover devise filters to prune the combinatorial space of the ILP, which ensures that it can be solved in an online system. Experiments with synthetic and real-life data demonstrate the viability of our method.

Note that we assume that searchers are indifferent to the varying ordering in the ranking except for the utility loss. In practice, searchers might be confused if they see very different results to the same queries. While this thesis does not tackle this problem, possible solutions might include incremental updates to ranking changes, or controlling that varying results to a given query are shown to different searchers.

This chapter makes the following novel contributions:

- To the best of our knowledge, we are the first to formalize the problem of individual equity-of-attention fairness in rankings, and define measures that capture the discrepancy between the deserved and received attention.
- We propose online mechanisms for fairly amortizing attention over time in consecutive rankings.
- We investigate the properties and behavior of the proposed mechanisms in experiments with synthetic and real-world data.

4.2 Equity-of-attention fairness

We now formally define equity of attention accounting for *position bias*, which determines how attention is distributed over the ranking positions. We consider a sequence of rankings at different time points, by different criteria or on request of different users.

4.2.1 Notation

We use the following notation:

- $u_1, ..., u_n$ is a set of subjects ranked in a system,
- $\rho^1, ..., \rho^m$ is a sequence of rankings,
- r_i^j is the [0..1]-normalized relevance score of subject u_i in ranking ρ^j,
- a_i^j is the [0..1]-normalized attention value received by subject u_i in ranking ρ^j,
- A denotes the distribution of cumulated attention across subjects, that is, $A_i = \sum_{j=1}^{m} a_i^j$ for subject u_i,

- R denotes the distribution of cumulated relevance across subjects, that is, $R_i = \sum_{j=1}^{m} r_i^j$ for subject u_i.

4.2.2 Defining equity of attention

Our fairness notion in this work is in the spirit of the *individual fairness* proposed by Dwork et al. [2012], which requires that "similar individuals are treated similarly", where "similarity" between individuals is a metric capturing suitability for the task at hand. In the context of rankings, we consider *relevance* to be a measure of subject suitability. Further, in applications where rankings influence people's economic livelihood, we can think of rankings not as an end, but as a means of achieving distributive justice, that is, fair sharing of certain real-world resources. In the context of rankings, we consider the *attention of searchers* to be a resource to be distributed fairly.

There exist different types of distributive norms, one of them being *equity*. Equity encodes the idea of proportionality of inputs and outputs [Walster et al. 1973], and might be employed to account for "differences in effort, in productivity, or in contribution" [Yaari and Bar-Hillel 1984].

Building upon these ideas, we make a translational normative claim and propose a new notion of individual fairness for rankings called *equity of attention*, which requires that *ranked subjects receive attention that is proportional to their worthiness in a given search task*. As a proxy for worthiness, we turn to the currently best available ground truth, that is, the system-predicted relevance.

Definition 1 (Equity of Attention)**.** *A ranking offers equity of attention if each subject receives attention proportional to its relevance:*

$$\frac{a_{i1}}{r_{i1}} = \frac{a_{i2}}{r_{i2}}, \ \forall u_{i1}, u_{i2}. \tag{4.1}$$

Note that this definition is unlikely to be satisfied in any single ranking, since the relevance scores of subjects are determined by the data and the query, while the attention paid to the subjects (in terms of views or clicks) is strongly influenced by position bias. The effects of this mismatch will be aggravated if multiple subjects are similarly relevant, yet obviously cannot occupy the same ranking position and receive similar attention.

To operationalize our definition in practice, we propose an alternative fairness definition that requires *attention to be distributed proportionally to relevance, when amortized over a sequence of rankings.*

Definition 2 (Equity of Amortized Attention)**.** *A sequence of rankings $\rho^1, ..., \rho^m$ offers equity of amortized attention if each subject receives cumulative attention proportional to her cumulative relevance, i.e.:*

$$\frac{\sum_{l=1}^{m} a_{i1}^l}{\sum_{l=1}^{m} r_{i1}^l} = \frac{\sum_{l=1}^{m} a_{i2}^l}{\sum_{l=1}^{m} r_{i2}^l}, \ \forall u_{i1}, u_{i2}. \tag{4.2}$$

Observe that this modified fairness definition allows us to permute individual rankings so as to satisfy fairness requirements over time. The deficiency in the attention received by a subject relative to her relevance in a given ranking instance can be compensated in a subsequent ranking, where the subject is positioned higher relative to her relevance.

4.2.3 Equality of attention

In certain scenarios, it may be desirable for subjects to receive the same amount of attention, irrespective of their relevance. Such is the case when we suspect the ranking is biased and cannot confidently correct for that bias, or when the subjects are not shown as an answer to any query but need to be visually displayed in a ranked order (e.g., a list of candidates on an informational website for an election). In such scenarios, the desired notion of fairness would be *equality of attention*. We observe that this egalitarian version of fairness is a special case of equity of attention, where the relevance distributions are uniform, i.e., $r_{i1} = r_{i2}\ \forall u_{i1}, u_{i2}$. As equity of attention subsumes equality of attention, we do not explicitly discuss it further in this thesis.

4.2.4 Relation to group fairness in rankings

To our knowledge, all prior works on fairness in rankings have focused on notions of *group fairness*, which define fairness requirements over the collective treatment received by all members of a demographic group like women or men. Our motivation for tackling fairness at the individual level stems from the fact that position bias affects all individuals, independently of their group membership. It is easy to see, however, that when equity of attention is achieved for individuals, it will also be achieved at the group level: the cumulated attention received by all members of a group will be proportional to their cumulated relevance.

Prior works on fairness in rankings [Celis et al. 2018; Yang and Stoyanovich 2017; Zehlike et al. 2017] has mostly focused on diversification of the results. These approaches are geared for one-time rankings, and, as any static model, will steadily accumulate equity-of-attention unfairness over time. Since they were developed with a different goal in mind, they are not directly comparable to our dynamic approach.

Parallel with our work, Singh and Joachims [2018] have explored similar ideas of how position bias influences fairness of exposure. Their probabilistic formulations are possibly a counterpart of our amortization ideas, and it will be interesting to see to what extent these formulations are interchangeable. In line with other prior works on fairness in rankings and different from our work, however, they focus on satisfying constraints on group rather than individual fairness, and on notions of equality rather than equity.

4.3 Rankings with equity of attention

4.3.1 Measuring (un)fairness

To be able to optimize ranking fairness, we need to measure to what extent a sequence of rankings $\rho^1, ..., \rho^m$ violates Definition 2. Since the proposed fairness criterion is equivalent to the requirement that the empirical distributions A and R be equal, we can measure unfairness as the distance between these two distributions. A variety of measures can be applied here, including KL-divergence, or L1-norm distance. In this work, we measure fairness using the latter:

$$unfairness(\rho^1, ..., \rho^m) = \sum_{i=1}^{n} |A_i - R_i| = \sum_{i=1}^{n} \left| \sum_{j=1}^{m} a_i^j - \sum_{j=1}^{m} r_i^j \right|. \tag{4.3}$$

L1-norm is minimized with a value of 0 for distributions satisfying the fairness criterion from Definition 2, and is thus useful as an optimization objective. However, since the measure is cumulative and indifferent to the exact distribution of unfairness among individuals, other measures could be developed to *quantify* unfairness in the system at any given point.

4.3.2 Measuring ranking quality

Permuting a ranking to satisfy fairness criteria can lead to a quality loss when less relevant subjects get ranked higher than more relevant ones. We propose to quantify ranking quality using measures that draw from IR evaluation. Traditionally, ranking models are evaluated in comparison with ground-truth rankings based on human-given relevance labels. Here, we are interested in quantifying the divergence from the original ranking. Thus, we consider the *original ranking* ρ to be the ground-truth reference for evaluating the quality of a *reordered ranking* ρ^*. We assume that the ground truth scores are the relevance scores returned by the system, and that these scores reflect the best ordering of subjects. These considerations lead to the following definitions.

Discounted cumulative gain (DCG) quantifies the quality of a ranking by summing the relevance scores in consecutive positions with a logarithmic discount for the values at lower positions. The measure thus puts an emphasis on having higher relevance scores at top positions.

$$DCG@k(r) = \sum_{i=1}^{k} \frac{2^{r(i)} - 1}{log_2(i+1)} \tag{4.4}$$

This value can be further normalized by the DCG score of a perfect ranking ordered by the ground truth relevance scores. The normalized discounted cumulative gain (NDCG)-based quality measure can be thus expressed as:

$$NDCG\text{-}quality@k(\rho, \rho^*) = \frac{DCG@k(\rho^*)}{DCG@k(\rho)} \tag{4.5}$$

This measure is maximized with a value of 1 if the rankings do not differ or if swaps are only made within ties (i.e., subjects with equal relevance). Other measures, like Kendall's Tau or appropriately defined $MAP\text{-}quality$, could be applied as well.

4.3.3 Optimizing fairness-quality tradeoffs

As discussed in the previous section, there is "no free lunch": to improve fairness, we need to perturb relevance-based rankings, which might lead to lower ranking quality. To address the tradeoff, we can formulate two types of constrained optimization problems: one where we minimize unfairness subject to constraints on quality (i.e., lower-bound the minimum acceptable quality), and another where we maximize quality subject to constraints on unfairness (i.e., upper-bound the maximum acceptable unfairness measure). In this thesis, we focus on the former, since at the moment ranking quality measures are more interpretable, and so are the constraints on quality.

4.3.3.1 Offline optimization

Let $\rho^1, ..., \rho^m$ be a sequence of rankings where the subjects are ordered by the relevance scores. These rankings induce zero quality loss. We wish to reorder them into $\rho^{1*}, ..., \rho^{m*}$ so as to minimize the distance between the distributions A and R with constraints on NDCG-quality loss in each ranking:

$$\begin{aligned} &\text{minimize} \quad \sum_i |A_i - R_i| \\ &\text{subject to} \quad \textit{NDCG-quality@}k(\rho^j, \rho^{j*}) \geq \theta,\ j = 1, \ldots, m. \end{aligned} \tag{4.6}$$

where A_i and R_i denote the cumulated attention and relevance scores that subject u_i has gained across all the m rankings.

Instead of thresholding the loss in each individual ranking, an alternative would be to threshold the average loss over m rankings.

4.3.3.2 Online optimization

In practice, ranking amortization needs to be done in an online manner, one query at a time. Without the knowledge of future query loads, the goal is then to reorder the current ranking so as to minimize unfairness over the cumulative attention and relevance distributions in rankings seen so far, subject to a constraint on the quality of the current ranking. Thus, in the l-th ranking we want to :

$$\begin{aligned} &\text{minimize} \quad \sum_i |A_i^{l-1} + a_i^l \ - \ (R_i^{l-1} + r_i^l)| \\ &\text{subject to} \quad \textit{NDCG-quality@}k(\rho^l, \rho^{l*}) \geq \theta \end{aligned} \tag{4.7}$$

where A_i^{l-1} and R_i^{l-1} denote the cumulated attention and relevance scores that subject u_i has gained up to and including ranking ρ^{l-1}.

4.3.4 An ILP-based fair ranking mechanism

4.3.4.1 ILP for online attention amortization

The optimization problem defined in Sec. 4.3.3.2 can be solved as an integer linear program (ILP). Assume we are to rerank the l-th ranking in a series of rankings. We introduce n^2 decision variables $X_{i,j}$ which are set to 1 if subject u_i is assigned to the ranking position j, and set to 0 otherwise. At the time of reordering the l-th ranking, the following values are constants:

- relevance scores for each subject u_i in the current ranking: r_i^l,
- attention values assigned to ranking positions: w_j,
- relevance scores accumulated up to (and excluding) the current ranking for each subject: R_i^{l-1},
- attention values accumulated up to (and excluding) the current ranking for each subject: A_i^{l-1},

- IDCG@k value computed over the current ranking ρ_l, which is used as a normalization score for NDCG-quality@k.

For each subject u_i, the accumulated attention and relevance are initialized as $A_i^0 = 0$ and $R_i^0 = 0$ for all u_i.

The ILP is then defined as follows:

$$\begin{aligned} \text{minimize} \quad & \sum_{i=1}^{n}\sum_{j=1}^{n} |A_i^{l-1} + w_j - (R_i^{l-1} + r_i^l)| \cdot X_{i,j} \\ \text{subject to} \quad & \sum_{j=1}^{k}\sum_{i=1}^{n} \frac{2^{r_i^l} - 1}{log_2(j+1)} X_{i,j} \geq \theta \cdot IDCG@k \\ & X_{i,j} \in \{0,1\},\ \forall_{i,j} \\ & \sum_i X_{i,j} = 1,\ \forall_j \\ & \sum_j X_{i,j} = 1,\ \forall_i \end{aligned} \tag{4.8}$$

The first constraint bounds the loss in ranking quality, in terms of the NDCG-quality measure, by the multiplicative threshold $0 \leq \theta \leq 1$. The other constraints ensure that the solution is a bijective mapping of subjects onto ranking positions. The terms $A_i^{l-1} + w_j$ and $R_i^{l-1} + r_i^l$ encode the updates of the cumulative attention and relevance, respectively, if and only if u_i is mapped to position j.

It is worth noting that:

- When $\theta = 1$, we do not allow any quality loss. This, however, does not mean that the ranking will remain unchanged. Subjects can be reordered within ties to minimize unfairness.
- When $\theta = 0$, any permutation of the ranking is allowed striving to minimize unfairness in the current iteration.

4.3.4.2 ILP with candidate pre-filtering

The ILP operates on a huge combinatorial space, with the number of binary variables being quadratic in the number of subjects. Real systems deal with millions of subjects, and the optimization needs to be carried out each time a new ranking is requested. Such a problem size is a bottleneck for ILP solvers, and in practice the optimization needs to use approximation algorithms, such as LP relaxations or greedy-style heuristics. This is one of the directions for further research.

To deal with the issue in this work, instead of reranking all subjects in each iteration, we rerank only subjects from a prefiltered candidate set. Different strategies are possible for selecting the candidate sets. On the one hand, prefiltering the top-ranked subjects by relevance scores would let us satisfy the quality constraints, but may entail small fairness gains, especially for near-uniform relevance distributions. On the other hand, prefiltering based on the objective function might lead to situations where the ILP cannot find any solution without violating the constraints.[1]

[1] Without prefiltering, the ILP always has at least one feasible solution (the original ranking).

Our strategy thus is as follows. Assume we want to select a subject candidate subset D of size t to be reranked, and we constrain the quality in Eq. 4.8 at rank k. Since the attention weights w_j are positive, the biggest contributors to the objective function are the subjects with the smallest values of $A_i - (R_i + r_i)$. These are the subjects with the highest deficit (negative value) of fair-share attention. We always select k subjects with the highest relevance scores in r^l, to make sure we satisfy the quality constraint, plus other $t - k$ subjects with the lowest $A_i - (R_i + r_i)$ values, who are most worthy of being promoted to high ranks. As a result, when no feasible solution can be found by reranking the most worthy subjects, the ILP will default to choosing the top-k candidates by relevance scores.

4.3.4.3 Extensions

Granularity. The presented model assumes that attention and relevance are aggregated per ranked subject. It is straightforward to extend it to handle higher-level actors such as product brands or Internet domains, by summing the relevance and attention scores over the corresponding subjects. As a consequence of this modification, bigger organizations would obtain higher exposure. Deciding whether this effect is fair is a policy issue.

Handling dynamics. In a real-world system, the size of the population will vary over time, with new subjects joining and existing ones dropping out. Our model is capable of handling this kind of dynamics, since new users starting with no deserved attention will be positioned in between the users who got more than they deserved and those who got less. Moreover, ranking quality constraints will prevent such users from being positioned too low in rankings where they are highly relevant.

4.4 Experiments

4.4.1 Data

The datasets we use are either synthetically generated or derived from other publicly available resources. They are freely available to other researchers.

4.4.1.1 Synthetic datasets

We create 3 synthetic datasets to analyze the performance of the model in a controlled setup under different relevance distributions. We assume the following distribution shapes: (i) **uniform**, where every user has the same relevance score, (ii) **linear**, where the scores decrease linearly with the rank position, and (iii) **exponential**, where the scores decrease exponentially with the rank position. Each dataset has 100 subjects.

4.4.1.2 Airbnb datasets

To analyze the model in a real-world scenario, we construct rankings based on Airbnb[2] apartment listings from 3 cities located in different parts of the world: Boston, Geneva, and

[2] https://www.airbnb.com/

Hong Kong. Airbnb is a two-sided sharing economy platform allowing people to offer their free rooms or apartments for short-term rental. It is a prime example of a platform where exposure and attention play a crucial role in the subjects' financial success. The data we use is freely available for research.[3]

Rankings are constructed using the attribute *id* as a subject identifier, and various review ratings as the ranking criteria, with the rating scores serving as relevance scores. Such crowd-sourced judgments serve as a good worthiness-of-attention proxy on this particular platform, although one has to have in mind that rating distributions tend to be skewed towards higher scores, which is confirmed by our experimental analysis.

For each of the 3 datasets, we run the amortization model on two types of ranking sequences:

1. **Single-query**: We examine the amortization effects when a single ranking is repeated multiple times. To construct the rankings, we use the values of the *review_scores_rating* attribute, which corresponds to the overall quality of the listing.

2. **Multi-query**: We examine the behavior of the model when a sequence of rankings, each with a different relevance distribution, is repeated multiple times. To this end, for each city, we construct 7 rankings based on different rating attributes:
   review_scores_rating,
   review_scores_accuracy,
   review_scores_cleanliness,
   review_scores_checkin,
   review_scores_communication,
   review_scores_location,
   and *review_scores_value*.

The datasets for Boston, Geneva, and Hong Kong contain 3944, 1728, and 4529 subjects, respectively.

Note that, for the purpose of model performance evaluation, the queries themselves become irrelevant once the relevance is computed. Since the values of the aforementioned attributes serve as relevance scores, the queries are abstracted out.

4.4.1.3 StackExchange dataset

We create another dataset from a querylog and a document collection synthesized from the StackExchange dump by Biega et al. [2017b], please refer to the original paper for details. We choose a radom subset of users and order their queries by timestamps, creating a workload of around 20K queries. We use Indri[4] to retrieve 500 most relevant answers for each query, and treat the author of the answer as the subject to be ranked. Using this dataset helps us gain an insight into the performance of the method in core IR tasks and with different sets of subjects ranked in each iteration.

[3] Downloaded from `http://insideairbnb.com/`
[4] `https://www.lemurproject.org/indri/`

4.4.2 Position bias

Our model requires that we assign a weight to each ranking position, denoting the fraction of the total attention the position gets. These weights will depend on the application and platform, and may be estimated from historical click data. In this thesis, we study the behavior of the equity-of-attention mechanism under generic models of attention distribution. We focus on the following distributions:

1. **Geometric**: The weights of the positions are distributed geometrically with the parameter p up to the position k, and are 0 for positions lower than k. Geometrically distributed weights are a special case of the cascade model [Craswell et al. 2008], where each subject has the same probability p of being clicked. Setting the weights of lower positions to 0 is based on an assumption that low-ranked subjects are not inspected.

$$w_j = \begin{cases} p(1-p)^{j-1} & j \leq k \\ 0 & j > k \end{cases} \tag{4.9}$$

2. **Singular**: The top-ranked subject receives all the attention. This is a special case of the geometric attention model with parameters $p = 1, k = 1$. Studying this attention model is motivated by systems such as Uber, which present only top-1 matches to the searchers by default.

$$w_j = \begin{cases} 1 & j = 1 \\ 0 & j > 1 \end{cases} \tag{4.10}$$

Before being passed on to the model, the weights are rescaled such that $\sum_j w_j = 1$. Studying the effects of position bias on individual fairness under more complex attention models is future work.

4.4.3 Implementation and parameters

We implement the ILP-based amortization defined in Section 4.3.4 using the Gurobi software.[5] Constraints are set to be satisfied up to a feasibility threshold of $1e-7$. We prefilter 100 candidates for reranking in each iteration, as described in Section 4.3.4.2.

In the singular attention model, since all the attention is assumed to go to the first ranking position, the ILP constrains the NDCG-quality at rank $k = 1$. We construct the geometric attention model with $p = 0.5$ and $k = 5$, and in this case the ILP constraints the NDCG-quality at rank $k = 5$.

In the single-query mode, where a single ranking is repeated multiple times, we set the number of iterations to $20K$. In the multi-query mode, with a repeated sequence of different rankings, we repeat the whole sequence $3K$ times, which leads to a total of $21K$ rankings.

Relevance scores in the framework need to be normalized to form a distribution. In this work, we assume relevance is a direct proxy for worthiness and rescale the rating scores linearly. Note, however, that if additional knowledge is available to the platform regarding

[5] http://www.gurobi.com/

the correspondence between relevance and worthiness, other transformations can be applied as well.

4.4.4 Mechanisms under comparison

We compare the performance of the ILP-based online mechanism against two baseline heuristics.

1. **Relevance**: The first heuristic is to allow only relevance-based ranking, completely disregarding fairness.

2. **Objective**: The second heuristics is an objective-driven ranking strategy, which orders subjects by the increasing priority value: $A_i - R_i - r_i$ (see Sec. 4.3.4.2) for each ranking. Since all position weights w_j are positive, assigning highest weights to subjects with the lowest preference value is in line with the minimization goal. This ranking strategy aims at strong fairness amortization without any quality constraints, and is expected to perform similarly to the ILP with $\theta = 0$.

4.4.5 Data characteristics: relevance vs. attention

Figure 4.1 shows the relevance score distributions in the single-query Airbnb datasets for Boston, Geneva, and Hong Kong. The seemingly flatter shape of the Boston and Hong Kong distributions is the result of a bigger size of these datasets when compared to the Geneva dataset, where each individual has, on average, a larger fraction of the total relevance. Overall, the distributions have a shape which complements the uniform, linear, and exponential shapes of distributions in the synthetic datasets.

Figure 4.2 presents an example strongly motivating our research. Namely, it compares the distribution of relevance in the Geneva dataset with the distribution of attention according to the geometric model with $p = 0.5$, where the weights closely follow the empirical observations made in previous position bias studies [Joachims and Radlinski 2007]. Observe that the relevance distribution plotted in green is the same as that in Figure 4.1. There is a huge discrepancy between these two distributions, while, as argued in this thesis, they should ideally be equal to ensure individual fairness. Similar discrepancy exists in the two other Airbnb datasets.

4.4.6 Performance on synthetic data

Singular attention model. Figure 4.3 reveals a number of interesting properties of the mechanism for the Uniform relevance distribution. We plot the iteration number on the x-axis, and the value of the unfairness measure defined by Equation 4.3 on the y-axis. First, since reshuffling does not lead to any quality loss when all the relevance scores are equal, all the reshuffling methods perform equally well irrespective of θ. Their amortizing behavior should be contrasted with the black line denoting the relevance baseline. Unfairness for this method always increases linearly by a constant factor incurred by the single ranking. Second, amortization methods periodically bring unfairness to 0. The minimum occurs every

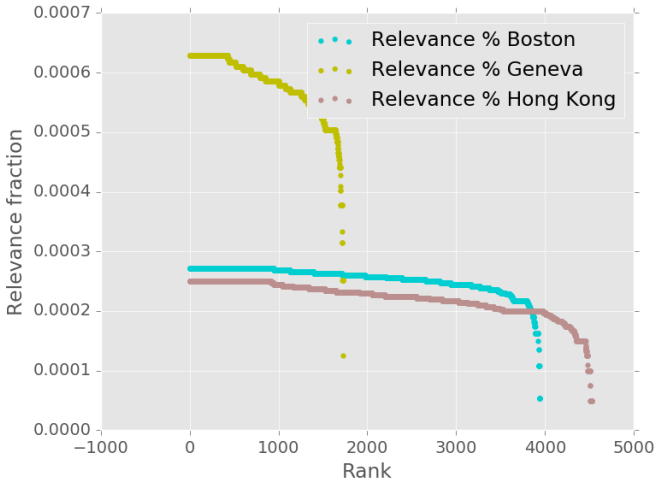

Figure 4.1: Relevance distributions in the Airbnb datasets.

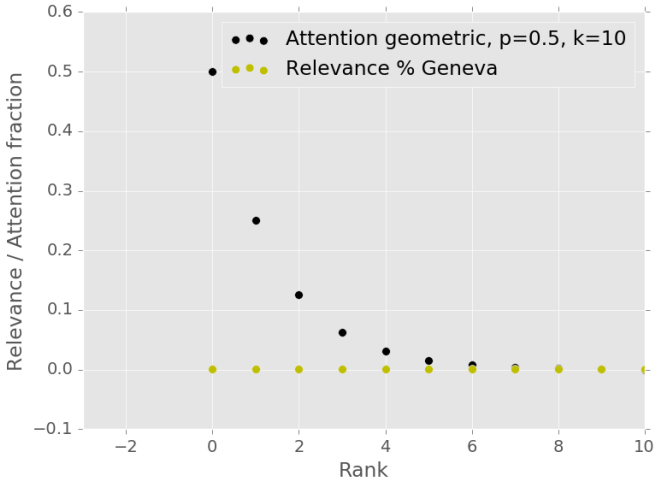

Figure 4.2: Comparison of the attention and relevance distributions for the top-10 ranking positions in the Geneva dataset. Note that the relevance distribution presented here is the same as in Fig. 4.1. To satisfy equity-of-attention fairness, the two distributions would have to be the same.

n iterations, where n is the number of subjects in the dataset. Within the cycle, each subject is placed in the top position (receiving all the attention) exactly once.

Figure 4.4 with the results for the Linear dataset, confirms another anticipated behavior. With no ties in the relevance scores, it is not possible to improve fairness without incurring quality loss. Thus, all methods with $\theta > 0$ lead to higher unfairness when compared to the Objective baseline, although the unfairness is still lower in ILP with $\theta < 0.8$ than in the Relevance baseline.

When the relevance scores decrease exponentially (Figure 4.5), the ILP is not able to satisfy the quality constraint with any $\theta >= 0.5$, and thus these rerankings become equivalent to those of the Relevance heuristic.

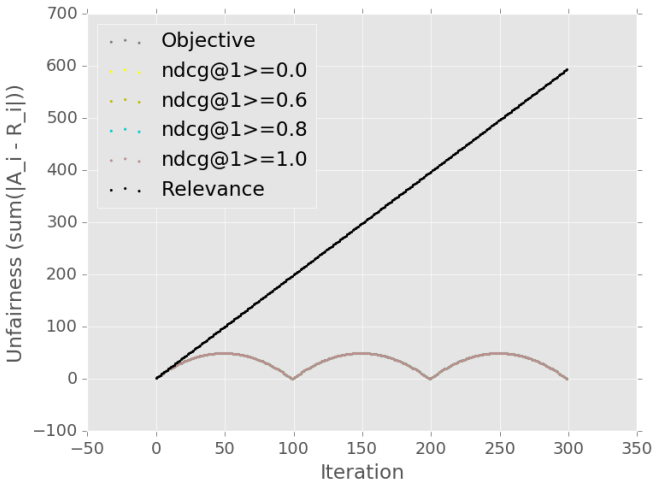

Figure 4.3: Model performance on the synthetic Uniform dataset. Attention singular.

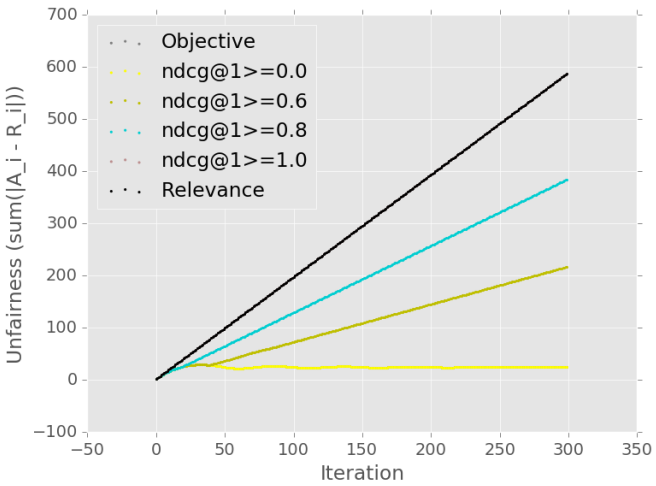

Figure 4.4: Model performance on the synthetic Linear dataset. Attention singular.

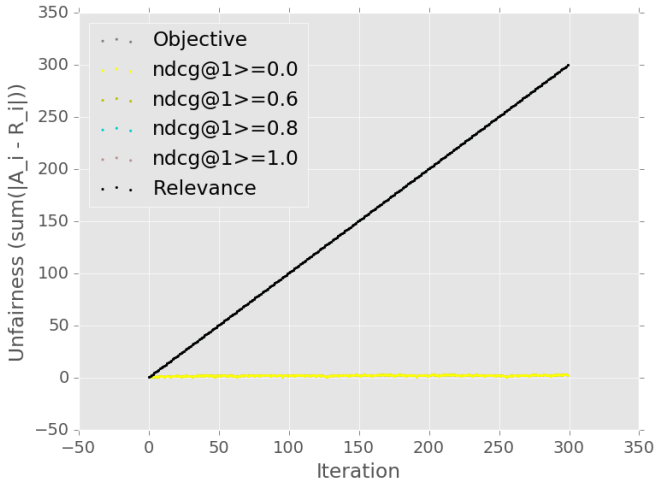

Figure 4.5: Model performance on the synthetic Exponential dataset. Attention singular.

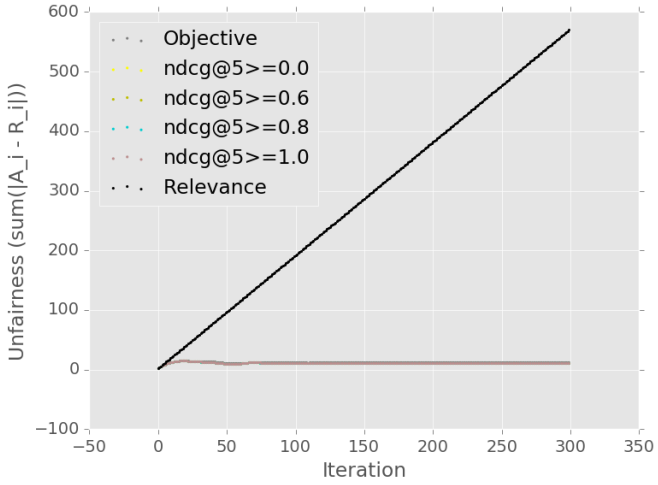

Figure 4.6: Model performance on the synthetic Uniform dataset. Attention geometric.

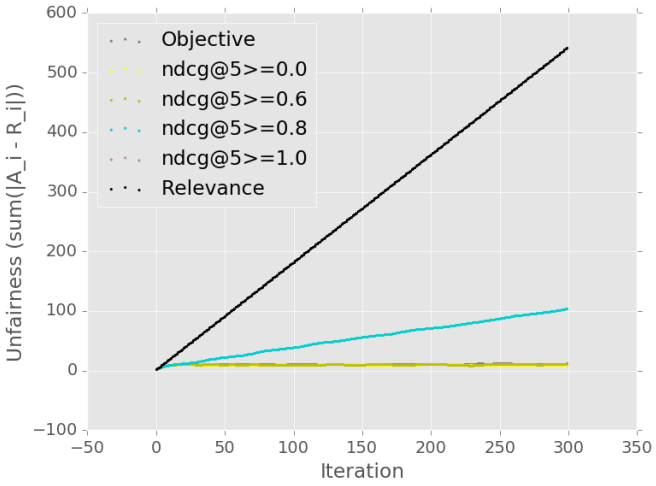

Figure 4.7: Model performance on the synthetic Linear dataset. Attention geometric.

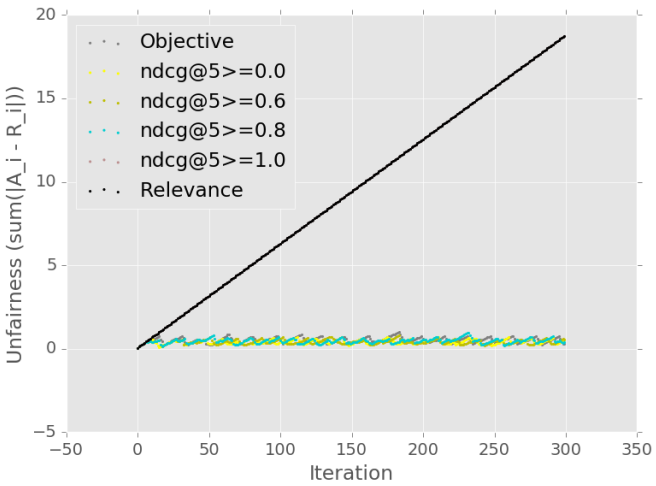

Figure 4.8: Model performance on the synthetic Exponential dataset. Attention geometric.

Geometric attention model. As shown in Figures 4.6, 4.7, and 4.8, the periodicity effect becomes less pronounced under the general geometric attention model. Figure 4.9 helps to understand this behavior by showing the unfairness values achieved by the Objective heuristic with different values of the attention cut-off k (see Equation 4.9). With $k = 1$, the model is equivalent to Singular. As we increase k, the distribution of the position weights becomes smoother, smoothing also the periodicity of the unfairness values.

The very good performance of the ILP-based rerankings with any $\theta < 1$ in Figure 4.8, stems from the fact that the relevance and attention distributions are almost the same (the only difference being that the scores in the relevance distribution are non-zero for more positions). Our results show that in this case the ILP performs a reordering only every now and then, when the subjects ranked lower than position 5 in the original ranking gather enough deserved attention. This causes the unfairness to go up and down periodically.

4.4.7 Performance on Airbnb data

4.4.7.1 Single-query, singular attention

We first analyze the model performance on the Airbnb datasets where a single ranking is repeated multiple times, and the attention model is set to singular. The results are shown in Figures 4.10, 4.11, 4.12 for Boston, Geneva, and Hong Kong, respectively. As in the analysis with the synthetic data, we plot the iteration number on the x-axis, and the value of the unfairness measure defined by Equation 4.3 on the y-axis. There are a number of observations:

- As noted before, the loss in the Relevance baseline (plotted in black) increases linearly by the constant unfairness factor incurred by the single ranking.

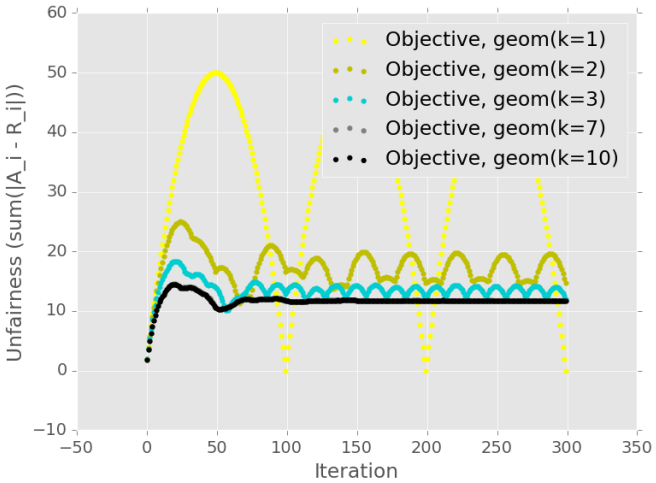

Figure 4.9: Performance of the Objective heuristic on the synthetic Uniform dataset under the geometric attention model with different attention cut-off points.

- Relaxing the quality constraint by decreasing θ allows us to achieve lower unfairness values in the corresponding ranking iterations.

- The Objective heuristic with no quality constraints and the ILP where $\theta = 0$ are able to amortize fairness over time well, with no significant growth of unfairness over time.

- The periodicity effect we observed on synthetic uniform data appears here as well. This is due to the relative closeness of the relevance distributions in the Airbnb data to the uniform distribution. Unfairness achieved by the amortizing methods is close to 0 every n iterations. The frequency of the minimum indeed corresponds to the size of the respective datasets.

- In some methods unfairness starts to grow linearly after a certain number of iterations (see, e.g., the blue curve in Figure 4.10). This is a side effect of the candidate prefiltering heuristic we chose. When the ILP receives a filtered candidate set where no subjects filtered based on the objective can be placed at the top of the ranking without violating the quality constraint, the ILP defaults to placing the most relevant subjects at the top, which causes the quality loss to be 0 and the unfairness growing linearly. This effect persists until some of the more relevant subjects gather enough deserved attention to be pre-selected - note the variability that occurs in the blue curve again starting around the 17K-th iteration.

- For a number of iterations at the beginning (equal to the number of ties at the top of the ranking), all the methods perform the same, irrespective of the quality constraints. This is due to the fact that unfairness is minimized by reshuffling the most deserving relevant subjects first, which does not incur any quality loss.

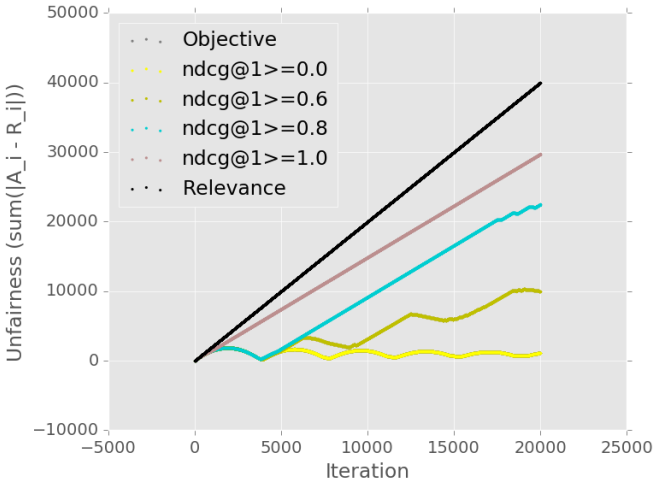

Figure 4.10: Model performance on the single-query Boston dataset. Attention singular.

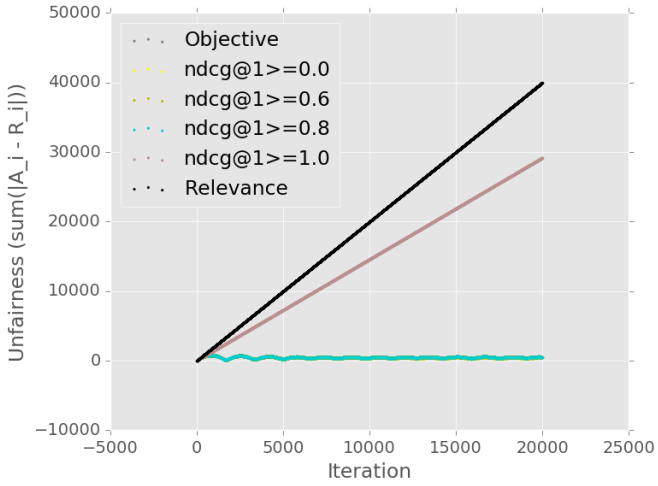

Figure 4.11: Model performance on the single-query Geneva dataset. Attention singular.

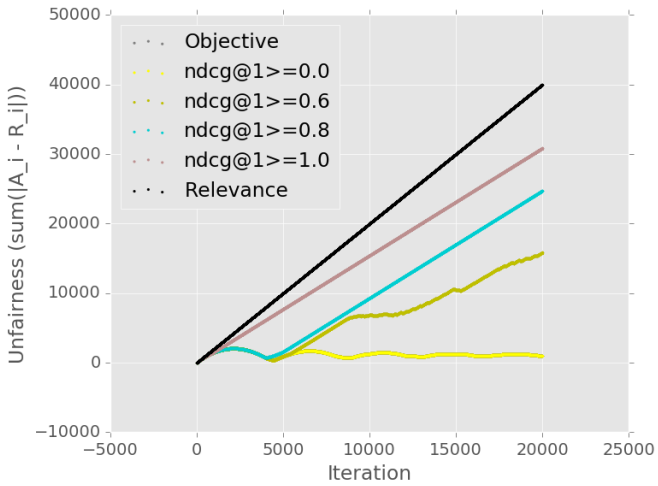

Figure 4.12: Model performance on the single-query Hong Kong dataset. Attention singular.

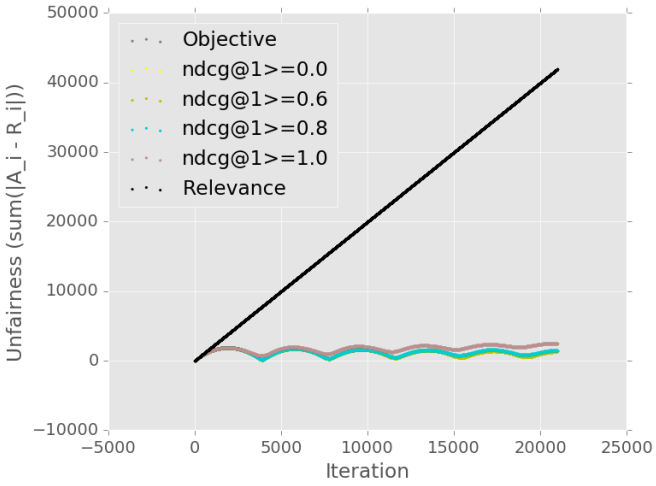

Figure 4.13: Model performance on the multi-query Boston dataset. Attention singular.

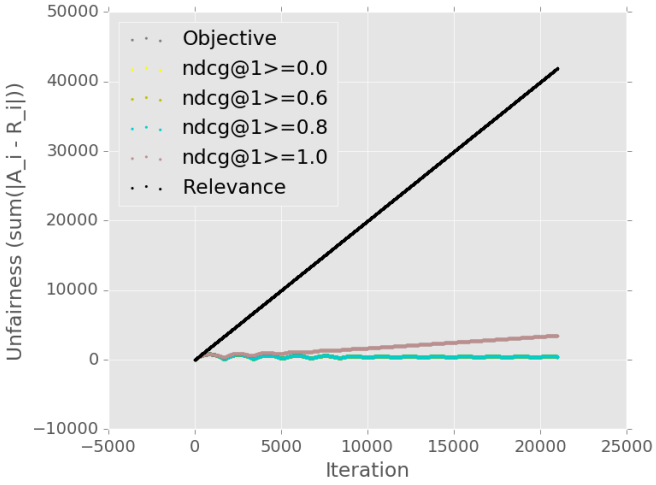

Figure 4.14: Model performance on the multi-query Geneva dataset. Attention singular.

4.4.7.2 Multi-query, singular attention

Our methods amortize fairness better (achieving lower unfairness) on the Airbnb multi-query datasets (Figures 4.13, 4.14, and 4.15) when compared to the single-query datasets for two reasons. First, the variability in subject relevance and ordering in different iterations is a factor helpful in smoothing the deserved attention distributions over time. Second, distributions of the rating attributes in the Airbnb datasets used to construct the rankings are more uniform than the global rating score, and have more ties at the top of the ranking. These relevance distribution characteristics enable methods with conservative quality constraints (even the ILP with $\theta = 1$) to perform very well.

4.4.7.3 Single-query, geometric attention

The general geometric attention distribution is closer to the relevance distributions in the Airbnb datasets than the singular distribution is. As noted in the analysis with synthetic data, the closeness of the two distributions helps amortize fairness at a lower quality loss. We can observe a similar effect in Figure 4.16, with more ILP-based methods reaching the performance of the Objective heuristic. Note, however, that the improved performance here is also partly due to the fact that we constrain the quality at a higher rank when assuming the geometric attention, which is easier to satisfy.

4.4.7.4 Unfairness vs. quality loss

The results presented so far show the performance of the ILP-based fairness amortization under different quality thresholds. Since the thresholds bound the maximum quality loss over all iterations, the actual loss in most cases might be lower. To investigate these effects, we plot the actual NDCG-quality values of the rerankings done by different methods on the Boston dataset under the Singular attention model in Figure 4.17. The results confirm that the actual loss is often lower than the threshold enforced by the ILP. Observe that

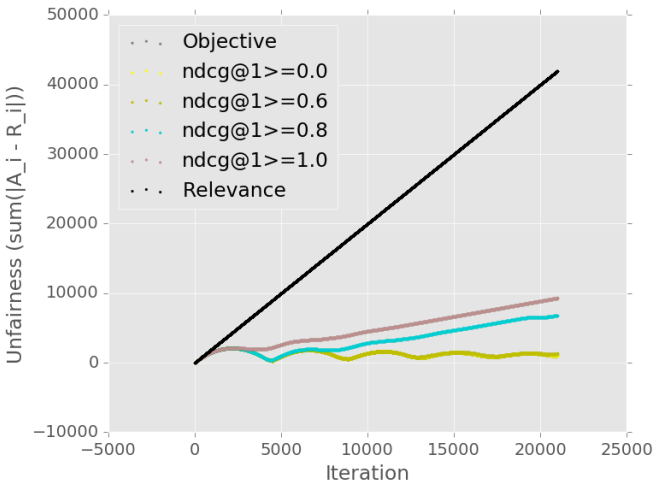

Figure 4.15: Model performance on the multi-query Hong Kong dataset. Attention singular.

NDCG-quality is 1 for a number of initial iterations in all the methods. This is where reshuffling of the top ties happens. The quality starts decreasing as less relevant subjects gather enough deserved attention, and periodically goes back to 1, when the top-relevant subjects gain priority again. Similar conclusions regarding the absolute loss hold under the general geometric attention model.

Note that without explicit control, the results with lower utility could be consistently delivered to the same users, leading to unfairness in the search quality for searchers. Mitigating this problem would require a two-sided fairness model for searched subjects and searchers.

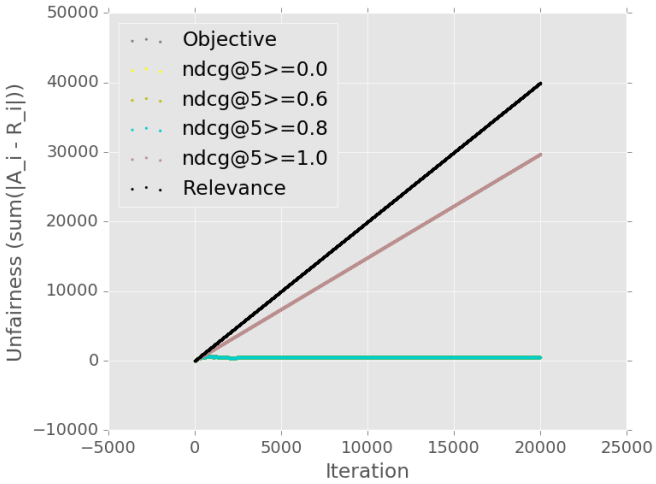

Figure 4.16: Model performance on the single-query Boston dataset. Attention geometric. Results are similar for the Geneva and Hong Kong datasets.

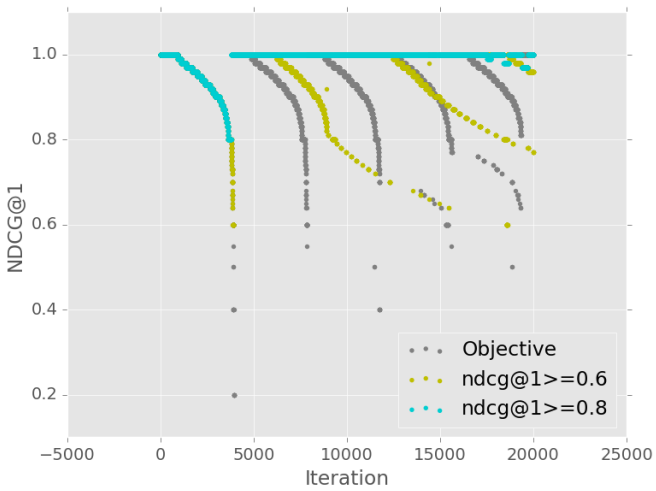

Figure 4.17: Actual values of ranking quality. Boston dataset, attention singular.

4.4.8 Performance on StackExchange data

The relative trends in the performance of our method are the same here as in the results for other datasets. One of the characteristics that distinguish the StackExchange dataset is that each individual subject occurs in relatively few rankings. An observation that follows is that longer amortization timeframe is necessary under such conditions - a subject obviously needs to appear in a number of rankings so that the model can reposition them to fairly distribute attention.

4.5 Related work

Fairness. The growing ubiquity of data-driven learning models in algorithmic decision-making has recently boosted concerns about the issues of fairness and bias. The problem of discrimination in data mining and machine learning has been studied for a number of years [Pedreschi et al. 2008; Kamishima et al. 2012; Romei and Ruggieri 2014]. The goal there is to analyze and counter data bias and unfair decisions that may lead to discrimination. Much prior work has centered around various notions of group fairness: preserving certain ratios of members of protected vs. unprotected groups in the decision making outcomes, with the groups derived from discrimination-prone attributes like gender, race, nationality, etc. [Feldman et al. 2015; Hardt et al. 2016]. For example, the criterion of statistical parity requires that a classifier's outcomes do not depend on the membership in the protected group. State-of-the-art mechanisms for dealing with such group fairness requirements are to solve constrained optimization, e.g. maximize prediction accuracy subject to certain bounds on group membership in the output labels. This has led to classification models with fairness-aware regularization (e.g., [Zafar et al. 2017]). Beyond the fairness of outcomes, researchers have looked into the fairness of process in the decision-making systems [Grgic-Hlaca et al. 2018b].

Individual fairness [Dwork et al. 2012] requires that individual subjects who have similar attributes should, with high probability, receive the same prediction outcomes. Literature to this end has so far focused on classification and selection problems [Zemel et al. 2013; Kearns et al. 2017].

Other lines of work investigate mechanisms for fair division of resources [Abebe et al. 2017], or how automated systems can assist humans in decision making [Kleinberg et al. 2017a].

Fairness in rankings. Prior work on fair rankings is scarce and recent. Some proposals show how to incorporate various notions of group fairness into ranking quality measures [Yang and Stoyanovich 2017]. There have been approaches that diversify the ranking results in terms of presence of members of different groups in ranking prefixes, at the same time keeping the ranking quality high [Zehlike et al. 2017]. This problem has also been studied from a theoretical perspective with the results provided for the computational complexity of the problem [Celis et al. 2018]. All of these approaches consider static rankings only, and all focus on group fairness. Parallel with our work, Singh and Joachims [2018] have proposed a notion of group fairness based on equality of exposure for demographic groups. While

technically complementary and similar in spirit to our approach, this method is also geared for a purpose different than individual fairness, and does not aim at binding attention to relevance.

Bias in IR. The existence of position bias in rankings of search results has been revealed by a number of eye-tracking and other empirical studies [Craswell et al. 2008; Dupret and Piwowarski 2008; Guo et al. 2009]. Top-ranked answers have a much higher probability of being viewed and clicked than those at lower ranks. The effect persist even if the elements at different ranks are randomly permuted [Joachims and Radlinski 2007]. These observations have led to a variety of click models ([Chuklin et al. 2015] provide a comprehensive survey), and several methods for bias-aware re-ranking [Wang et al. 2016; Joachims et al. 2017]. However, position bias has been primarily studied in the context of document ranking and no prior work has investigated the influence of the bias on the fairness of ranked results. A large search engine has been investigated for presence of differential quality of results across demographic groups [Mehrotra et al. 2017]. Similar studies have been carried out on other kinds of tasks such as credit worthiness or recidivism prediction [Adler et al. 2018].

Relation to other models. Fairness dimension has been considered for job dispatching at the OS level, for packet-level network flows [Ghodsi et al. 2012], for production planning in factories [Ghodsi et al. 2011], and even for two-sided matchmaking in call centers [Armony and Ward 2010]. Fairness understood as envy-freeness is also investigated in computational advertising, including generalized second-price auctions [Edelman et al. 2007]. In the context of rankings, a potential connection between fair rankings and fair queuing has recently been suggested [Chakraborty et al. 2017].

4.6 Conclusion

This thesis argues for equity of attention – a new notion of fairness in rankings, which requires that the attention ranked subjects receive from searchers is proportional to their relevance. As this definition cannot be satisfied in a single ranking because of the position bias, we propose to amortize fairness over time by reordering consecutive rankings, and formulate a constrained optimization problem which achieves this goal.

Our experimental study using real-world data shows that the discrepancy between the attention received from searchers and the deserved attention can be substantial, and that many subjects have equal relevance scores. These observations suggest that improving equity of attention is crucial and can often be done without sacrificing much quality in the rankings. Incorporating such fairness mechanisms is especially important on sharing economy or two-sided market platforms where rankings influence people's economic livelihood, and our work addresses this gap.

Equity of attention opens a number of interesting directions for future work, including calibration of ranker scores in economically-themed applications, all the way down the IR stack to properly training judges to provide relevance labels with fairness in mind.

Chapter 5

Sensitive Search Exposure

Contents

Search engines in online communities such as Twitter or Facebook not only return matching posts, but also provide links to the profiles of the authors. Thus, when a user appears in the top-k results for a sensitive keyword query, she becomes widely exposed in a sensitive context. The effects of such exposure can result in a serious privacy violation, ranging from embarrassment all the way to becoming a victim of organizational discrimination.

In this chapter, we propose the first model for quantifying search exposure on the service provider side, casting it into a reverse k-nearest-neighbor problem. Moreover, since a single user can be exposed by a large number of queries, we also devise a learning-to-rank method for identifying the most critical queries and thus making the warnings user-friendly. We develop efficient algorithms, and present experiments with a large number of user profiles from Twitter that demonstrate the practical viability and effectiveness of our framework.

5.1 Introduction

Motivation. A query search engine in online communities, such as Twitter or Facebook, not only returns matching posts, but also identifies the users who have written the posts. This *search exposure risk* is particularly pronounced when a user's post appears in the top-k results for a sensitive keyword query.

Note that exposure is different from just having contents visible within a community. When Facebook introduced the *News Feed* feature, a lot of users responded with outrage. They felt their privacy was being violated, even though the new feature only meant that newly generated content would be broadcasted to people who would have access to that content anyway [Boyd 2008]. Analogously, in the context of search systems, while a user may be fine with posting about health problems, controversial political issues or using swearwords, she may feel very uncomfortable with the posts being returned as top-ranked results. Content found this way could be used, for example, in stories written by journalists or bloggers, and attract uninvited attention to the user's account. Beyond topically sensitive queries, there are also risks regarding search exposure by unique strings. An adversary could search for people posting urls of sensitive domains, such as pirate websites, or certain price tokens, such as $\$1K$. An adversary with a list of e-mails could issue these to find answers to security questions necessary to reset passwords. An adversary with a list of generated credit card numbers could issue these as queries to find other personal information necessary for credit card transactions.

State of the art and limitations. Despite the existence of such threats, to the best of our knowledge, there is no support for users to find out about their search exposure risks. The only way would be to try out all possible queries and inspect their top-k results, yet this is all but practical. The service providers – search engines or social network platforms – do not provide such support at all.

Work in the broad area of exposure has been tangibly motivated by a study showing the discrepancy between the expected and actual audience of user-generated content [Bernstein et al. 2013]. Exposure has been addressed in other contexts so far, including information exposure among friends in social networks [Mondal et al. 2014], location exposure [Shokri et al. 2011], longitudinal information exposure [Mondal et al. 2016], controlled information sharing [Schlegel et al. 2011], or exposure with respect to sensitive topics [Biega et al. 2016]. The importance of exposure control has led service providers to introduce features such as Facebook's *View As*, which informs a user how her profile appears to other people. However, this does not quantify the exposure, and the problem of search exposure in particular has been disregarded completely.

Problem and challenges. To the best of our knowledge, this dissertation is the first to address the problem of modeling, analyzing and quantifying search exposure risks. As the risk is most significant when a user is spotted in the top-k results of a query, our goal is to identify these top-k exposing queries for each user. Such information can then be used to guide the user, for example, in deleting posts or restricting their visibility. In an online

setting, a tool based on our model could even alert the user about the exposure before submitting a post.
The search exposure problem poses a number of challenges:

- *Efficiency:* A user could possibly be found by millions of distinct queries. An algorithm to identify the critical queries thus faces a huge scalability and efficiency problem.
- *Dynamics:* With the high rate of new online contents, the critical queries cannot simply be computed offline from a query log. The exposure of users keeps continuously shifting.
- *Usability:* Showing all queries for which a user appears in the top-k results would in many cases flood the user with millions of irrelevant or small-risk queries and miss the point of guiding the user. Thus, it is crucial that the queries are ranked by an informative measure of, possibly user-specific, sensitivity.

An interesting thing to note is that from the perspective of a user, reducing search exposure can be seen as a problem of *"inverse search engine optimization"*, inverse SEO for short. SEO aims to push a user to the top-ranked results for certain queries. Here, the goal is the opposite – the users would like to be moved to the low-ranked tail of answers, or even completely removed from the search results of particularly sensitive queries.

Approach and contributions. We model search exposure as a problem of reverse search – instead of starting with a query and finding top-k documents relevant to the query, we start with a document and want to find all the queries that return the document in the top-k results. If we then think of keyword search with answer ranking as the problem of finding the top-k nearest-neighbor posts according to a given similarity function, search exposure becomes a *reverse k-nearest-neighbor* problem (RkNN).

To assist a user in understanding her search exposure risks, we devise an algorithm for ranking the queries in the user's RkNN set, which potentially contains hundreds of queries. To this end, we combine informative features ranging from topical sensitivity (e.g., usually higher for queries about health problems than for those about movies), through query selectivity and entropy (e.g., higher for queries containing birth dates, or social security numbers), to user surprisal (e.g., high for queries matching a post about a user's children in an otherwise professional profile). The salient contributions of this chapter are:

- A model of the search exposure problem;
- A learning to rank method with informative features for ranking the queries in the exposure sets according to a new notion of search exposure relevance;
- An experimental study with a large set of Twitter profiles, providing insights on the exposure sets and the effectiveness of our query ranking methods.

5.2 Problem statement

Preliminaries. Assume we have a set of users U and a set of documents D posted by the users. We denote the fact that a post d is written by the user u by $d \in u$. The profile of

each user is defined as the set of all documents she has posted in a community.

Search exposure. The problem of search exposure of user u can be formalized as finding all the reverse k-nearest neighbors of u, i.e, the set of all the queries for which any of the posts of u comes among the top-k results. We call the sets of such queries as *exposure sets*.

Generation of exposure sets. Before defining the exposure sets of all users, we first define $RkNNs(d)$ for each document d as follows:

$$RkNNs(d) = \{(q, r) | q \in Q \wedge d \text{ is the } r\text{th NN of q} \wedge r \leq k\} \tag{5.1}$$

where Q is the set of all queries and r is the rank of d for the query q ($r = rank(q, d)$). According to Emrich et al., the above equation is equivalent to the definition of a bichromatic RkNN [Emrich et al. 2015].

Accordingly, we define the exposure set of each user as the union of the exposure sets of all the documents in her profile. We denote the exposure set of the user u by $RkNNs(u)$ which is defined as follows:

$$RkNNs(u) = \{(d, q, r) | (q, r) \in RkNNs(d) \wedge d \in u\} \tag{5.2}$$

An efficient algorithm for generating exposure sets developed by Biega et al. [2017a] is not a contribution of this thesis.

Ranking of queries in exposure sets. Exposure sets of certain users might be big and dominated by rare, non-informative, or non-critical queries. On the other hand, exposure by certain sensitive queries might leave the user uncomfortable. Therefore, to make the exposure sets user-friendly, we want to rank the triples in $RkNNs(u)$ such that *the queries the users would not want to be searched by appear at the top*. This defines the notion of relevance in our ranking problem, termed *search exposure relevance*. We discuss the exposure set ranking methods and our notion of relevance in Sec. 5.4.

5.3 Generating exposure sets

Biega et al. [2017a] develop an efficient algorithm for computing exposure sets defined by Equation 5.2 under the assumption that the search engine uses a ranking mechanism based on language models. This algorithm is not a contribution of this thesis. Instead, to present the further contributions of this chapter, we assume we have the exposure sets (Equation 5.2) computed for all the users in a system. Since considering all possible queries makes the problem intractable, we similarly limit the considered queries to: (i) unigram and bigram queries, (ii) and only those queries for which there exists at least one document in the underlying collection containing all the query terms.

Note that such exposure sets can be generated by the algorithm proposed by Biega et al. [2017a], or in a brute-force manner, by first computing top-k rankings for all considered queries and then re-aggregating the results.

aim oshtitsbaj, asleep oshtitsbaj, http://ra*.com/teh_ba splash, mac vanilla, suck wake, mood sick
emma sun, watch xxxxxx, forget toast, @jeff* fall, heavyweight ladder, omg tan, alcohol nice
blown death, comin lake, parilla wait, bathroom wanna, crush hannah, friend lord, record woman

Table 5.1: Example queries from unprocessed exposure sets.

5.4 Ranking of queries in exposure sets

Generating the exposure sets is not enough for the results to be presentable to end users for two reasons. First of all, for many users the size of their $RkNN$ set is simply too big for easy consumption. Our experiments on a sample of $50K$ user profiles from Twitter later confirm this – even when only unigram and bigram queries are considered, more than $35K$ users are exposed by more than 100 queries, with some users exposed by millions. Figure 5.1 in the upcoming experimental section of this chapter shows the distribution of exposure set size for users from the sample.

Moreover, since we do not a priori exclude queries such as infrequent or numerical tokens most RkNN sets will end up dominated by garbage queries. Leaving such queries in during the generation phase is a design choice motivated by the 'worst case scenario' principle that often guides privacy and security research. While most users will find these queries uninformative, for some people it might be important to know they are searchable by certain URLs (e.g., when the domain is known to contain sensitive content) or numbers (e.g., their year of birth or the prices of products they buy). Table 5.1 shows examples of the top queries in the raw exposure sets where queries are ordered by the rank position of the corresponding user post. These examples illustrate the need for ranking the queries before presentation to end users – raw sets are uninformative when mostly garbage queries are shown to the users first.

5.4.1 Learning to rank the exposing queries

Recall the search exposure sets defined by Eq. 5.2. We want to rank the triples within these sets according to search exposure relevance, i.e., such that the queries the users would not want to be searched by appear at the top. The traditional IR learning to rank setup, in which the learned function orders the documents by relevance to queries, is replaced by one where we rank the queries according to relevance to users.

Each user-document-query triple can be represented as a feature vector $\Phi(u, d, q)$. For each user, together with the relevance score annotations, these form partial rankings determining pairwise relevance constraints between the data points (e.g., for a user u, an exposing query q_1 matching a document d_1 should be ranked higher than the query q_2 matching a document d_2.) We want to learn a ranking function that minimizes a loss measure over these partial training rankings. For example, when learning to rank using SVM^{rank} [Joachims 2006], it is the number of violated pairwise constraints that is minimized, which implicitly leads to

maximization of Kendall's τ between the golden and learned rankings.

We describe the features and relevance scores we used to learn the ranking function in the following two sections.

5.4.2 Features

5.4.2.1 Semantic features

The meaning of words plays an important role in determining criticality of search exposure. In a similar context, user studies have shown topical sensitivity to be useful in the context of privacy risk quantification from text [Biega et al. 2016]. To capture the coarse-grained semantics of the queries, we annotate them with categories from the LIWC dictionaries [Tausczik and Pennebaker 2010]. LIWC categorizes words into 80 linguistically and psychologically meaningful categories such as *positive emotion* (love, nice, sweet), *affective processes* (happy, cry, abandon), *swear words* (damn, piss, fuck), *anxiety* (worried, fearful, nervous), or *sexual* (horny, love, incest). We create one binary feature based on each category, with a value of 1 if any of the query words matches any of the words from the category.

5.4.2.2 Uniqueness of queries

While any query generated from a community's text contents search-exposes some of its members, from the perspective of a single user, these are the rare tokens that are more likely to lead to exposure. While a considerable portion of rare queries are simply meaningless noise, it is possible that there are meaningful infrequent tokens with the potential to violate privacy. Recall some of our motivating examples where an adversary searches for information associated with a given sensitive domain, or an e-mail address.

We propose two features to capture how rare a query is: query selectivity and query entropy. We define the query selectivity as the number of documents matching the query exactly:

$$selectivity(q) = |\{d : q \in d\}| \tag{5.3}$$

This measure will be low for queries which appear infrequently.

Another aspect of a query being unique is how skewed the distribution of the relevance scores is. We capture this by measuring the entropy over the distribution of ranking scores of the top-k returned results. Let R be the distribution of the relevance scores of the top-k results. We measure the entropy of the query as:

$$entropy(q) = H(R), \text{ where } R(i) = \frac{score(q, d_i)}{\sum_j^k score(q, d_j)} \tag{5.4}$$

Note that these measures are not dependent on a given user, but are dependent on the community as a whole, i.e., the relative rankings of queries in different communities might differ. For instance, while the query *Lyme borreliosis* might be an infrequent query on Twitter, it could be more popular in a medical Q&A forum.

5.4.2.3 User surprisal

The lexical context of a user might also matter when determining the criticality of a query. Imagine a user with a Twitter profile where she posts mostly professional content. It would not be surprising, and perhaps even desirable, that the user's posts are returned as top results to the queries from that professional domain. However, if it turns out that the user profile comes up at the top only to the query *funny cats* that matches that single post the user has ever made outside of the professional domain, this might be both unexpected and undesirable.

We propose to capture this intuition using surprisal, which is measured by reversing the probability of the query being generated from a user's vocabulary distribution estimated from the posts:

$$surprisal(q,u) = log\left(\frac{1}{P(q|u)}\right) = log\left(\frac{1}{\prod_{w\in q} P(w|u)}\right) \tag{5.5}$$

To account for the sparsity of user profiles, we compute these probabilities using Dirichlet smoothing.

5.4.2.4 Document surprisal

Even though these are the queries that are ranked, the users might not want to be matched to a non-critical query when it exposes a critical post. Similarly to surprisal of queries, we define the surprisal of posts that are matched by the exposing queries by replacing q by d in Eq. 5.5.

5.4.2.5 RkNN features

Two traditional methods for ranking the reverse nearest neighbors by relevance to the user are the proximity of the reverse neighbor to the user and the rank of the reverse neighbor. While not likely to be useful when the relevance is defined as criticality, we include these features for comparison. We measure proximity using the probability of generating the query from the posting history of u:

$$proximity(q,u) = log\left(P(q|u)\right) \tag{5.6}$$

Let d_u be the post of a user u that is returned as an answer to query q at position $rank(q,d_u)$:

$$rankposition(q,u) = rank(q,d_u) \tag{5.7}$$

5.4.2.6 Syntactic features

We also introduce a number of binary post-dependent features that characterize emotional display or content the users might not want to be exposed by through search. These include:

- *has_url* (set to 1 if the post contains a URL),
- *has_at_mention* (set to 1 if the post mentions another user),

- *has_hashtag* (set to 1 of the query contains a hashtag).
- *has_emoticon* (set to 1 if the post contains an emoticon),
- *has_repeated_punctuation* (set to 1 is any token in the post ends with a double exclamation mark, double question mark, or an ellipsis),
- *has_repeated_vowels* (set to 1 if any token in a post contains a vowel repeated at least three times in a row),
- *has_laughter* (set to 1 if any token contains a substring like *haha*, with different vowels).

5.4.3 Relevance

Search exposure relevance differs in many ways from the topical relevance of traditional IR tasks. A query might be relevant not only because it's topically sensitive, but also because it could embarrass, offend, or otherwise violate the privacy of the exposed person. The subjective nature of such judgments makes the manual collection of relevance at scale an extremely time-consuming task, especially if done by external evaluators. To decide which queries would be relevant, a judge would have to put themselves in the shoes of the evaluated user, imagine who that person is based on the contents of the profile, and decide which queries would concern her. Moreover, a judge would have to come up with likely threat scenarios. It is a non-trivial task to prime the judges regarding these issues without biasing them. With all these considerations, we derive implicit relevance scores from other user-generated signals that indicate reluctance to be associated with a given content. Implicit relevance signals, especially in the form of clickthrough patterns, are commonly used in traditional retrieval tasks [Carterette and Jones 2007]. The remainder of this section presents our method for synthesizing the search exposure relevance scores.

User score. If a user deletes a post, it is a signal she does not want to be associated with its content. Thus, a query matching a post that got deleted after publication receives a user score of 1, whereas a query matching a non-deleted post receives a user score of 0. While a service provider quantifying exposure would have a direct access to this information, there are also ways for collecting it outside of the system [Mondal et al. 2016]. We describe our collection method in more detail in the experimental section.

Community score. The deletion information is a noisy signal, however, as users delete posts for a variety of reasons, including language or double posting errors. We want to sanitize these scores using stronger, community-wide signals that encode the differences in language distributions in anonymous and non-anonymous communities. These linguistic differences have been observed, for instance, when comparing posts from Twitter and Whisper (an anonymous microposting platform) [Correa et al. 2015]. Having estimated the vocabulary distributions in an anonymous (P_{anon}) and a non-anonymous ($P_{non-anon}$) community, we treat the relative probability of a query being generated from these distributions as a

community-wide signal that users do not want to be associated with the keywords. More precisely, we set the community score of a query to:

$$community\_score(q) = \frac{P_{anon}(q)}{P_{non-anon}(q)} = \prod_{w \in q} \frac{P_{anon}(w)}{P_{non-anon}(w)} \tag{5.8}$$

Golden ranking. Finally, we derive the relevance as a linear combination of both scores:

$$score(q) = \alpha \cdot user\_score(q) + (1 - \alpha) \cdot community\_score(q) \tag{5.9}$$

Combining both scores allows us to discount the relevance of noisy queries that match deleted posts, as well as add relevance to sensitive queries matching posts that did not get deleted, as the user perhaps did not have any privacy concerns in mind.

5.5 Experiments

In this section, we discuss our insights into the search exposure problem through evaluation of the ranking methods, as well as an analysis of user perceptions regarding exposing queries collected in an experiment on Amazon Mechanical Turk.

5.5.1 Dataset

For our experiments, we use a sample of Twitter profiles from the longitudinal exposure study by Mondal et al. [2016]. It consists of 51,550 user profiles with a total of about 5.5 million tweets posted over the year 2009.

5.5.2 RkNN generation

The experiments on RkNN Generation reported by Biega et al. [2017a] are not a part of this thesis. However, we use the exposure sets the authors generated as an input for the ranking algorithm experiments presented in this thesis. We thus report on the data preparation techniques applied before generating the exposure sets. The cleaning included stop words removal, lemmatization and stemming, resulting in around 2 millions unique tokens. The query filtering strategy described in Section 5.3 resulted in 45 million queries considered for exposure sets.

Figure 5.1 shows the distribution of the size of exposure sets of the users in the dataset for different values of k. We assume $k = 10$ for the experiments in this chapter.

5.5.3 Query ranking in exposure sets

5.5.3.1 Exposure sets cleaning

For the evaluation results to be meaningful, we excluded the following queries from the exposure sets: queries with tokens shorter than 3 letters, queries for which none of the tokens is an English word, queries with numerical tokens, urls, and references to other accounts. We also excluded users whose posts are primarily written in a language other than English. While all of these queries could be search exposure relevant in certain contexts, it is unlikely

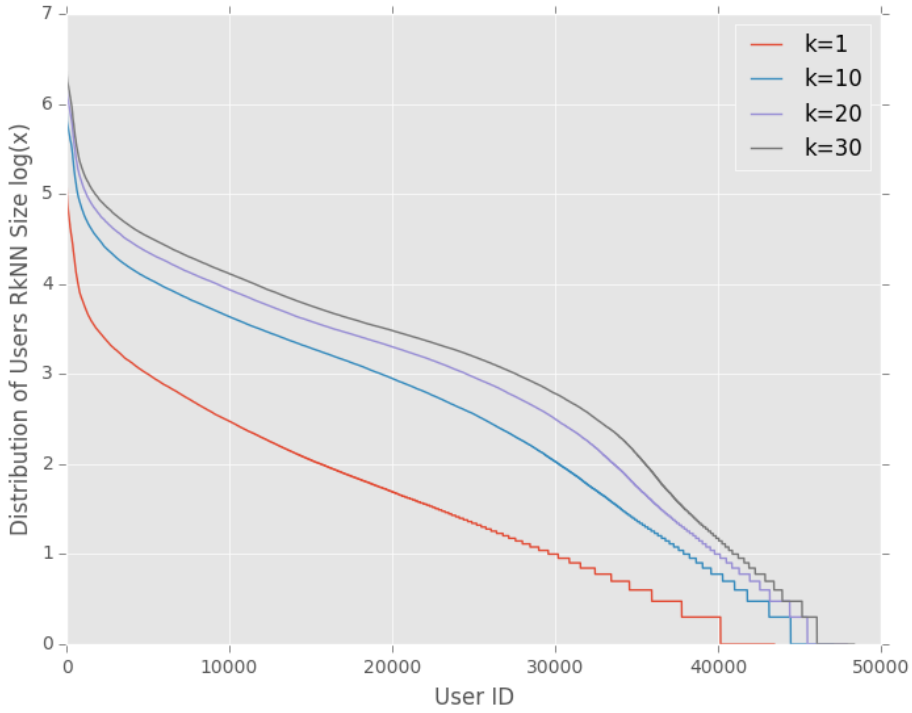

Figure 5.1: Distribution of the size of exposure sets. The values on Y axis are in logarithmic scale with base 10.

that human judges who evaluate the ranking outputs would be able to associate any meaning with them.

5.5.3.2 Relevance score statistics

We construct the relevance scores as described in Sec. 5.4.3. Community scores are derived from the Whisper dataset collected by Correa et al. [2015] and the Twitter dataset collected by Mondal et al. [2016]. The Twitter dataset, moreover, comes with the information regarding tweet deletion. More precisely, by querying the Twitter API using a subset of the tweet IDs, the authors were able to determine which tweets got deleted after publication. This information was collected for 11M tweets, 400K of which turned out to have been removed. We use these signals as the user score.

Because the information about post deletion is limited, the ground truth provides us with only a partial ranking over RkNN queries. We therefore exclude the queries for which we cannot infer relevance from the evaluation in this part. These include: (i) the queries matching posts for which we do not have the deletion information, (ii) the queries with only a partial overlap with the source post (it might happen that a post is returned in the top-k results for a query even though not all query words appear in the post; for such queries we do not assume the deletion information signals not wanting to be associated with the words). Excluding exposure sets with less than 30 queries, which do not need ranking to be

Feature	Example queries
Sexual	gross kiss, breast whitney, gay shirt
Humans	dumbest guy, girl xoxo, chess kid
Friend	pal wife, fellow fool, ummm honey
Anger	buy weapon, mad scientist, idiot vegetarian

Table 5.2: Most important semantic features learned by the L2R model together with example queries.

presentable to users, we are left with around $15K$ profiles under evaluation.

5.5.3.3 Ranking algorithm

To learn the ranking function, we use SVM^{rank} [Joachims 2006] with the linear kernel. Parameter C is tuned on a random sample of 10% RkNN sets, and the rest of the data is used to evaluate the L2R method in a 10-fold cross-validation.

5.5.3.4 Feature analysis

The weights of the decision boundary vector learned using SVM with a linear kernel can be interpreted as feature importance weights. Table 5.2 lists the most important features learned by our model together with example queries exhibiting the features. The model captures well that the categories related to personal issues are the ones people feel more uncomfortable sharing. High importance of words related to sexuality stems from the bias of the Whisper data – a large majority of anonymous posts from this community regard sexuality. However, the methodology we propose is general enough to handle different types of anonymous contents. For instance, as an alternative, it would be possible to collect anonymous posts from more general Question & Answer communities such as Quora.

5.5.4 User-study evaluation

Because the relevance scores used for training the algorithm constitute noisy signals for search exposure relevance, we evaluate the reranked exposure sets in a user study. The leading question is whether users themselves would find the output useful, feeling that exposure by top-ranked keywords would make them feel uncomfortable. This section provides the details of the study.

5.5.4.1 Evaluation setup

To evaluate the rankings, we sample a number of exposure sets and a number of queries from each.

User sampling. The first important thing to note is that not all of the exposure sets contain sensitive queries. To account for this and make sure we cover the sensitive users in the evaluation, we sample users non-uniformly in the following way. Queries within exposure sets are ordered by the predicted relevance scores. The score of the highest ranking query within a set can be thought of as an indicator of how sensitive the exposure set is overall (i.e.,

the lower the highest score, the less sensitive content there is overall). For evaluation, we choose the 50 most sensitive exposure sets, and 50 exposure sets sampled from the remaining tail with the probability proportional to the predicted relevance of the highest scoring query. We thus evaluate 100 exposure sets in total.

Query sampling. To construct assignments with reasonable workloads, we evaluate 50 queries from each of the sampled exposure sets. Having the queries ranked by the L2R method under evaluation, we choose 25 highest scoring queries (to see how useful the top of the ranking is), and 25 queries chosen uniformly randomly from the remaining tail (to control if the head of the ranking does not miss critical queries).

5.5.4.2 AMT survey

Each set of 50 sampled queries was shown to 3 Amazon Mechanical Turk workers. The queries were displayed in a random order. We required that the workers have a master qualification (to ensure the quality of annotations) and are located in the USA (to prevent language misunderstanding). Upon explaining the basic pipeline of the Twitter search engine and priming the users about what exposure is, the survey asked the following question:

> *Would you feel concerned (uncomfortable, embarrassed, privacy-violated, or threatened) if your tweet was returned as one of the top answers to these search terms? (Yes/No)*

Having three people evaluate each query leads to a 4-graded (0..3) relevance scale, based on how many people chose *Yes*.

Out of $5K$ evaluated queries, 10% had a score of 3, 12% had a score of 2, 24% had a score of 1, 54% had a score of 0. Inter-annotator agreement measured by Fleiss' κ was 0.376, which corresponds to a fair agreement.

5.5.4.3 Results

We report the values for NDCG@[5,10,20] and Kendall's τ. Moreover, since the collected scores offer good interpretability in terms of binary relevance as well, we also report Precision@[5,10,20], assuming a query is search exposure relevant if it was marked by at least one judge.

Table 5.3 shows the results of the user-study evaluation. Note that, although the queries were sampled from the L2R-ranked exposure sets, the collected judgments also let us evaluate other ranking heuristics. We use the rankings based on the values of several high-level features as baselines. Majority of these perform significantly worse than the L2R method – differences significant by a paired t-test with $p < 0.05$ are marked with the $*$ symbol. The strongest heuristics include document surprisal and selectivity. Both of these quantities capture a different aspect of the rareness of the content, and thus shine in situations where, for instance, the judges thought that exposure by a typo might lead to embarrassment. We also observed that a number of query tokens are typos that can be mapped to a sensitive word. Such queries were often marked by the judges as relevant, and because of their rareness, heuristics such as selectivity gain in performance.

	Prec@5	**Prec@10**	**Prec@20**	**NDCG@5**	**NDCG@10**	**NDCG@20**	**Kendall's τ**
L2R	**0.636**	**0.566**	**0.509**	**0.515**	**0.496**	**0.530**	**0.107**
Surprisal	0.448*	0.449*	0.447*	0.210*	0.245*	0.316*	-0.035*
Document Surprisal	0.480*	0.471*	0.470*	0.229*	0.262*	0.350*	0.016*
Entropy	0.472*	0.489*	0.494	0.209*	0.262*	0.347*	0.026*
Selectivity	0.548*	0.508*	0.489	0.278*	0.305*	0.378*	0.037*
Rank	0.460*	0.463*	0.466*	0.204*	0.248*	0.330*	0.018*

Table 5.3: Exposure set ranking user-study results averaged over all users. Methods marked with * perform significantly worse than L2R on a given metric (paired t-test, $p < 0.05$).

5.5.4.4 Anecdotal examples of sensitive exposure sets

Table 5.4 presents examples of exposure sets with the top-10 queries ranked by the L2R method and is meant as an overview of the types of sensitive keywords a user might be exposed by in Twitter. Queries were generated from the contents of user posts, which means that each presented word combination matches at least one post in our sample. We resort to showing a manually chosen subset of examples, as the top sensitive exposure sets were highly explicit and offensive.

5.6 Insights into search exposure relevance

5.6.1 Tweet context

An interesting question regarding search exposure relevance is whether it is influenced by the context of the returned tweet. It might happen that a query that looks sensitive is constructed from words that do not form a coherent context within a post, thus being a false alarm. On the other hand, innocent looking queries might bring out posts that do contain sensitive content otherwise.

To gain preliminary insight into this problem, we conducted a second survey on AMT, in which the workers assessed the relevance of queries, also knowing the tweet that is being returned as a result; the rest of the setup remained analogous. Comparison of these two surveys is summarized in Figure 5.2. Existence of dark squares outside of the diagonal suggests indeed that the context might change the exposure relevance judgement. This happens both ways, suggesting that both scenarios we mentioned in the previous paragraph are plausible. We believe that investigating the factors that influence the search exposure relevance is an interesting topic for future work.

5.6.2 Search exposure relevance vs topical sensitivity

Topical sensitivity is a concept introduced for studying privacy risks of text, in particular for quantifying R-Susceptibility (Rank-Susceptibility) in communities where user profiles consist of textual contents [Biega et al. 2016]. It measures how likely the presence of words from different topics (understood as distributions over words) leads to privacy risks, irrespective of the user or community context. We want to understand if there is a correlation between topical sensitivity defined this way and the search exposure relevance. We thus annotate each query from our evaluation set using the topical sensitivity annotations $sensitivity(t)$ collected in the R-Susceptibility paper [Biega et al. 2016]. We define the sensitivity of a query as:

$$sensitivity(q) = \frac{1}{|q|} \sum_{w \in q} \sum_{t} sensitivity(t) \cdot P(w|t) \tag{5.10}$$

where $P(w|t)$ is the probability of a word w in the topic t.

We measure the correlation between these sensitivity-annotations and the collected relevance scores using the Pearson correlation coefficient. We find a strong correlation between these scores in case of the relevance collected for queries without the tweet context

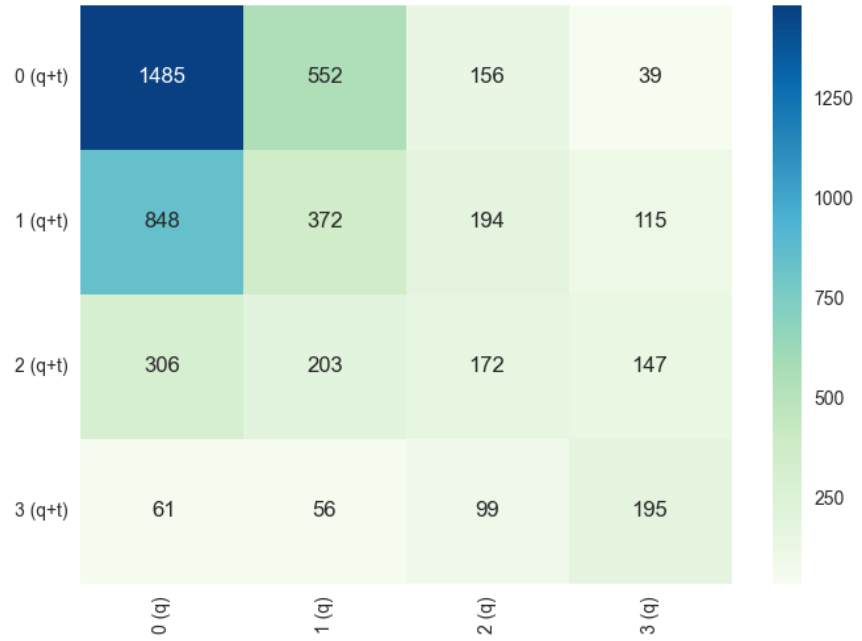

Figure 5.2: Influence of the tweet context on search exposure relevance. The number in a square $x(q)$, $y(q+t)$ denotes the number of tweets that received the score of x in the study with queries only, and the score of y in the study with queries in context.

(Pearson coefficient of 0.44), and a little lower correlation (Pearson coefficient of 0.32) in case of the relevance judgments for queries with the tweet context. This result reconfirms the findings from the evaluation of the L2R method – the meaning of the query is an important factor in determining search exposure relevance, and topical sensitivity is a viable alternative for implicit relevance scores.

5.7 Related work

Exposure. Although, to the best of our knowledge, the problem of search exposure has not been addressed in the past, there are different aspects of user and data exposure that have been studied in the prior literature. Mondal et al. [2016] proposed exposure control as an alternative solution to access control in social networks [Mondal et al. 2014], and later devised solutions for longitudinal exposure control. [Biega et al. 2016] quantify privacy risks for sensitive topics in rankings based on textual posts using the notion of R-Susceptibility. Exposure has also been studied in the context of individual attribute leakage, such as location [Shokri et al. 2011]. Another interesting problem is that of usability of exposure warnings. Example solutions include depicting the current size of content audience by the size of a displayed pair of eyes [Schlegel et al. 2011].

blame gay, dutch gay, gay rabo, gay guy, blame dutch, suck teacher, start tht, rider tour, donald duck, attack bad
gay racist, fuckin young, deal gay, simon watchin, fuckin kinda, guy note, kind sum, live net, bcoz gay, dnt season
adopt convert, adopt religion, convert religion, convert essay, essay religion, bon convert, bon religion, bon essay, river tonight, adult love
lesbian pregnant, lesbian music, lesbian live, boat lesbian, gay norway, end lesbian, fritt gay, bell page, star trend, lesbian uuum
oooh virgin, virgin wen, video virgin, crack oooh, oooh wen, normal tom, normal smith, smith tom, swine year, outfit xoxo
gay israel, bit web, gay gunman, michael pant, hat pant, attack bit, e-mail match, israel wtf, china tale, obama recov
david queer, queer ted, asian queer, queer warhol, david folk, model race, keith york, driver rule, kind remix, jean odd
camera stick, stick tape, rep usa, china rep, stick tehran, governor tehran, governor stick, prayer tehran, prayer tehran, israel rep
detail obama, alex detail, alex obama, box long, bloomberg flash, dubai investor, dubai investor, june real, june real, alybi*@gmail.com investor
u.s. union, mexico union, canada union, american union, demand democrat, agenda reform, nasa obama, american borderless, nazi obama, demand overhaul

Table 5.4: Top-10 sensitive exposing queries returned by the L2R model for a subset of users.

Privacy-preserving IR. Problems studied in privacy-preserving IR include sanitization of query logs prior to a release [Götz et al. 2012; Zhang et al. 2016a], or obfuscation of query histories through broadened or dummy queries [Gervais et al. 2014; Wang and Ravishankar 2014]. A number of works also investigate the viability of personalized search under privacy constraints [Chen et al. 2011; Shen et al. 2007; Xu et al. 2007; Zhu et al. 2010].

User protection and internal audits. Service providers increasingly come under close scrutiny by external organizations and observers, including journalists and researchers. This pressure encourages the SPs to perform proactive, internal audits to improve their services and infrastructure. New solutions for increased privacy are constantly introduced to mitigate the threats for users from external adversaries in services like maps [Huang et al. 2017]. User data itself has also been analyzed, for example, to deliver better security protections in the context of account recovery personal questions [Bonneau et al. 2015]. Beyond privacy, there are also other societal issues that press SPs to audit their services, including the issues of fairness and bias [Feldman et al. 2015], or user satisfaction with search results [Mehrotra et al. 2017].

Search exposure can be seen as another dimension for internal audits. Along these lines, we believe more work can be done to examine which types of search queries should be blocked altogether, and which search results should be removed to protect against finding users in

sensitive contexts. While certain ad-hoc protections are already in place (for instance, it seems impossible to explicitly query for credit card numbers in Google, Twitter, or Facebook, since these tokens get post-processed and end up matching other numerical tokens as well), there is a need for a more direct examination and protection mechanisms regarding the exposure of users in search systems.

(Reverse) k-nearest-neighbors problems. The problem of finding reverse k nearest neighbors has been studied in scenarios different from the one proposed in this thesis. These scenarios include matching the user preferences to products [Vlachou et al. 2011], or assigning new publications to subscribers [Basik et al. 2015; Chen and Cong 2015]. Reverse k nearest neighbors problem can be studied in two different setups: monochromatic or bichromatic. The setup is determined by whether sets of queries and sets of reverse nearest neighbors are the same (monochromatic) or disjoint (bichromatic) [Emrich et al. 2015] The model proposed in this thesis is an instance of a bichromatic RkNN, since the sets of queries and documents are disjoint.

Existing algorithms for finding reverse k nearest neighbors are not sufficient for application in the context of search exposure due to high dimensionality, large cardinality and sparsity of the query space. Most approaches heavily depend on geometric properties of the underlying space to perform efficient pruning [Vlachou et al. 2011].

5.8 Conclusion

This chapter introduces the problem of quantifying user search exposure, that is, finding the queries for which any of the user's posts is returned as a top-ranked result in a given search system. We cast the problem formally as reverse search and propose a method for ranking the queries in the resulting exposure sets to make the output user-friendly. The ranking task, moreover, uses a newly defined concept of search exposure relevance, which we studied in a series of AMT surveys.

We believe there are a number of fascinating research questions that could be studied as an extension to the work presented in this chapter. On the generation side, considering various ranking models, expanding the query length and efficient stream processing of search exposure requests, including parallel computation, caching and request partitioning would be necessary in a real-world deployment. On the usability and ranking side, further understanding of exposure relevance, designing better ranking methods, incorporating the probabilities of queries being asked to the overall setup, or detecting exposure in black-box systems, are only a few of such extension possibilities. Finally, further investigating layman perceptions regarding search exposure, as well as developing the expert understanding of the possible threats, would give us a better grip of this newly defined privacy question.

Chapter 6

Rank-Susceptibility

Contents

Privacy of Internet users is at stake because they expose personal information in posts created in online communities, in search queries, and other activities. An adversary that monitors a community may identify the users with the most sensitive properties and utilize this knowledge against them (e.g., by adjusting the pricing of goods or targeting ads of sensitive nature). Existing privacy models for structured data are inadequate to capture privacy risks from user posts.

This chapter presents a ranking-based approach to the assessment of privacy risks emerging from textual contents in online communities, focusing on sensitive topics, such as being depressed. We propose ranking as a means of modeling a rational adversary who targets the most afflicted users. To capture the adversary's background knowledge regarding vocabulary and correlations, we use latent topic models. We cast these considerations into the new model of R-Susceptibility, which can inform and alert users about their potential

for being targeted, and devise measures for quantitative risk assessment. Experiments with real-world data show the feasibility of our approach.

6.1 Introduction

Motivation and background. The goal of this chapter is to provide privacy risk assessments from textual data for users in online communities. An online post may directly or indirectly disclose personal information, such as gender, age, political affiliation, or interests. An adversary can combine such observations with his background knowledge of correlations between different attributes to infer privacy-sensitive information and discriminate against users. We argue that existing privacy models for structured data, such as k-anonymity [Sweeney 2002b], l-diversity [Machanavajjhala et al. 2007], t-closeness [Li et al. 2007], membership privacy [Li et al. 2013] and differential privacy [Dwork 2008], are inherently inappropriate to capture these situations. One reason is that user posts in social media are mostly of textual form, inducing a high-dimensional data space of word-level or phrase-level features. A second reason is that users might not want to be prevented from posting contents, but instead be selectively warned about emerging privacy risks. In our setting, certain assumptions also differ from the assumptions of prior work on privacy-preserving data publishing [Fung et al. 2010]: users do want to post information, but they should be aware of possible exposure and targeting risks. For these reasons, we pursue an IR-centric approach to privacy in this chapter, making novel use of topic models and ranking.

Scenario. To understand why adversaries and user risks are different from the privacy concerns for structured databases, consider the following scenario. An unscrupulous drug company wishes to advertise its new anxiety-reducing drug to Facebook users. It decides to target ads at a million users that are most susceptible to be afflicted by depression within the 1 billion population of Facebook. The company plans to infer users' demographics by text mining their posts and combine it with the background knowledge correlating demographics and certain vocabulary usage with depression, obtained from text mining an archive of medical journals. In such a scenario, how can a Facebook user estimate her risk of being targeted? Similar issues arise also within specialized online communities such as `healthboards.com` or `patient.co.uk`. Although these have a much smaller scale, a smart adversary would still target only a subset of highly susceptible users to avoid the impression of mass spamming.

Targeted ads of sensitive nature constitute one kind of risk, but there are even more severe threats with real cases reported: scoring users for financial credit worthiness or insurance payments, factoring a user's social-media posts in assessing her job application, and more. Despite these being big trends, most users do not need hard guarantees regarding privacy (e.g., preventing de-anonymization by all means), and perfect anonymity cannot be guaranteed without severely diminishing the utility of social media. For example, someone who always posts using a one-off anonymous identity cannot build up a reputation as a credible information source. Conversely, even making all posts under a pseudonym is insufficient to prevent tracking-and-rating companies (e.g., `www.spokeo.com`) from linking

user accounts across different social platforms. Therefore, we focus on the assessment of privacy *risks* and on alerting users to support their awareness, rather than pursuing the elusive goal of enforcing privacy.

Existing privacy models fail to capture theses issues, along the following dimensions:

- *Data model*: Privacy models like k-anonymity or differential privacy are primarily geared for structured data or content that can be cast into low-dimensional feature spaces. Capturing risks from textual contents in online communities faces the problem of high-dimensional feature vectors (e.g., word bigrams). Prior work that coped with text in specific settings, such as predicting sensitive posts [Peddinti et al. 2014], sanitizing the information from query logs [Carpineto and Romano 2015; Navarro-Arribas et al. 2012], or publishing high-dimensional datasets [Day and Li 2015]. Our goal, on the other hand, is to be able to quantify privacy risks from text in a generic way.

- *Adversary's background knowledge:* Prior work on privacy assumes computationally powerful adversaries, but disregards or makes special assumptions about the background knowledge that an adversary may have beyond the dataset at hand. However, adversaries may easilly tap into many datasets including large text corpora, thus obtaining a model of the typical vocabulary used by potential targets as well as semantic dependencies or statistical correlations between topics.

- *Disclosure vs. discrimination risk:* Existing privacy models focus on limiting information disclosure, but they do not capture the exposure within a community with regard to sensitive properties. Standing out in a community this way may result in discriminatory treatment, such as being rejected for loans or job applications, or receiving ads of sensitive nature.

Approach and challenges. This thesis introduces *R-Susceptibility*: a ranking-based privacy risk model for assessing users' privacy risks in online communities, accompanied by IR-style risk measures for quantifying risks from textual contents. The model is very versatile: in this chapter we demonstrate how it can capture user posts or search queries, but it can also be used with click streams, and other online activities. Semantic dependencies and statistical correlations among words and sensitive topics are represented using latent topic models, such as LDA [Blei et al. 2003] or Skip-grams [Mikolov et al. 2013]. This way, we anticipate adversaries with rich background knowledge. Adversaries are assumed to be rational: they target only a fraction of "promising" users. Therefore, we model the risk of a user as the ranking position in the community when all the users are ordered by the relevance of their contents to sensitive topics, such as pregnancy, depression, financial debts, etc. This ranking-based model is meant to alert the users whenever critical situations arise. We posit that users might be then guided to selectively post anonymously. Our model addresses several technical challenges:

- *Sensitive vs. general topics:* A trained latent topic model does not indicate which of the topics are privacy-sensitive. We carried out a crowdsourcing study to identify

sensitive topics. Our study differs from the prior work of [Peddinti et al. 2014] as the latter relied on explicit categories.

- *Personal vs. professional interest:* A user who posts about a sensitive topic may merely have a professional or educational interest without being personally afflicted. To be able to rank such users lower, our model introduces the notion of topical *breadth* of interest, complementing the user's strength of interest in a sensitive topic.
- *Personal interest vs. curiosity:* A user may become interested in a topic out of curiosity, perhaps prompted by an external event (e.g., a celebrity scandal). To be able to rank such non-critical users lower, our model also considers the *temporal variation* of interest in a topic.

The chapter's salient contributions are:

- a novel approach to privacy risks focusing on exposure in user rankings within online communities, and emphasizing risk awareness;
- a framework for quantifying privacy risks from textual contents in online communities, based on latent topic models and user rankings;
- measures for computing risk scores with regard to sensitive topics based on users' posts or search queries.

6.2 R-Susceptibility model

6.2.1 Sensitive states and adversaries

We assess the risk of a user being perceived as afflicted by a *sensitive state*, such as depression, pregnancy, or financial debts. An adversary in our model attempts to find the most susceptible users, that is, the users who are most exposed with regard to a sensitive state. For instance, an adversarial insurance company might want to identify the users who are likely afflicted by certain diseases, an adversarial HR department of a company might want to screen for the users with likely drug or alcohol problems, while a seller of illegal anti-depressants might want to find the users most likely to be depressed, and thus prospective customers.

We therefore propose *ranking* as a means of modeling a rational adversary trying to identify the most susceptible users. To rank the users with respect to a given sensitive state, an adversary needs to choose a measure of quantitative risk assessment based on the contents of user profiles. We discuss several such measures in Section 6.3.

6.2.2 Sensitive topics

We associate sensitive states with a vocabulary distribution, i.e., distributional vectors of related words. For example, the topic *financial debts*, could be captured by related words and phrases like *loan, mortgage, money, problem, sorrows*, or *sleepless night*. Such salient phrases related to a sensitive state can be obtained by unsupervised or semi-supervised

training of latent topic models over external datasets such as news archives, digital libraries or large crawls of social media. This way we capture the adversaries' background knowledge about the vocabulary for a topic and about semantic dependencies and correlations.

Sensitive states might manifest themselves in the online contents of users. User posts can also be characterized as distributional vectors of salient words. Then, the similarity between the distributional vectors of the user's posts and a sensitive topic can be used to assess the user's susceptibility to being exposed with regard to that topic.

6.2.3 Background knowledge

An adversary in our model is assumed to be interested in a sensitive state and aims to target a fraction of the most afflicted users. The adversary has *background knowledge*, characterized by statistical language and topic models. This is a natural form of useful knowledge for a rational adversary who wants rank the users based on the textual contents, and to bound the cost of his targeting efforts.

In this dissertation, we consider three versions of adversary's background knowledge. The basic version is the knowledge of the most salient words for different topics, which is assumed in all the solutions we explore. The more advanced version assumes that the adversary is able to compute similarities between words, in the sense of semantic relatedness. Finally, in some of the solutions, we assume an adversary is able to assign latent topics to broader thematic domains, e.g., the topic of depression to the domain of psychiatry.

We believe that this model reflects a wide class of adversaries whose goal is to discriminate and target the most susceptible users in online communities.

6.2.4 R-Susceptibility

We propose R-Susceptibility (Rank-Susceptibility) as a measure of a user's privacy risk. To measure R-Susceptibility with respect to a sensitive topic, we first rank all users within an online community based on their decreasing susceptibility of being exposed with regard to a sensitive topic (as described above) and then compute the position where the user is ranked.

Intuitively, the R-Susceptibility model could also have the following IR interpretation: we rank the users according to the relevance of their posts to a query containing the words of a sensitive topic, and choose the top-ranked, who should be the most likely to be personally afflicted.

6.3 Risk assessment measures

Risk measures are plug-in components in the framework and orthogonal to the idea of R-Susceptibility. In this dissertation, we begin by investigating three kinds of risk scores, leaving an extended risk-measure study as future work.

The first two of the risk scores are baselines, inspired by standard measures in privacy research, namely, the entropy of attribute value distributions (as used in the t-closeness model) and the changes in the global probability distributions of attribute values incurred by the inclusion of an individual user's data (as used in the differential privacy model). The

third measure is a novel IR-centric score based on topic models, capturing lexical correlations and three different characteristics of user interest in a topic: the strength of interest, the breadth of interest, and the temporal variation of interest.

Desired properties. By considering the community and interpreting risk with respect to a user's rank in the community, our framework does not impose any restrictions on the absolute values or the value domains of valid risk measures. Intuitively, for the framework to function, we expect a good measure to correlate with human assessments on the sensitivity of user profiles: the more human observers agree that a user might be in a sensitive state, the higher the value of the risk score should be.

6.3.1 Entropy baseline measure

The entropy baseline measure is inspired by comparing a global probability distribution (for an entire community) against a local distribution (for an individual user) using relative entropy (aka KL divergence). We apply this measure to textual data as follows.

Let X be a sensitive topic, and $\{x_1, ..., x_j\}$ be the salient words and phrases of X. The knowledge of this vocabulary for different topics is assumed to be a part of the adversarial background knowledge (e.g., derived from latent topic models). We treat $x_1, ..., x_j$ as database attributes and represent users as database records where the value of an attribute x_i equals to 1 if the word appears in the user's contents, and to 0 otherwise.

Let U_0 be the user for whom we wish to compute the risk score with respect to X, and $U = \{U_1, .., U_k\}$ be the set of other users in the community. Let further be $U^* = \{U_0\} \cup U$, and let P_U, P_{U^*} denote the distributions of attribute values for U and U^*, respectively.

We compute the risk score by averaging the relative entropy of the univariate distributions P_U, P_{U^*} for the individual attributes $\{x_1, ..., x_j\}$. Note that measuring the relative entropy over the multivariate joint distributions of attributes could be an alternative, but we do not pursue this here because of the data sparseness that we would face.

Definition 3 (Entropy baseline risk score of topic X for U_0)**.** *The entropy baseline risk score of the user U_0 with respect to a topic X is:*

$$risk_{\text{ENT}}(U_0, X) = \frac{1}{j} \sum_i \sum_{v=\{0,1\}} P_U[x_i = v] log(\frac{P_U[x_i = v]}{P_{U^*}[x_i = v]}) \tag{6.1}$$

The ranking method based on this definition is being referred to as ENT.

Measure properties. It holds that $risk_{\text{ENT}}(U_0, X) \geq 0$. The lowest value of 0 is reached when the user does not have any of the topic's salient attributes in her observable contents. Otherwise, the risk score is lowest when half of the community's users exhibit an attribute in their contents and highest when all or none of the users have the attribute.

6.3.2 Differential-privacy baseline measure

The differential-privacy-based measure is inspired by the definition of differential privacy, that is calculating the changes of attribute probabilities incurred by the inclusion of a user's

data. Let X, $\{x_1, ..., x_j\}$, U_0, U, U^*, P_U, and P_{U^*} be defined as in the previous section. The differential privacy principle requires that:

$$P_U[x_i] \leq 2^{\varepsilon} P_{U^*}[x_i] \text{ and } P_{U^*}[x_i] \leq 2^{\varepsilon} P_U[x_i] \tag{6.2}$$

for a small $\varepsilon > 0$. To give an ε-differential-privacy guarantee, existing methods would perturb the data by Laplacian noise if the inequalities are not already satisfied. However, our "attributes" are words in user posts that the user intentionally chose and our goal is to quantify risk rather than perturb the data. We thus aim to determine the best possible value of ε for which the guarantee holds without perturbation. This is the minimum ε for each x_i, but the guarantee is only as strong as the weakest x_i, leading to the following formulation:

Definition 4 (Differential-privacy baseline risk score of topic X for U_0)**.** *The differential-privacy baseline risk score of the user U_0 with respect to a topic X is:*

$$risk_{\text{D-P}}(U_0, X) = \max_{x_i} \left(\max \left(log \left(\frac{P_U[x_i]}{P_{U^*}[x_i]} \right), log \left(\frac{P_{U^*}[x_i]}{P_U[x_i]} \right) \right) \right) \tag{6.3}$$

The ranking method based on this definition is being referred to as DIFF-PRIV.

Measure properties. It holds that $risk_{\text{D-P}}(U_0, X) \geq 0$. The risk value is lowest for a user who does not have any of the sensitive topic's salient attributes in her contents and highest for a user who has a critical attribute that is not present in the contents of any other user.

6.3.3 Topical risk measure

To this end, we construct a distributional representation of each of the sensitive topics X (e.g., financial debts), user contents U (e.g., from an online community such as `quora.com`), and each post P the user authors in the online community. We model X, P and U as vectors in a distributional vector space.

6.3.3.1 Distributional vectors for topics and users

Topic vectors. Topics are represented as vocabulary distributions found by collecting word statistics over suitably chosen corpora.

Definition 5 (Sensitive Topic Vector)**.** *For sensitive topic X, the topic vector is a distributional vector $\vec{X}$ constructed using words or bigrams weighted by topic relevance.*

For example, *hiv* and *positive* are salient for the topic of hiv infection. Such topics and their salient phrases can be automatically extracted by applying latent topic analysis to large, thematically broad text corpora.

User vectors. To be able to relate posts and users to topics, we map each user U and post P created by the user in an online community to a vector.

Definition 6 (User Post and User Vectors)**.** *The content of a post P of a user U is modeled as a distributional vector $\vec{P}$. User U in the context of a topic X is modeled as a distributional vector $\vec{U}$ defined as:*

$$\vec{U} = \max_{P \in U} cos(\vec{P}, \vec{X}) \tag{6.4}$$

Vector construction. The exact mapping of topics and posts to vectors depends on the vector space in which we are operating. We use three different configurations in our experiments: i) a bag-of-words model (BOW), ii) an LDA model (LDA), and iii) a Skip-gram model (W2V).

Note that the use of LDA here is to construct a lower-dimensional vector space; this is orthogonal to using LDA for obtaining topics with their salient phrases, which we discussed above.

In the BOW vector space, we create topic vectors directly over the characteristic topic words with binary scoring; we also use these words as features with tf-scoring for user and post vectors.

In the LDA model, topic vectors are indicator vectors of for the latent dimensions. Users and posts are treated as documents that LDA maps into its low-dimensional latent space.

The third technique that we consider, `w2v`, is a model based on learning word relatedness, which can be trained over large text corpora [Mikolov et al. 2013]. To create the topic vectors in this word-centric vector space, we compute a weighted sum of words from the previously computed sensitive topic distributions. Since there is no natural mapping of documents to vectors in this setting, the procedure for posts is similar. However, to discount the impact of words unrelated to the topics at hand, we introduce a topic-dependent weighting scheme for user vectors. Namely, for a topic X and a post containing the set of words $\{v_1, v_2, \dots\}$, the post vector is $\vec{P} = \sum_j cos(\vec{v_j}, \vec{X}) \cdot \vec{v_j}$.

Risk scoring. Given these vectors, we can now compare a user posting history against a sensitive topic by vector-based similarity measures, like the cosine similarity. An advantage of this risk measure is that, unlike the ENTROPY or DIFF-PRIV measures, it does not require any community-level data, as the risk score of a user is independent of other users' data. Thus, each user can compute her score locally and privately, and send the value to a server to obtain an R-Susceptibility rank.

In addition to quantifying the *strength* of user interest in a sensitive topic, we also capture the *breadth* and *temporal variation* of that interest. This is crucial to avoid erroneously ranking higher those users who have a professional interest in a topic without being personally afflicted, or are temporarily interested out of curiosity. In our previous preliminary work in this area, we identified these two components to be crucial for reducing classification error in a similar setup [Biega et al. 2014].

6.3.3.2 Strength of interest

Having a vector representation of a user U, we can now compute the similarity between U and a topic vector X.

Definition 7 (Topic-aware risk score)**.** *The strength-of-interest risk score for a user U with respect to a topic X is:*

$$risk(U, X) = \cos\left(\vec{U}, \vec{X}\right) \tag{6.5}$$

We further refer to methods based on this definition as BOW, LDA, and W2V.

Measure properties. It holds that $-1 \leq risk(U, X) \leq 1$. A high value of this measure means the user has at least one post with vocabulary related to the topic. Thus, the strength of interest is reflected by the presence of the topic's salient vocabulary in user posts.

6.3.3.3 Breadth of interest

When ranking users, an adversary might want to distinguish between users who show a focused interest in a topic and users who show a broad interest in many topics within a domain, ranking the former higher than the latter. Applying this strategy could help, for instance, to capture users who are not personally afflicted but rather showing educational, hobbyist or professional interest in a topic. For example, for the topic of financial debts, a bank agent or finance hobbyist could offer advice in Q&A communities; similarly, a medical doctor or student could engage herself in health forums.

The posts of a user with a broad interest should exhibit a diversity of topics within their respective domain. We aim to capture this behavior, by means of distributional vectors, assigning each topic X to a broader domain, like finance, medicine, psychology, etc.

Definition 8 (Domain Vectors)**.** *A domain D is a set of topics $X_1, ..., X_{|D|}$ and its vector representation is a set of corresponding topic vectors $(\vec{X}_1, ..., \vec{X}_{|D|})$.*

To assess the risk taking into account whether a user U has a focused or a broad interest in a topic X, we compute:

1. how similar $\vec{U}$ is to $\vec{X}$ and
2. how dissimilar $\vec{U}$ is to the domain D by computing the distances between $\vec{U}$ and $\vec{X}_j$ for $j = 1..|D|$ and taking the $\lceil k * |D| \rceil$-th largest value, for some $0 < k \leq 1$.

If both of these measures are high, then we conclude that U is personally afflicted by topic X.

Definition 9 (Domain-aware risk score)**.** *The domain-aware risk score for a user U with respect to a topic X from the domain D is:*

$$risk_{\mathrm{D}}(U, X) = \cos\left(\vec{U}, \vec{X}\right) - max_{\lceil k*|D| \rceil}\{\cos\left(\vec{U}, \vec{X}_j\right) |\}_{j=1..|D|} \tag{6.6}$$

We further refer to methods based on this definition as BOW-D, LDA-D, and W2V-D.

Measure properties. It holds that $-2 \leq risk_{\mathrm{D}}(U, X) \leq 2$. The value would be high for a user who has a post containing topic's salient vocabulary, but whose contents do not exhibit any vocabulary from other topics in the respective domain. A low value occurs in a situation where the user has not written any posts related to the topic at hand, but has contents related to other topics in the domain. Studying the relative importance of the two components in different online communities is an interesting topic of future work.

The intuition for parameter k is that a personally afflicted user would not have high posting activities in k-fraction of different topics within the same domain. The value of the parameter controls how large the domain coverage should be for the users to be considered broadly interested. In practice, setting this parameter requires the knowledge of the breadth of topics discussed in a particular community.

6.3.3.4 Temporal variation of interest

Being interested in users most likely afflicted by a given state, we would like to rank the users who exhibit recurring activity regarding a topic X higher than the non-afflicted (possibly curious or exploratory) users exhibiting a short-term interest in the topic. Such a bursty activity might be prompted by prominent news related to X, be it sex scandals in the press, or social campaigns about depression.

To capture this issue, rather than computing a user vector $\vec{U}$ over the entire user history, we divide the history into time buckets and compute a sequence of vectors $\vec{U_i}$ using the contents from each bucket i separately. In our model, bucketization may be realized at different granularity levels depending on the user observation period and the characteristics of the community.

We then identify the top-m time buckets with the highest risk level, representing m different time periods (such as days or weeks). Let us denote these buckets of the user model as $U_1^*, \ldots, U_m^*$. A user whose interest in X is clearly above the level of a bursty interest (signifying occasional curiosity) would consistently have high risk scores in all of the top-m buckets. This leads us to our next definition of a user's privacy risk regarding topic X.

Definition 10 (Time-aware risk score)**.** *The time-aware risk score for a user U in time period i with respect to a topic X is:*

$$risk_{\mathrm{T}}(U, X) = \mathrm{avg}_{i=1..m} \left\{ \cos\left(\vec{U_i^*}, \vec{X}\right) \right\} \tag{6.7}$$

We further refer to methods based on this definition as BOW-T, LDA-T, and W2V-T.

Measure properties. It holds that $-1 \leq risk_{\mathrm{T}}(U, X) \leq 1$. The value would be high for a user whose posts contain relevant topic vocabulary in at least m observation buckets, and low for a user who does not exhibit topic's vocabulary in their contents.

The choice of a particular value of the m parameter depends on the available observation timeline and the characteristics of a given community. The parameter controls how often the activity regarding a topic should occur in order to not be considered occasional.

6.3.3.5 Combining domain- and time-awareness

The final measure we introduce combines all the aforementioned dimensions of interest. Note that we use bucketized user contents for computing the temporal variation component, but the breadth-of-interest component is computed over the full contents.

Definition 11 (Domain- and time-aware risk score)**.** *The risk of user U in time period i for topic X in domain D is:*

$$risk_{\mathrm{DT}}(U, X) = \mathrm{avg}_{i=1..m} \left\{ \cos\left(\vec{U_i^*}, \vec{X}\right) \right\} - \cos\left(\vec{U}, \left(\vec{D} - \vec{X}\right)\right). \tag{6.8}$$

We further refer to methods based on this definition as BOW-DT, LDA-DT, and W2V-DT.

6.4 Identifying sensitive topics

To complete our framework, we need to train a background knowledge model and answer the remaining question of how to identify sensitive topics. Although our model is applicable to any topic irrespective of its sensitivity, in practice users would only be interested in their R-Susceptibility ranks for truly sensitive topics. There is indeed a systematic way of gathering such information in a reasonably inter-subjective manner: training a latent topic model on a background corpus and crowdsourcing sensitivity judgments for each topic. This section presents our results along these lines.

6.4.1 Experiments on topic sensitivity

Datasets. We trained 3 LDA models, using the Mallet topic modeling toolkit: i) with 500 topics, on 600K Quora posts we crawled ii) with 200 topics, on 3M posts from health Q&A online forums, and iii) with 500 topics, on a sample of 700K articles from the New York Times (NYT) news archive.

Crowdsourcing sensitivity and domain judgements. We collected human judgements regarding the sensitivity and the domains of topics using Amazon Mechanical Turk (AMT), employing only master workers from the USA, and collecting 7 judgements per topic. For each of the topics, the workers were shown the 20 most salient words computed by LDA, and asked whether they would consider a post in social media containing these words privacy-sensitive. We explained that by privacy-sensitive we mean that a person uses these words because he/she is in a privacy-sensitive situation (e.g, alcohol addicted), or that the usage of these words might lead to a privacy-sensitive situation (e.g., political extremism). The first condition can capture, for instance, words related to diseases, the second can capture words related to political or religious positions.

We computed Fleiss' Kappa to measure the inter-annotator agreement for this task, obtaining 0.241 for the Quora topics, 0.294 for the HF topics, and 0.157 for the NYT topics. These low values confirm that sensitivity is rather subjective. However, there is a considerable number of topics in all of these corpora, which were unanimously or almost unanimously rated as sensitive. These were mostly related to health, private relationships,

#judges	#topics Quora	#topics NYT	#topics HF
7	29	8	38
6	43	27	32
5	48	60	30
4	56	84	21
3	68	73	22
2	99	90	28
1	106	111	23
0	51	47	6

Table 6.1: #topics with #judges agreeing on the topic being sensitive.

Topic	Vocabulary
clinical depression	depression depress suicide feel depressed suffer suicidal commit
drug addiction	drug addiction addict cocaine heroin substance meth addictive
pregnancy	baby birth pregnancy pregnant mother woman born child
hiv and viral diseases	hiv disease aids virus spread infection cure vaccine
financial debts	debt loan pay student interest payment money owe

Table 6.2: Examples: vocabulary of sensitive topics.

political and religious convictions, personal finance, legal problems and others. Table 6.1 shows the numbers of topics on which certain numbers of judges agree on their sensitivity.

The judges were also asked to assign a topic to one of seven high-level categories. Six of these, potentially containing some sensitive topics, were chosen based on the top-level Microsoft Academic Search categories. The annotators could also choose a generic category *other*.

Topics for evaluation in Section 6.5. For our further experiments, to make the much more laborious and costly evaluation of user profiles feasible, we leverage the above study to restrict the evaluation to 5 topics from the group of the most sensitive topics. The choice of particular topics is guided by the reported cases of social media screening by insurance companies, employers, and credit companies mentioned in Section 7.1. These are: *clinical depression*, *drug addiction*, *hiv*, *pregnancy*, and *financial debts*, assigned to the domains of *psychology*, *medicine*, and *finance&economy*. Table 6.2 shows the most prominent words for each of the chosen topics from the Quora topic model.

6.5 Experiments

6.5.1 Setup

6.5.1.1 Data sources

To test our methods in a variety of scenarios, we constructed three datasets using online communitites of different nature. As a first data source, we used the AOL query log collected between March and May 2006. The resulting data source amounts to around $107K$ users and more than $13M$ queries. The second data source consisted of over 5M posts spanning 13 years (2000-2013) from `healthboards` and `ehealthforum` Q&A health communities. We also collected data from the Quora Q&A community over a period of three months, between February and May 2015. The crawl focused on Quora users who were active in categories related to the considered sensitive topics and their domains, and comprised more than 200K users and 1.3M posts.

Ethics. To adhere to ethical standards concerning incorporation of user data into research, we decided to only use data that is publicly available – either as online profiles (Quora, Health Q&A), or as datasets used in numerous other studies (AOL). We never attempted to identify the individuals whose profiles we analyzed.

6.5.1.2 User sampling

We created our datasets by sampling the users from the data sources described above. However, we encounter a technical challenge, as the number of sensitive users for a given topic is very small when compared to the size of the whole community. Sampling users uniformly would not constitute a good benchmark for risk scoring methods. For example, we could achieve high accuracy, in a misleading way, by the simplistic prediction that all users are non-sensitive.

What we want, though, is ranking evaluation – our goal is to see how sensitive the users are in different ranking regions. Therefore, our sampling method is non-uniform and proceeds as follows. We first rank all users for each of the datasets using our basic strength-of-interest method from Section 6.3.3.2, and then sample users from this ranking. To pick a user, the sampling procedure orders users by their score, then computes prefix sums Σ_i for all users up to user i, with Σ_n being the score sum for all users. Then we draw a random number between 0 and Σ_n. If the number falls between Σ_i and Σ_{i+1}, we choose user $i+1$ (with users numbered from 1 to n. However, given that risk scores are extremely skewed, this sampling does still not yield good coverage of all the ranking regions. Therefore, we transform the original risk score q into a^q, where constant a needs to be determined based on the score skew in a data source. The intuition is to give a higher probability of being sampled to users with higher scores, so that the final sample set has good coverage of users with both high and low scores. Figure 6.1 depicts the depression risk scores of our 100 samples from the AOL data vs. the scores of the original dataset of 170K users.

In our case, a value of $a = 10^2$ for the Health Q&A, and a value of $a = 10^3$ for the AOL were reasonable to compensate the skew. For each of these datasets, we sampled 100 users

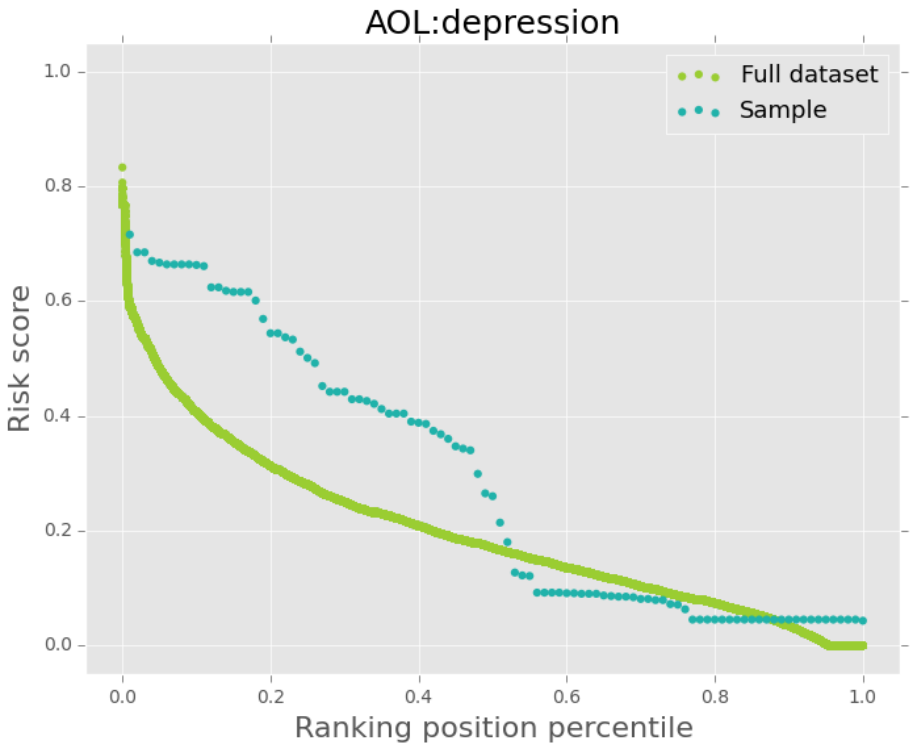

Figure 6.1: Example comparison of risk scores of sample vs. full data.

for each sensitive topic. We did not perform this kind of sampling for Quora, as our dataset was based on a focused crawl in the first place with focus on sensitive discussion threads. Since evaluating sizeable Quora profiles requires much more effort, for this data source we constructed smaller datasets with 40 users per topic. In total, our datasets comprised 1100 user profiles: personal histories of posts or queries.

6.5.1.3 User study for ground-truth labels

To assign sensitivity labels for user-topic pairs as ground truth, we used crowdsourcing and asked human judges to examine user profiles with chronologically ordered textual posts. Specifically, we asked whether based on the content of the profile, the judge suspects that the user (or a family member) is in a given sensitive state.

To evaluate the AOL and Health Q&A datasets, we employed AMT master workers from the USA and collected 5 judgements for each of the profiles. Since the majority of Quora profiles contain hundreds of posts, to ensure that proper care is given to evaluating them, we collected the judgements employing 19 students from our institution. We computed Fleiss' Kappa to quantify the global inter-annotator agreement across all the topics. The respective values for the AOL, Health Forums and Quora datasets were 0.442, 0.444, and 0.468 respectively, all corresponding to a moderate agreement. Table 6.3 shows the number of users who were marked by the human judges as sensitive by a majority vote.

	AOL	HF	Quora
depression	24	42	20
drugs	22	31	11
pregnancy	15	42	21
hiv	14	19	5
debts	24	n/a	11
TOTAL	99/500	134/400	68/200

Table 6.3: Number of sensitive users according to judges' assessments (x/y means x sensitive users out of y in total).

6.5.1.4 Configuration of methods

To evaluate the topic-model-based method, we used three different distributional vector spaces: a bag-of-word vector space, as well as two 500-dimensional vector spaces trained with (i) the LDA implementation from the Mallet toolkit [McCallum 2002] and (ii) `word2vec`[1] tool [Mikolov et al. 2013]. The latter two models were trained on NYT and Quora corpora described in Section 6.4.

In the breadth-of-interest model from Section 6.3.3.3, we set the parameter k to 0.3, i.e. we want a user of a broad interest in a domain to have at least a 30% coverage of topics from the domain.

In the temporal-variance-of-interest models described in Section 6.3.3.4, we compute the results using weekly time buckets and set the number of buckets parameter m to 3.

We later analyze the robustness of the ranking methods with respect to these parameters.

6.5.1.5 Ranking effectiveness metrics

- **R-Precision.** For a given sensitive topic, where r users were identified by the judges as sensitive, we compute the precision@r. When computing the average precision over all sensitive topics, we report both micro and macro average scores (summing over individual samples, and summing over topic precisions, respectively). To apply this measure, for each of the profiles we have to cast the five collected judgements to a binary score. We assume that an average of more than 0.5 classifies a user as sensitive. Note that r-precision imitates an adversary who, for instance, knowing that 1% of the population is depressed, ranks the users according to a depression-risk measure and chooses the top 1% of the users for further investigation.

- **Mean Average Precision (MAP).** For a given sensitive topic, we compute the average precision computed at the ranking positions of sensitive users.

- **Normalized Discounted Cumulative Gain (NDCG).** To asses the effectiveness of our methods using the actual non-binary judge assessments, we employ NDCG, which compares the rankings our methods yield with a perfect ranking obtained using the crowdsourced scores.

[1] `https://code.google.com/p/word2vec/`

6.5.1.6 Significance testing

The number of topics in our experiments is too small to perform significance tests over macro-averaged metrics. We thus resort to performing a paired t-test over r-precision differences on individual test samples within a dataset, marking the significance in the micro-averaged r-precision columns in the result tables. The $*$ symbols denote the case when the gain of a given ranking method over the baseline (ENT and DIFF-PRIV in Table 6.4, strength-of-interest baselines in Table 6.5) is statistically significant with a p-value < 0.05. This lets us conclude that a good r-precision score of a ranking method does not likely depend on the particular choice of user profiles.

6.5.1.7 Research questions

The remainder of the experimental section seeks to answer the following research questions.

RQ 1: Do the proposed topical risk measures perform better than the ENTROPY and DIFF-PRIV methods in predicting human risk judgements? (Sec. 6.5.2.)

RQ 2: Does the topical risk scoring measure perform better when extended with the breadth and temporal dimensions of user interest? (Sec. 6.5.3.)

RQ 3: How robust is the proposed method against changes in the parameter configuration and the background knowledge of the adversary? (Sec. 6.5.4.)

6.5.2 Traditional vs. IR risk scoring

We begin the risk scoring methods analysis by comparing the effectiveness of the baseline (ENTROPY, DIFF-PRIV) and the strength-of-interest topical risk scoring methods. Here, we choose the baseline IR-based methods for comparison, while exteding the measures with dimensions of interest will be addressed in the sections to follow.

Table 6.4 shows that the LDA risk scoring outperforms the alternatives (similar observation holds for W2V), which confirms that these methods are not naturally applicable to textual data in the context of risk scoring. The relatively good Precision@5 of these measures indicates that the most sensitive users tend to use highly salient words. However, operating on explicitly given salient attributes for each topic, the baseline measures do not capture any lexical correlations, an important prerequisite to capture users manifesting their sensitivity in a less direct way. This result validates the need to design new privacy risk measures better tuned to textual contents.

6.5.3 Risk scoring with dimensions of interest

We posited that extending the topical risk measures with the breadth and the temporal-variation dimensions of interest can help to predict sensitivity judgements better. Table 6.5 shows the evaluation results averaged over all topics, confirming that incorporating breadth and temporal variation into the risk score indeed improves the ranking performance.

We observe that breadth-of-interest is especially important for Quora, which is a Q&A community with a very wide variety of topics. Many Quora users seem to frequently post

	R-precision		Prec@5	MAP	NDCG
	micro	macro			
AOL					
ENTROPY	0.495	0.496	**0.760**	0.524	0.819
DIFF-PRIV	0.475	0.465	0.480	0.492	0.789
W2V	**0.556***	**0.533**	0.720	**0.589**	**0.836**
Health Forums					
ENTROPY	0.560	0.537	**0.750**	0.613	0.870
DIFF-PRIV	0.560	0.559	0.500	0.542	0.794
W2V	**0.664***	**0.634**	**0.750**	**0.696**	**0.894**
Quora					
ENTROPY	0.239	0.205	0.240	0.317	0.632
DIFF-PRIV	0.239	0.223	0.200	0.310	0.623
W2V	**0.343***	**0.341**	**0.280**	**0.352**	**0.637**

Table 6.4: Average metrics over all sensitive topics for different risk assessment measures

replies prompted by others rather than by their personal situation; hence the lower impact of the temporal component. Contrary, in AOL the temporal component takes over. With merely implicit cues in the form of queries, the temporal dimension is an important indicator of user sensitivity (also for the annotators). The breadth-of-interest component performs worse for AOL, possibly due to the short time span of the query log (3 months).

Note that in case of the proposed breadth-of-interest score, an underlying assumption is that an adversary is able to assign latent topics to broader thematic domains. Thus the best performing -DT methods imply a stronger background knowledge of an adversary.

Risk scoring for different topics Table 6.6 shows the values of r-precisions split by the topic, for different variants of LDA-based risk scoring. The trends observed in the results averaged over all topics can be seen here as well - there are consistent improvements across topics when incorporating the temporal and breadth dimensions. These results constitute anecdotal evidence that the proposed methods are general enough to be potentially applied to a variety of topics.

6.5.4 Robustness to configuration changes

Model changes. The BOW vector space models only an adversarial knowledge of salient words for different topics, whereas the latent vector spaces additionally enable an adversary to compute similarities between arbitrary words. The results presented in Table 6.5 show that this has a direct consequence in the risk ranking performance. The methods with the latent models as the background knowledge outperform the methods with the BOW background knowledge, while being comparable with each other. Thus the model seems resilient to rational background knowledge model changes, capturing a wide class of adversaries - the rational, cost-aware adversaries adopting latent models.

Training corpus changes. The results presented in the experimental section were obtained using the Quora topic model as the background knowledge model. We ran additional experiments using the NYT topic model described in section 6.4.1, and noticed that for the topics which were captured in the other latent model as well, we observe similar trends and dependencies in the results. This would suggest that an adversary has the freedom to choose among the inputs where his topics of interest are well captured.

Parameter changes in risk measures. The topical risk measures introduce two parameters: k for coverage of domain topics, and m for the number of (weekly) time buckets. Observing the values of r-precision and NDCG obtained when varying these parameters between $k = \{0.1, 0.2, ..., 1.0\}$, and $m = \{1, 2, ..., 12\}$, yields the following observations. First, when the parameters are set to values from the lower half of the ranges, we still observe improvements over the baseline strength-of-interest measure. Second, when the parameters are set to higher values, the results tend to deteriorate, possibly due to the incompleteness of user profiles in our datasets. Third, we observe higher sensitivity to parameters when a given dimension of interest is important for a given community (e.g. temporal for AOL, breadth for Quora). This result suggests that there is room for improvement within the framework of R-Susceptibility in that community-specific risk measures could be employed.

6.5.5 Discussion

The presented experimental results suggest that R-Susceptibility with appropriate risk measures is able to identify sensitive users with reasonable accuracy. The topical risk measures that quantify a user's exposure with respect to different topics work well, especially when the domain- and time-awareness components are included.

The R-Susceptibility framework allows the plugging of different risk measures, and in the future more advanced measures could be studied to address some of the limitations of this work. These could, for instance, model semi-experts, subtle vocabulary correlations, user contexts, or specific characteristics of a community.

6.5.5.1 User guidance

The R-susceptibility model and risk measures can work on a user history in a streaming manner, considering all contents up to a given point and periodically or continuously repeating the risk assessment. These methods could also be embedded in a privacy advisor tool that would help users assess their privacy risk, raising an alert when they become too exposed with regard to a sensitive topic.

6.5.5.2 Possible countermeasures by the platform

To prevent the risks describe in this chapter, the platform could prevent displaying the results related to certain sensitive topics. However, a countermeasure like this could also be considered as censorship. A middle-ground approach might be then for the platform to allow the users to select the topics for which they do not wish to be ranked.

	AOL				Health Forums				Quora			
	R-prec		MAP	NDCG	R-prec		MAP	NDCG	R-prec		MAP	NDCG
	micro	macro			micro	macro			micro	macro		
BOW	0.434	0.420	0.459	0.759	**0.642**	**0.625**	0.620	0.833	0.284	0.262	0.319	0.605
BOW-D	0.364	0.358	0.394	0.700	0.612	0.580	0.610	0.832	0.418*	0.398	0.400	0.672
BOW-T	**0.556***	**0.546**	**0.574**	**0.843**	0.582	0.574	0.619	0.873	0.284	0.285	0.317	0.605
BOW-DT	0.374	0.372	0.441	0.758	0.612	0.597	**0.667**	**0.894**	**0.463***	**0.444**	**0.440**	**0.688**
W2V	0.556	0.533	0.589	0.836	0.664	0.634	0.696	0.894	0.343	0.341	0.352	0.637
W2V-D	0.414	0.395	0.427	0.738	0.642	0.600	0.647	0.874	**0.493***	**0.465**	**0.532**	**0.776**
W2V-T	**0.586**	**0.580**	**0.645**	0.859	0.619	0.616	0.643	0.884	0.313	0.312	0.401	0.695
W2V-DT	0.545	0.530	0.601	**0.860**	**0.687**	**0.678**	**0.768**	**0.939**	0.463*	0.434	0.489	0.763
LDA	0.525	0.518	0.557	0.796	0.649	0.636	0.724	0.913	0.358	0.362	0.428	0.715
LDA-D	0.566	0.563	0.557	0.803	**0.716***	0.703	0.772	0.921	**0.493***	**0.485**	**0.489**	**0.752**
LDA-T	0.576	0.567	**0.655**	**0.879**	0.709	0.704	0.748	0.925	0.299	0.264	0.378	0.669
LDA-DT	**0.616***	**0.616**	0.649	0.859	**0.716***	**0.709**	**0.825**	**0.957**	0.418	0.389	0.481	0.751

Table 6.5: Results averaged over all sensitive topics.

6.6 Related work

Data-centric privacy. Methods for privacy-preserving data publishing [Fung et al. 2010] aim at preventing the disclosure of individuals' sensitive attribute values, while maintaining data utility, e.g., for data mining [Bertino et al. 2008], using concepts like k-anonymity [Sweeney 2002b], l-diversity [Machanavajjhala et al. 2007], t-closeness [Li et al. 2007], and membership privacy [Li et al. 2013]. All these models are geared for and limited to dealing with structured data, and this holds also for the most powerful and versatile privacy model, differential privacy [Dwork 2008]. In the field of Private Information Retrieval the goal of retrieving data from a database without revealing the query is mainly addressed by query encryption or obfuscation [Yekhanin 2010]. Generating dummy queries to obscure user activity is another technique studied in this area [Pang et al. 2012].

Sensitivity prediction. There is little research on characterizing what constitutes a sensitive topic. The recent work of Peddinti et al. [2014] analyzed features of posts and user behavior in Quora, and developed a classifier that can predict the sensitivity of individual posts. However, the solution is largely based on explicit categories (rather than latent embeddings) and the "go anonymous" posting option that users may choose. In contrast, our work aims to understand the sensitivity of any latently represented topic, and provide assessment for risk understood as topical exposure in a community.

Query log sanitization. This line of work tackles the challenge of an adversary using session information to infer user identities from queries [Adar 2007]. A variety of techniques have been proposed for anonymizing query logs, e.g., hashing tokens, removing identifiers, deleting infrequent queries, shortening sessions, and more [Cooper 2008; Fan et al. 2014; Hong et al. 2012; Korolova et al. 2009; Kumar et al. 2007]. Götz et al. [2012] compared different methods for publishing frequent keywords, queries and clicks, and showed that most methods are vulnerable to information leakage.

User-centric privacy. Stochastic privacy [Singla et al. 2014] is one of the few works that focus on users rather than data. This model introduces a user-defined threshold for sharing data to be obeyed by the platform provider. Closest in spirit to our approach is the work of Biega et al. [2014], which uses probabilistic graphical models to infer sensitive user properties, but is very limited in scope.

Linkability and de-anonymization. Privacy research for social networks has demonstrated the feasibility of linking user profiles across different communities [Goga et al. 2013] and de-anonymizing users [Narayanan et al. 2012; Narayanan and Shmatikov 2009; Zhang et al. 2014]. To prevent such attacks, a family of methods eliminates joinable attributes from published datasets [Vatsalan et al. 2013].

User behavior modeling. It has been shown that search queries can often be used to predict identity of users, as well as their gender, location, and other demographic attributes [Jones et al. 2007; Hu et al. 2007; Weber and Castillo 2010]. Such information can be

	AOL				Health Forums				Quora			
	LDA	LDA-D	LDA-T	LDA-DT	LDA	LDA-D	LDA-T	LDA-DT	LDA	LDA-D	LDA-T	LDA-DT
depression	0.542	0.542	0.625	**0.667**	0.762	0.762	**0.833**	0.762	**0.650**	0.550	**0.650**	0.600
drugs	0.545	**0.636**	0.591	**0.636**	0.710	**0.806**	0.677	0.774	0.364	**0.545**	0.273	0.364
pregnancy	0.533	0.667	0.533	**0.733**	0.571	**0.667**	0.619	**0.667**	0.095	**0.429**	0.095	0.381
hiv	0.429	0.429	**0.500**	**0.500**	0.526	0.579	**0.684**	**0.632**	**0.400**	**0.400**	0.200	**0.400**
debts	0.542	0.542	**0.583**	0.542	n/a	n/a	n/a	n/a	0.300	**0.500**	0.100	0.200

Table 6.6: Comparison of r-precision of LDA, LDA-D, LDA-T and LDA-DT for different topics.

harnessed for personalization but may also incur privacy threats. Pennacchiotti and Popescu [2011] analyzed Twitter profiles and network information to predict the political affiliation and race of users.

Expertise identification and trust analysis. Expert and trustworthy users can be identified based on their questions/answers contents and community votes [Adamic et al. 2008] or by analyzing user interaction graphs [Jurczyk and Agichtein 2007; Zhang et al. 2007]. Unlike in these works, our aim is not to identify experts, but to push the users who have a broad interest in a domain down the privacy risk ranking.

6.7 Conclusion

This chapter proposes a framework for quantifying privacy risks from textual contents of user profiles in online communities. By employing IR techniques such as ranking and latent topic models, it specifically addresses the risk of exposure with respect to sensitive topics and targeting by a rational adversary with rich background knowledge about topic vocabulary and word correlations.

Although more large scale studies of adversarial risk scoring strategies are needed, our experiments constitute a proof of concept that the approach is a viable basis for privacy risk assessment for users who want to post about sensitive topics but would like to be warned when the risk of being targeted becomes high.

In the future, R-Susceptibility can be extended to incorporate other forms of online activities, and be integrated in a framework for risk mitigation through appropriately guided user actions. Our vision is a trusted personal privacy advisor which assesses risks, alerts the user when critical situations arise, and guides her in appropriate countermeasures.

Chapter 7

Privacy through Solidarity

Contents

Online service providers gather vast amounts of data to build user profiles. Such profiles improve service quality through personalization, but may also intrude on user privacy and incur discrimination risks. As the next contribution of the thesis, we propose a framework which leverages solidarity in a large community to scramble user interaction histories. While this is beneficial for anti-profiling, the potential downside is that individual user utility, in terms of the quality of search results, may severely degrade. To reconcile privacy and user utility and control their trade-off, we develop quantitative models for these dimensions and effective strategies for assigning user queries to Mediator Accounts. We demonstrate the viability of our framework by experiments using a querylog with rich user profiles synthesized from the StackExchange Community Question Answering (CQA) forum.

7.1 Introduction

Motivation. Users are profiled and targeted in virtually every aspect of their digital lives: when searching, browsing, shopping, or posting on social media. The gathered information is used by service providers to personalize search results, customize ads, provide differential pricing, and more [Hannak et al. 2013; Teevan et al. 2005]. Since such practices can greatly intrude on an individual's privacy, the goal of our research is to devise a mechanism to *counter* such extensive *profiling*.

A careful user can largely preserve her privacy by taking measures like anonymizing communication or using online services only in a non-linkable manner (for instance, by changing accounts or pseudonyms on a regular basis). However, this comes at the cost of greatly reducing utility, both for the service providers and the user. On the one hand, the service provider will miss out on learning from the same user's long-term behavior, which may result in less effective systems. This issue of system-level utility has been studied in the past research on privacy [Krause and Horvitz 2010; He et al. 2014]. On the other hand, the individual user will experience degraded service quality, such as poor search results, as the service provider would not understand the user's interests and intentions. This notion of user-level utility has not been extensively explored in prior work. This dissertation formalizes the *trade-off* between a user's *profiling privacy* and her *individual utility*.

State of the art and its limitations. Research in privacy has primarily addressed the disclosure of critical properties in data publishing [Bertino et al. 2008; Chen et al. 2009; Fung et al. 2010]. Common techniques include coarsening the data so that different users become indistinguishable (e.g., k-anonymity [Sweeney 2002b], l-diversity [Machanavajjhala et al. 2007], and t-closeness [Li et al. 2007]), or perturbing the answers of an algorithm so that the absence or presence of any record does not significantly influence the output – the principle of differential privacy [Dwork 2008]. These methods consider notions of utility that reflect a system-level error in an analytical task, such as classification. In contrast, our goal is to prevent detailed profiling and targeting while keeping the *individual user utility* as high as possible, for example, in terms of the quality of personalized search results or product recommendations.

For privacy-preserving search, many approaches have been proposed based on *query obfuscation* [Gervais et al. 2014; Peddinti and Saxena 2014]. In these solutions, queries are generalized to hide user intentions, or additional dummy queries are generated to prevent accurate profiling. Both techniques come at the cost of largely reducing user utility. However, none of the prior work addressed the trade-off between privacy and user utility in a quantitative manner. A few methods [Chen et al. 2011; Peddinti and Saxena 2014] have considered an entire user community as a means for query obfuscation. This idea is related to our approach in this dissertation – we generalize it and make it applicable in the context of anti-profiling.

Approach and contribution. Our approach to reconcile privacy and user utility builds on the following observation: service providers often do not need a complete and accurate user profile to return personalized results. Thus, in accordance with the need-to-know principle,

we assign user requests to *Mediator Accounts* (MA) mimicking real users, such that (i) individual user profiles are scrambled across MAs to counter profiling, while (ii) coherent fragments of a user's profile are kept intact in the MAs to keep user utility high. We call this paradigm *privacy through solidarity*. Specifically, MAs are constructed by *split-merge assignment* strategies: splitting the interaction history of a user and merging pieces of different users together. Mediator Accounts are meant as an intermediate layer between users and the service provider, so that the provider *only sees MAs* instead of the real users.

Ideas along these lines have been around in the prior literature [Reiter and Rubin 1998; Santos et al. 2008; Xu et al. 2009; Goodrich et al. 2012; Rebollo-Monedero et al. 2012], but the formalization of the trade-off between privacy and user-utility has never been worked out. In particular, to make this idea viable, one needs to devise quantitative measures for the effects of Mediator Accounts on privacy and utility. In addition, a strategy is needed for assigning user requests to such accounts. The simplest approach of uniform randomization would be ideal for privacy but could prove disastrous for user utility. This dissertation addresses these challenges within a framework of Mediator Accounts. Our ideas are general enough to be applied to search engines, recommender systems, and other online services where personalization is based on the user interaction history. The salient contributions of this chapter are:

- a model with measures for quantifying the trade-off between profiling privacy and user utility;
- the Mediator Accounts framework together with strategies for assigning user interactions to MAs;
- comprehensive experiments using a large query log derived from the StackExchange CQA community.

7.2 Framework overview

7.2.1 Architecture

The architecture of the Mediator Accounts framework is shown in Fig. 7.1. It consists of three parties: users, a service provider (SP), and a Mediator Accounts proxy (MA-proxy). A *user profile* consists of a set of *objects*, such as queries, product ratings or other forms of user interactions with the SP. Instead of issuing objects directly to the SP, users pass them on to the MA-proxy together with some context information. The goal of the MA-proxy is to redistribute the incoming objects on to mediator profiles mimicking real users. The MA-proxy assigns each incoming object to a Mediator Account offering the right context for the current object and user, and issues the object to the SP from the chosen MA. Upon receiving a response (for example, a result page or a product recommendation) from the SP, the MA-proxy passes it back to the user. When an interaction is over, the MA-proxy discards all linking information about the original user and the object and remembers only the association between the mediator account and the object. As a result, the original user

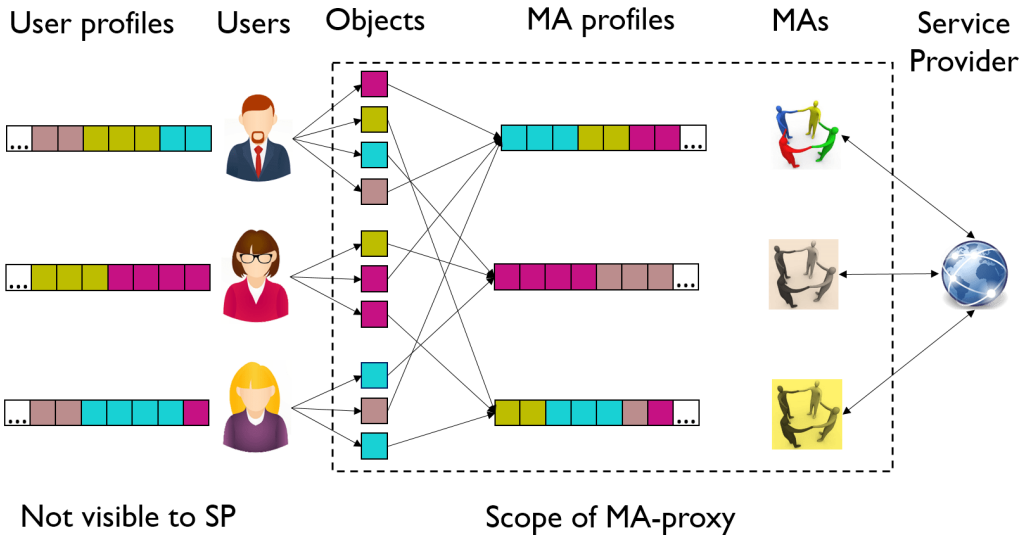

Figure 7.1: Overview of the MA framework.

profiles are scrambled across multiple MAs, and each MA consists of data from multiple users.

7.2.2 Incentives of participating parties

Users. The goal of a user participating in an MA system is to be able to get high-quality personalized results, while not letting any online provider (neither SPs nor the MA-proxy) keep her interaction history and link it to her as an individual. The MA-proxy has the user interaction history scrambled across multiple accounts, and no links between the objects and the real users are stored.

Users of anonymous services that do not offer topical personalization, such as the DuckDuckGo, Startpage or Qwant search engines, may be open to trading off some privacy for enhanced results through the MA framework.

Non-profiling service providers. The incentive of a non-profiling service provider would be to enhance personalization in the results, without compromising on the non-profiling principle.

Profiling service providers. A big question is whether profiling service providers would allow a third-party like an MA-proxy to mediate between them and the users. While examples of such third-parties already exist (the Startpage search engine uses Google as a source of search results), we believe that (i) an MA-proxy being able to group objects into realistic profiles that yield similar analytics results for the SP, and (ii) an MA-proxy being able to attract privacy-wary users who would not otherwise use the profiling SP, would be viable incentives for an SP not to block an MA service.

MA-proxy. An MA-proxy could be set up by individuals, or cooperatives of non-profiling SPs (to provide personalization without accumulating real user profiles), or by non-governmental organizations that promote online privacy. The Electronic Frontier Foundation is such an organization – a non-profit organization that has built privacy-preserving solutions like Privacy Badger.

7.2.3 Trusted and adversarial parties

MA-proxy. Users opting for an MA service would need to trust that it scrambles their profiles across mediator accounts, and discards the original profiles as well as any identifying information once an interaction (a single request or a session) is complete. A standard approach to gain such trust would be to make the MA solution open-source, enabling the code to be vetted by the community. A real implementation of an MA framework would have to take into account secure end-to-end communication channels between users and SPs via the MA-proxy. These issues may be resolved using encryption and security techniques (e.g., secure browser, onion routing, etc.), and are outside the scope of this thesis.

Provider. The service provider is not exactly distrusted, but there have been cases where user-related information has been leaked or passed on beyond the original intentions – by sabotage, acquisition by other companies, or enforcement by government agencies. By detaching users from profiles and limiting their accuracy, the potential damage is bounded.

Other risks might result from service providers displaying privacy-sensitive personalized ads, such as ads related to pregnancy or health issues, especially when observed by others on a user's screen. The architecture would allow an MA-proxy to support filtering ads and adjusting them to users' topical interest. Such a configuration has indeed been found to be a preferable ad-serving setup in a user study [Agarwal et al. 2013]. Ad filtering, however, is orthogonal to this research.

Third parties. Profiling companies that operate outside the user-provider connections are considered untrusted. The same holds for agglomerates of providers that aggregate and exchange user data. A conceivable attack could be to guess a user's attribute (e.g., whether she is pregnant) by combining (i) observations on the MAs and (ii) observations on a set of accounts in a social network, using statistical inference methods. The MA framework aims to keep such risks low by breaking observable associations between MAs and real users, and limiting the profiling accuracy of the split-merge superpositions of different users that cannot be easily disentangled.

7.3 Assignment model

The core of the MA framework is an algorithm for assigning user objects to Mediator Accounts. To guide it on the privacy-utility trade-offs and to assess the quality of the output, we need measures for quantifying the effect of an assignment on privacy and user utility. This section presents such measures, and the algorithm for object assignment based on the split-merge concept.

7.3.1 Concepts and notation

We use the following notations:

- A set U of users $u_1 \dots u_p$.
- A set O of objects $o_1 \dots o_s$ issued by users; the objects are treated as unique even if they represent the same content. For instance, a query `folk music` issued by user u_i is treated as an object distinct from the same query issued by u_j. Analogously, a product rating for a book (`Folk Music History`, 3.0) by u_i is distinct from the rating by u_j for the same book, irrespective of the rating value.
- A set M of mediator accounts $m_1 \dots m_t$ to which objects are assigned by the MA-proxy.

We reserve the symbols i, j, k for subscripts of users, objects, and MAs. If user u_i issues object o_j, we write $o_j \in u_i$. Similarly, if o_j is assigned to MA m_k, we write $o_j \in m_k$.

Assignments. An *assignment* of objects on to MAs can be denoted as an $s \times t$ matrix A of 0-1 values, where $A_{ij} = 1$ means that o_i is assigned to m_j. If we think of the Cartesian product $O \times M$ as a bipartite graph, then the assignment can be conceptualized as a subgraph $S \subseteq O \times M$ where each node of type O has exactly one edge with one of the M nodes.

7.3.2 Objective

In a real application, an MA-proxy has to assign objects to accounts in an online manner, one object at a time as input arrives. In this chapter, we focus on analyzing the model and assignments in an offline setting, although the algorithm we devise can be applied in both offline and online scenarios. The offline case is useful for two reasons. First, it is a foundation for understanding the underlying privacy-utility trade-offs. Second, performing offline assignment on a set of initial user profiles can address the cold-start problem that a new MA-proxy would face. Using the notation from Sec. 7.3.1, the *MA offline assignment problem* can be defined as follows:

- Given a set of objects O belonging to a set of users U, and the set of mediator accounts M, compute an assignment matrix A that optimizes a desired objective function for the privacy-utility trade-off.

The *MA online assignment problem* is:

- Given an assignment A of past user objects to MAs and a newly arriving object o of user u, find the best MA to which o should be assigned with regard to a desired goal for the privacy-utility trade-off.

7.3.3 Measuring privacy gain

An ideal situation from the perspective of privacy is when the objects from a user profile are spread across MAs uniformly at random – this minimizes the object-level similarity of any MA to the original profile. We thus measure privacy as the entropy of the user distribution over MAs, formalizing these notions as follows.

Entropy. We introduce for each user u_i an *MA-per-user vector* $\vec{mu}_i \in (\mathcal{N}_0)^t$ with one counter (≥ 0) per MA, written as $\vec{mu}_i = \langle x_{i1} \dots x_{ij} \dots x_{it} \rangle$ where x_{ij} is the number of objects by user u_i in account m_j (such that $\sum_{j=1}^{t} x_{ij} = |u_i|$). We can cast this into an *MA-per-user probability distribution* $\mathbf{\Phi_i} = \langle \phi_{i1} \dots \phi_{ij} \dots \phi_{it} \rangle$ by setting $\phi_{ij} = x_{ij}/|u_i|$ followed by smoothing (e.g., Laplace smoothing) so that $\phi_{ij} > 0$ for each j and $\sum_{j=1}^{t} \phi_{ij} = 1$.

The degree of u_i's profile fragmentation can be captured by the entropy of the distribution $\mathbf{\Phi_i}$. We can define the *MA-per-user entropy* as a measure of *privacy gain* (gain over having each user exhibit her full individual profile):

$$privacy{-}gain(u_i) = H_i = -\sum_{j=1}^{t} \phi_{ij} \log \phi_{ij} \tag{7.1}$$

This quantifies the spread of the user's objects across accounts. The higher the entropy value, the higher the gain in profiling privacy.

Profile overlap. If a use-case requires a more user-interpretable measure of privacy, an alternative is to minimize the *maximum profile overlap*. For a user u_i, this measure can be expressed as:

$$O_i = \max_{j=1}^{t} \frac{|\{o \in u_i \cap m_j\}|}{|u_i|} \tag{7.2}$$

This measure of overlap can directly tell a user how much "error" could be made by an adversary, who assumes one of the MAs is the user's profile. The optimum for this measure, as with entropy, is achieved when the objects are uniformly spread across accounts. Thus, in the following, we use entropy as our privacy measure, and leave maximum profile overlap as a design alternative.

7.3.4 Measuring user utility loss

User utility loss measures to what extent an object o_k of user u_i is placed *out of context* by mapping it to account m_j. We define a real-valued function $sim(\cdot, \cdot)$ to measure the coherence of user and MA profiles: $sim(o_i, o_j) \in [0, 1]$ is a symmetric measure of the relatedness between objects represented by o_i and o_j. In practice, different notions of relatedness can be used, based on object properties or usage. In settings where labels for topics or categories are available, we can set $sim(o_i, o_j) = 1$ if o_i and o_j are issued by the same user and have the same topic/category label, and 0 otherwise. Generally, we assume that *sim* measures are normalized with values between 0 and 1.

The objects of user u_i form a *context*, typically with high pairwise relatedness among the objects. When considering sets of objects as a whole (rather than time-ordered sequences of object posts), we can measure the *normalized context coherence* of an object o_k in the profile of user u_i by:

$$coh(o_k, u_i) = \frac{\sum_{o_l \in u_i, k \neq l} sim(o_k, o_l)}{|u_i| - 1} \tag{7.3}$$

When o_k is placed in MA m_j, we analogously define:

$$coh(o_k, m_j) = \frac{\sum_{o_l \in m_j, k \neq l} sim(o_k, o_l)}{|m_j| - 1} \tag{7.4}$$

The utility loss of u_i in a given MA assignment is then measured as an average coherence loss over all user objects:

$$utility{-}loss(u_i) = \frac{\sum_{o_k \in u_i} [coh(o_k, u_i) - coh(o_k, m_j)]}{|u_i|} \tag{7.5}$$

where m_j is the account containing o_k in the given assignment.

The normalization helps to account for varying sizes of user profiles. As a result, coherence values are always between 0 and 1, and utility loss is normalized to take values between -1 and 1. Note that our utility measure assumes that the context coherence can increase if an object is assigned to an MA with more similar objects. Coherence increase will result in negative utility loss.

7.3.5 Assignment algorithms

The role of an assignment algorithm is to scramble user objects across accounts so as to satisfy a desired privacy-utility tradeoff or optimize a corresponding objective function. In this chapter, we experiment with a number of assignment algorithms and study their output quality.

7.3.5.1 Optimal assignment (offline)

The trade-off can be expressed as a joint non-linear optimization problem as follows:

$$\max_{A} \min_{u} [\alpha \cdot privacy\_gain(u) - (1-\alpha) \cdot utility\_loss(u)] \tag{7.6}$$

Alternatively, one could optimize one of the two measures with a constraint on the other. Solving this problem exactly, however, is computationally expensive. If we use the less complex overlap privacy measure, we could cast the problem into a Quadratic Integer Program. However, this would have millions ($|M| \cdot |O|$) of variables; so it would remain intractable in practice. We thus do not pursue this direction in this dissertation and instead consider a number of heuristics. The following are also suitable for the *online case*.

7.3.5.2 Profiling-tradeoff assignment

We aim to approximate the combined objective function as follows. Let o be an object we want to assign to one of the accounts m_j. If we want to optimize for privacy (i.e., entropy), we should choose an MA at random from a uniform distribution over MAs:

$$P_{priv}(m_j|o) = \frac{1}{|M|} \tag{7.7}$$

If we want to optimize for utility, we could choose an MA that offers the best coherence:

$$P_{util}(m_j|o) = \begin{cases} 1, & \text{if } m_j = m_{max} \\ 0, & \text{otherwise} \end{cases} \tag{7.8}$$

where $m_{max} = \arg\max_{m_k} coh(o, m_k)$.

Let α be a parameter that controls the trade-off between privacy and utility. We sample an MA according to the distribution:

$$P(m_j|o) = \alpha \cdot P_{priv}(m_j|o) + (1 - \alpha) \cdot P_{util}(m_j|o) \tag{7.9}$$

In the offline case, we may choose an arbitrary order of objects to feed into this assignment heuristic. In the online case, we process objects ordered by the timestamps in which they are issued to the MA-proxy. It is also worth noting that in an online setting users could choose different α for each object, deciding that some should be assigned randomly, and some with the best possible context.

7.3.5.3 Random assignment

In this assignment, objects are assigned to accounts uniformly at random. This is a special case of the Profiling-tradeoff algorithm with $\alpha = 1$. This assignment maximizes privacy.

7.3.5.4 Coherent assignment

Personalization is usually based on semantically coherent parts of user profiles. If we retain such coherent fragments of a profile within the accounts, individual utility should be preserved better than in a completely random assignment. The mode in which we assign an object to the account that offers the best coherence is a special case of the Profiling-tradeoff algorithm, in which we set $\alpha = 0$. We refer to this method as Coherent. This assignment explicitly aims for the best utility only, yet some privacy is gained as chunks of user profiles get assigned to MAs randomly.

7.4 Mediator accounts in search systems

By analyzing query-and-click logs, search engines can customize results to individual users. Such user profiling, however, may reveal a detailed picture of a person's life, posing potential privacy risks. At the same time, personalization of a single query is often based on only a subset of a user's history. Thus, as a first use case, we apply the MA framework in a search engine setting, scrambling the query histories of different users across accounts.

7.4.1 Framework elements

In the search scenario, the elements of the framework described in Sec. 7.3 are instantiated as follows. The objects are *keyword queries*, and user profiles consist of sets (or sequences) of queries, possibly with timestamps. Accounts contain re-assigned queries of different users. Object similarity can be understood as topical similarity between queries, with topics being either explicit such as categories or classifier labels, or latent, based on embeddings. As a query is characterized by a set (or weight vector) of topics, the similarity can be computed, for instance, using (weighted) Jaccard overlap or vector cosine. The service provider in this setting is a search engine, which, upon receiving a query from a given user profile, returns a ranked list of documents personalized for that user. User utility is measured by the quality of the result list.

7.4.2 Service provider model

The ability of the MA framework to preserve utility while splitting user profiles across accounts depends on a retrieval model for ranking query answers. We use the language-model-based retrieval technique [Croft et al. 2009], as described below.

Let $o \in u$ be a query of user u consisting of a number of words $w \in o$, and D be the document collection. The model retrieves the results in two steps. First, it fetches a set of top-k documents $D_o \subseteq D$, each document $d \in D$ being scored by the query-likelihood model with Dirichlet smoothing (parameter μ_D [Croft et al. 2009]):

$$score(o,d) = \log P(o|d) = \sum_{w \in o} \log \left(\frac{tf_{w,d} + \mu_D \cdot P(w|D)}{|V_d| + \mu_D} \right) \tag{7.10}$$

where $tf_{w,d}$ is the count of w in d, $P(w|D)$ is the probability that w occurs in D, and $|V_d|$ is the count of all words in d. For every user u, we compute a personalization score as the log-probability of the document d being generated from the user language model using Dirichlet smoothing with parameter μ_U, where U is the set of all users (or equivalently, the collection of their search histories):

$$score(d,u) = \log P(d|u) = \sum_{w \in d} \log \left(\frac{tf_{w,u} + \mu_U \cdot P(w|d)}{|V_u| + \mu_U} \right) \tag{7.11}$$

where $tf_{w,u}$ is the count of w in the search history of u, $P(w|d)$ is the probability that w occurs in d, and $|V_u|$ is the count of all words in the search history of u.

In the second step, documents $d \in D_o$ are re-ranked using a linear combination of the two scores:

$$score_u(o,d) = \gamma \cdot score(o,d) + (1-\gamma) \cdot score(d,u) \tag{7.12}$$

In practice, γ would be set to a low value to put more importance on personalization.

When we use the MA framework, the computations are similar. The notion of a user is simply replaced by an account m. The personalization stage is adjusted as follows: we compute $score(d,m)$ using $P(d|m)$, which in turn is computed using $tf_{w,m}$, μ_M and $|V_m|$ with Eq. 7.11. Definitions of these quantities are analogous to their user counterparts.

7.5 Experiments

7.5.1 Experimental setup

7.5.1.1 Dataset

For lack of publicly available query logs with user profiles, we created a query log and a document collection using the data from the Stack Exchange Q&A community (dump as of 13-06-2016). We excluded the large software subforums from outside the Stack Exchange web domain (such as StackOverflow), as they would dominate and drastically reduce the topical diversity. The final dataset consists of ca. $6M$ posts of type 'Question' or 'Answer' in 142 diverse subforums (e.g., Astronomy, Security, Christianity, Politics, Parenting, and Travel).

Document collection. We use all posts of type 'Answer' as our collection. The resulting corpus contains $3.9M$ documents.

User query histories. We construct a query log from posts of type 'Question', as these reflect users' information needs. Each question is cast into a keyword query selecting the top-l question words with the highest TF-IDF scores, where l is a random integer between 1 and 5. We consider only users with at least 150 questions, which yields a total of 975 users and $253K$ queries. Each query is assigned a topical label, used for object similarity. We set this label to the *subforum* where the original question was posted.

7.5.1.2 Service provider

For reproducible experiments, we base our search engine model on the open-source IR system Indri [Strohman et al. 2005]. Indri ranks query answers based on state-of-the-art statistical language models with Dirichlet smoothing [Croft et al. 2009]. We use Indri to retrieve the top-100 results for every query from the entire corpus, and implement user-personalized re-ranking ourselves (see Sec. 7.4.2). We compute per-user language models from the original questions to tackle sparsity. The Dirichlet smoothing parameter is set to the average document length (56 words), and γ is set to 0.1.

7.5.1.3 Empirical measures

Privacy Gain. The model entropy reflects how scrambled the user profiles are. Yet from the perspective of a profiling adversary it is rather the distribution over semantic topics that matters. Empirically, a proper way to measure privacy then is to compare the original topic distribution per user against the topic distributions of the MAs. The minimum KL-divergence between pairs of these distributions signifies the privacy level:

$$emp{-}priv{-}gain(u_i) = \min_{m_j \in M} D_{KL}(P^{u_i} \parallel Q^{m_j}) \tag{7.13}$$

where P^{u_i} and Q^{m_j} refer to the user and MA profile distributions over topics with add-one Laplace smoothing. We use subforums as explicit labels for topics.

Utility Loss. Rankings of documents d for a query are derived from $score_u(o, d)$ and $score_m(o, d)$ (Eq. 7.12), respectively, where the former refers to the query being issued by user u and the latter to the query being issued by the mediator account m (see Sec. 7.4.2). We quantify the empirical utility loss as the divergence between the two rankings. We compute two measures: the loss in Kendall's Tau over the top-100 document rankings: $1 - KTau@100$ (as the personalization step considers the top-100 documents), and the loss in Jaccard similarity coefficient over the first 20 ranking positions: $1 - Jaccard@20$ (as end-users typically care only about a short prefix of ranked results). For each user, we average these scores over all queries.

α	**M-Priv-Gain** (Entropy)	**M-Util-Loss** (Coherence Loss)	**E-Priv-Gain** (Min. KL-div.)	**E-Util-Loss** (1 - KTau@100)
Original	0.000	0.000	0.000	0.000
0.0 (Coh)	1.180	0.178	0.320	0.170
0.2	2.208	0.293	0.319	0.203
0.4	3.130	0.389	0.346	0.228
0.6	3.975	0.463	0.389	0.246
0.8	4.731	0.515	0.494	0.260
1.0 (Rand)	5.287	0.535	0.863	0.266

Table 7.1: Results with trade-off parameter α for the model (M) and empirical (E) measures.

7.5.1.4 Assignment methods

Object similarity. We set $sim(o_i, o_j) = 1$ if both o_i and o_j belong to the same user and to the same topic, and 0 otherwise. During the assignment, this measure helps to keep related parts of a user profile together.

Assignment algorithms. We run the Profiling-Tradeoff algorithm varying α between 0 and 1 with a 0.1 increment, and setting the number of MAs to be the number of users (975). With the chosen object similarity, the special case of $\alpha = 0$, i.e. the Coherent assignment, results in splitting user profiles into subforum chunks and assigning each chunk to a randomly chosen account.

7.5.2 Results and insights

Aggregate trends. Table 7.1 presents the results on the model measures and empirical measures for different values of the assignment trade-off parameter, macro-averaged over users. Recall that $\alpha = 0.0$ and $\alpha = 1.0$ correspond to the special cases of Coherent and Random assignments, respectively. These results need to be contrasted with the baseline, denoted *Original* in the table, where each original user forms exactly one account (i.e., no scrambling at all). Compared to the baseline, all numbers are statistically significant by paired t-tests with $p < 0.01$. For empirical utility loss, we report Kendall's Tau; the results for the Jaccard coefficient are similar.

The results show that the Profiling-Tradeoff assignments improve privacy over the *Original* baseline (the topical KL-div. between original users and MAs is increased) while keeping the utility loss low. This is largely true regardless of the exact choice of α. So the MA framework provides a fairly robust solution to reconciling privacy and utility, supporting the observation that high-quality topical personalization does not require complete user profiles.

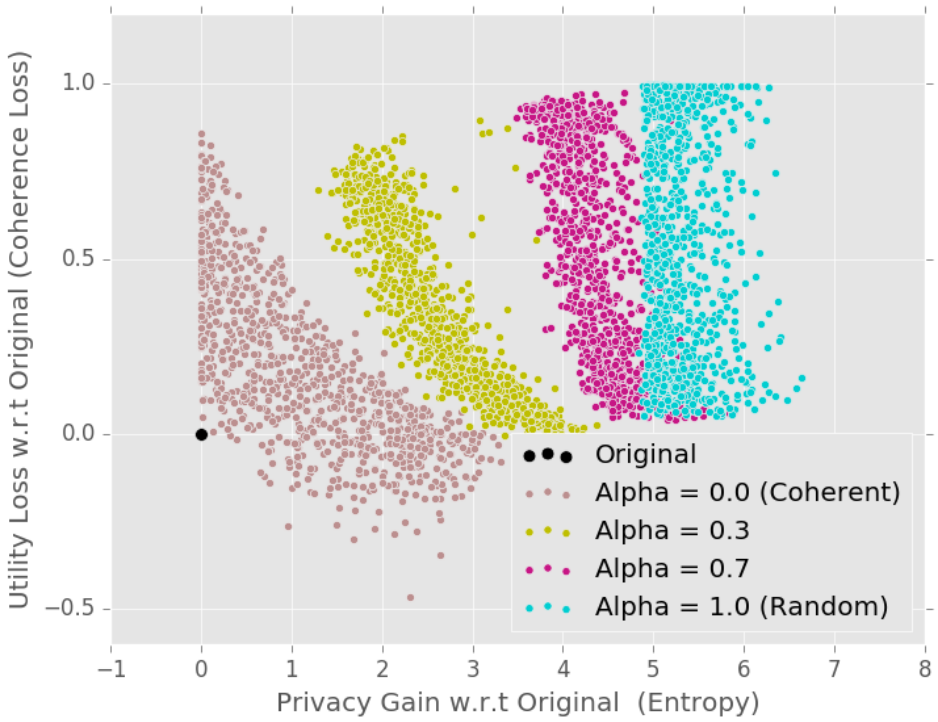

Figure 7.2: Model measures per user.

With the α increasing, assignments become more random, so the privacy increases and utility is reduced (but with a low gradient). In this regard, the empirical measures reflect the expected behavior according to the model measures well.

Results per user. Figs. 7.2 and 7.3 show privacy and utility values of each user, for the model and empirical measures, respectively. Different colors represent different assignments, and each dot represents a user, with measures averaged over the user's queries. We have several observations:

- Higher privacy gain is correlated with higher utility loss. The Original assignment maps each user to the origin (0 utility loss, but also 0 privacy gain). No assignment reaches the bottom-right area of the chart – which would be an ideal.
- Varying α not only tunes the privacy-utility tradeoff at the community aggregate level, but also affects the variance over individual user scores. This suggests that we should further explore choosing α on an individual per-user basis (which is easily feasible in our framework, but is not studied in this thesis).
- Even the Random assignment ($\alpha = 1.0$) keeps utility reasonably high. This is due to the fact that random MAs – sampled from queries in the community – end up being averaged rather than random profiles.

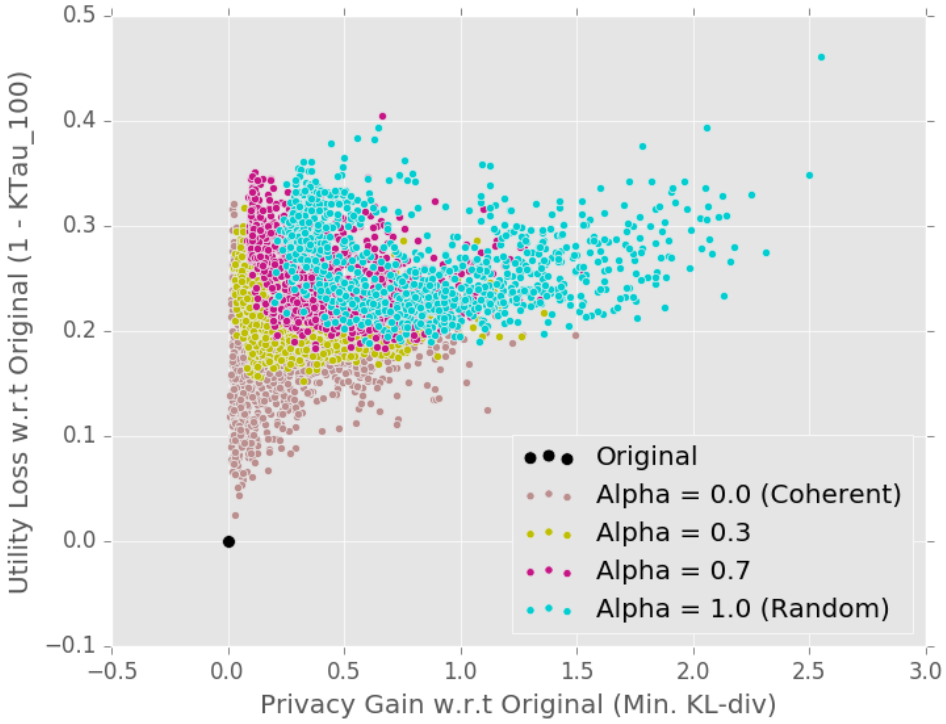

Figure 7.3: Empirical measures per user.

- Some users achieve high privacy gains without losing hardly any utility, and vice versa. We investigate this further below.

Effect of profile size and diversity. We analyze how different user profile characteristics affect the assignment results. Figure 7.4 presents the empirical trade-offs for the Coherent (top row) and Random (bottom row) assignments, where each dot is a user and the dot color represents (i) the logarithm of the number of queries in the user profile (left column), or (ii) the diversity of the profile measured by the entropy of the distribution of queries across topics (right column). We make the following observations:

- Users with more queries (darker dots) in the Coherent assignment clearly gain privacy at the cost of losing utility, whereas for the smaller profiles (lighter dots), the trade-off is not as pronounced. In the Random assignment this trade-off is less pronounced irrespective of the size of the profile.

- In the right column, one can see the lighter dots (profiles with little diversity) moving from the bottom-left for the Coherent assignment (little privacy gain, little utility loss) to the top-right for the Random assignment (higher privacy gain, higher utility loss). This suggests that our framework does not offer much help to the users with uniform and focused interests. This is an inherent limitation, regardless of which privacy protection is chosen. Such homogeneous users cannot hide their specific interests,

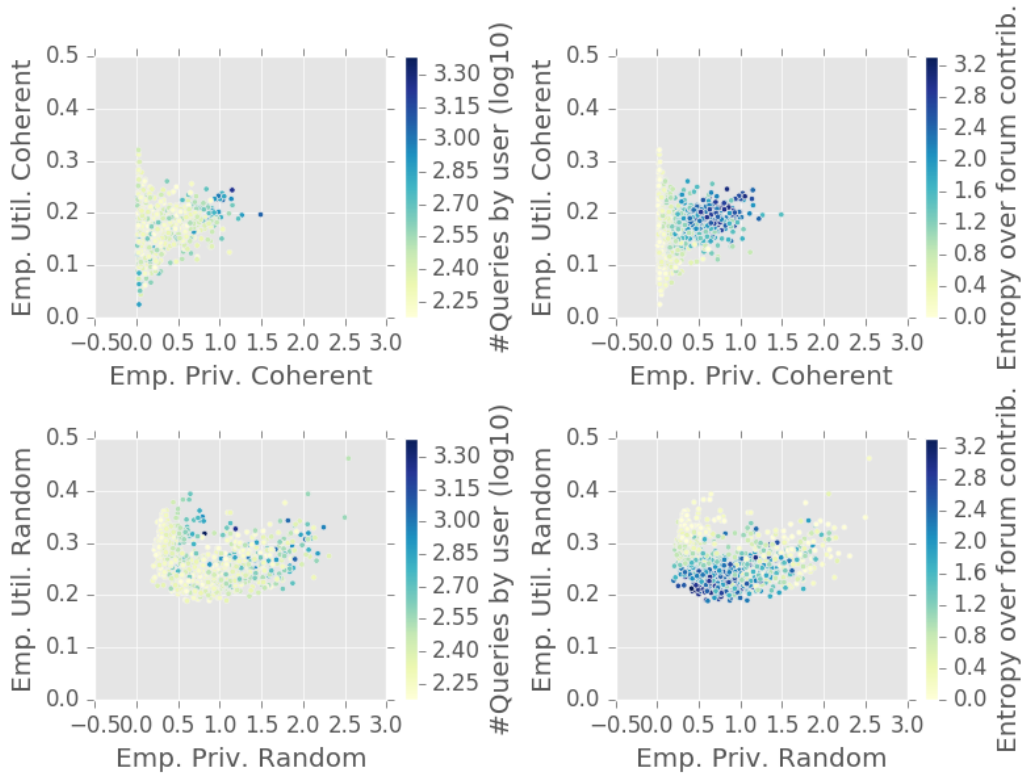

Figure 7.4: Effect of profile size and diversity.

unless they give up on personalization utility.

- Our split-merge assignments offer good results for users with high diversity. As suggested by the darker dots, the Coherent assignment leads to a lower utility loss and higher privacy gain for users with diverse profiles, when compared to the Random assignment. This is because such users have more independent and internally coherent chunks that can be split without affecting utility. This class of users is exactly where the right balance of utility and privacy matters most, and where we can indeed reconcile the two dimensions to a fair degree.

7.6 Related work

Grouping for privacy. The idea of masking the traces of individual users by combining them into groups has been around since the *Crowds* proposal by Reiter and Rubin [1998]. However, this early work solely focused on anonymity of web-server requests. Narayanan and Shmatikov [2005] devised an abstract framework for group privacy over obfuscated databases, but did not address utility. For search engines specifically, Jones et al. [2008] proposed a notion of query bundles as an implicit grouping of users, but focused on countering de-anonymization in the presence of so-called vanity queries. The short paper by Zhu et al. [2010] sketches a preliminary approach where semantically similar queries by different users

are grouped for enhancing privacy. Aggregation of users' website-specific privacy preferences through a centralized server [Yu et al. 2016], can also be perceived as a type of *privacy through solidarity*. The principle of solidarity has moreover been explored through a game-theoretic framework over recommender systems [Halkidi and Koutsopoulos 2011].

Tracking and profiling. A good body of work investigates to what extent and how users are tracked by third parties in web browsers [Lerner et al. 2016; Meng et al. 2016b; Yu et al. 2016], or through mobile apps [Meng et al. 2016a]. These are primarily empirical studies with an emphasis on identifying the tracking mechanisms. The interactions with service providers, where users log in and leave extensive traces, have been largely disregarded. In contrast, our framework helps counter both tracking and individual profiling by detaching users from online accounts.

To reduce the scale of profiling, a model called stochastic privacy has been proposed to selectively sample user profiles for use by personalizing algorithms [Singla et al. 2014]. To counter profiling by search engines in particular, Xu et al. [2007] has proposed to issue queries anonymously, but provide the engine with a coarse topical profile for answer quality. On the tracking front, the Non-Tracking Web Analytics system reconciles users' need of privacy and online providers' need of accurate analytics [Akkus et al. 2012]. Although these various works address the privacy-utility trade-off, no explicit control mechanism has been proposed for user utility.

Privacy-preserving IR. The intersection of privacy and IR has received some attention in the past years [Yang et al. 2016]. One of the key problems studied in the field is that of post-hoc log sanitization for data publishing [Cooper 2008; Götz et al. 2012; Zhang et al. 2016a]. Online sanitization, on the other hand, aims at proactively perturbing and blurring user profiles. Techniques along these lines typically include query broadening or dummy query generation [Shen et al. 2007; Balsa et al. 2012; Peddinti and Saxena 2014; Wang and Ravishankar 2014]. It has also been proposed to perturb user profiles by making users swap queries and execute them on behalf of each other [Rebollo-Monedero et al. 2012]. Very few of these prior works consider the adverse impact that obfuscation has on utility, and the usual focus is on the utility of single query results. To the best of our knowledge, none of them focuses on personalization utility or offers quantitative measures for the trade-off.

Another privacy concept studied in IR is that of exposure. Recently, the notions of R-Susceptibility and topical sensitivity have been proposed to quantify user exposure in sensitive contexts within a given community [Biega et al. 2016].

Privacy-preserving data mining. There is a vast body of literature on preserving privacy in mining data for rules and patterns and learning classifiers and clustering models [Fan et al. 2014; Aggarwal 2015]. In this context, utility is measured from the provider's perspective, typically an error measure of the mining task at hand (e.g., classification error) [Bertino et al. 2008]. In the context of recommender systems, rating prediction accuracy [McSherry and Mironov 2009; Nikolaenko et al. 2013] and category aggregates [Shen and Jin 2016] are typically used as proxies for utility. Techniques for user profile perturbation have also been

studied for utility-preserving differentially-private recommeders [Guerraoui et al. 2015].

7.7 Conclusion

We presented Mediator Accounts (MAs): a framework to counter user profiling while preserving individual user utility as much as possible. The framework enables decoupling users from accounts, making direct targeting impossible, and profile reconstruction or de-anonymization much harder. At the same time, users are still able to benefit from personalization by service providers. The versatility of the framework has been demonstrated in experiments using a large query log synthesized from the StackExchange platform. While the application of the framework to recommeder systems is not a contribution of this thesis, Biega et al. [2017b] have additionally demonstrated the applicability of the framework in that scenario.

While our model allows for flexible trade-offs between privacy and utility, a key question in our empirical study has been to understand how well the MAs can preserve the utility in terms of high-quality search results. The experiments show that the split-merge approach with Coherent assignment improves the privacy, while incurring little user utility loss. These benefits are most pronounced for users with larger profiles (i.e., more activity) and higher diversity of interests. Open issues for future work include practical deployment, handling of other personalization features, and exploring the options for tuning assignments and framework parameters to the specific needs of individual users. On top of that, analyzing the three-dimensional trade-off between user privacy, user utility and the traditional service provider utility could help ensure that the resulting mediator profiles are a useful source for user analytics, making an MA proxy a tolerable component of the online landscape.

Finally, we would hope that the MA proposal stirs up the investigation of how the need-to-know principle could be implemented in case of personalized online services.

Chapter 8

Conclusions and Outlook

This thesis broadly investigates privacy and fairness problems of search system users, including both searchers and searched subjects. The models we propose together with the experimental results suggest that there is scope for search systems to provide better experience for their users in terms of privacy and fairness, without the need to sacrifice much of the search utility. The results from Chapter 4 show that systems could provide more equitable exposure to search subjects without much reduction in ranking utility, especially since in many existing scenarios there are numerous subjects who are equally relevant to various queries. Chapters 5 and 6 exemplify how systems could provide their users with more privacy awareness by computing queries which expose users in search results. Annotators in our user studies found exposure by topically sensitive queries to be especially problematic, while media reports about privacy breaches in search engines highlight the problems with exposure by unique queries. Finally, Chapter 7 shows that search engines do not need to accumulate rich query histories per user to deliver quality personalization, particularly for users with topically diverse profiles and interests. Going forward, there are a number of fascinating questions that need to be answered to design privacy-friendly and fair search engines.

Fair exposure. To create holistically fair systems, we need to gain a deeper understanding of the intrinsic properties of ranking and human relevance feedback mechanisms that might lead to bias and unfairness. Moreover, the existence of position bias calls for redesigning display interfaces to try to minimize the biasing effects of visual ranking perceptions in applications such as people search. Beyond exposure measured through ranking position, interface design and the results of people search should also be examined for representational harms. It might be possible, for instance, that various demographic groups have different information highlighted in the snippets on the search results. To aid the algorithm and system design, inspired by social comparison effects, we should also better understand human perceptions regarding the fairness of their positions in ranking.

Sensitive exposure. To mitigate the consequences of exposure without reducing the utility of the systems, topical sensitivity and search exposure relevance should be contextualized. To this end, we need to better understand which textual contents are sensitive for which types of users in which situations. On the mechanism side, developing external black-box methods for auditing search exposure is an important complement to existing privacy support methods, as well as a means of incentivizing service providers to develop more comprehensive internal tools. Beyond user perceptions, it is also vital to develop

expert understanding of the risks, studying the consequences of exposure by different types of queries and proposing reasonable mitigation solutions beyond awareness. Last but not least, to be able to tangibly raise awareness of these problems, we should investigate user perceptions regarding search exposure, as well as user understanding of the underlying mechanisms and coping strategies.

Minimizing profiling. Our results revealed that often only a fraction of a user's interaction history is needed to deliver quality personalization. We can turn this observation around to arrive at a more general question: What is the minimum amount of information needed to maintain personalization quality? To answer this question, more work is needed to define generalized framework-independent notions of profiling privacy and user utility, and inspect the interplay between user utility and system utility. Moreover, not less important is understanding which solutions would provide enough incentives for service providers to be adopted and to what extent users would be willing to trade some personalization utility for more privacy.

Infrastructure supporting research in privacy and fairness. Ironically, developing solutions for user privacy and fairness often requires access to user data or even users themselves if privacy and fairness-related annotations from the data owners are necessary. While more readily available in the industry, access to real user data is limited for academic researchers. Developing infrastructure for research in privacy and fairness is thus a crucial factor fostering progress in these areas. Such infrastructure includes, for instance, protocols and data sanitization algorithms for data release, methods for synthesizing realistic data (this thesis contributes to such infrastructure by creating a synthetic query log), and providing data in shared benchmarks.

We envision search engines that serve their users equitably, that offer support mechanisms for their users to understand and control the use of their data and their exposure in the search results, and that collect the minimum amount of data necessary to deliver the service.

Appendix A

AMT User Study: Topical Sensitivity

This appendix documents the details of the Amazon Mechanical Turk[1] user study described in Chapter 6, Section 6.4.1. The goal of the study was to collect judgments on privacy sensitivity of different topics in a topic model (topic models were trained using Mallet [McCallum 2002]). Each topic was represented by its 20 most salient words, that is, the words with the highest probability from the topic.

Setup details:

- **Title:** A survey about sensitivity of words in online posts (WARNING: This HIT may contain adult content. Worker discretion is advised.)
- **Description:** We'd like to collect your judments on how privacy-sensitive different sets of words might be when used in online posts.
- **Keywords:** online posts, sensitive words, privacy, sensitivity of topics
- **Reward per assignment:** $2
- **Number of assigments per HIT:** 7
- **Time alloted per assignment:** 90 minutes
- **Master workers:** Yes
- **HIT Approval Rate for all Requester's HITs:** >= 90%
- **Location:** is United States
- **Number of HITs Approved:** >= 100

Instructions: In this task, we want to collect your judgements on whether you would consider public posts on social media (e.g., Facebook or Twitter or blog) containing certain sets of words to be potentially "privacy sensitive". We consider the usage of a set of words to be privacy sensitive if any of the following is true:

- A person is likely to use these words because they are in a situation that is privacy-sensitive. For example, if you use words related to diseases it might mean you are sick, which is a privacy-sensitive situation.

[1] https://www.mturk.com/

- The usage of these words might create a privacy-sensitive situation or some problems. For example, if you use vocabulary indicating your religious views, the information might easily be used against you, leading to a violation of privacy.

Additionally, for each set of words, please choose a best suiting thematic domain.

Question: Each HIT consisted of 50 sets of questions of the following form:

1. Do you think a post in social media containing these words can be privacy-sensitive? (Yes/No)
20 most salient words from the topic were displayed here

2. What thematic domain do these words come from?
(Law and Politics, Humanities, Psychiatry and Psychology, Health and Medicine, Economy and Finance, Other)

Appendix B

AMT User Study: Search Exposure

This appendix documents the details of the Amazon Mechanical Turk[1] user study described in Chapter 5, Section 5.5.4. The goal of the study was to collect judgments on the search exposure relevance of different queries (more privacy-critical queries have higher search exposure relevance.)

Setup details:

- **Title:** A study of profile exposure through keyword search (WARNING: This HIT may contain adult content. Worker discretion is advised.)
- **Description:** We'd like to collect your opinions regarding sensitivity of different search terms on Twitter.
- **Keywords:** exposure of tweets, keyword search
- **Masters has been granted :** Yes
- **Qualification Requirement:** HIT Approval Rate (%) for all Requesters' HITs greater than or equal to 95
- **Location:** is United States
- **Number of HITs Approved:** >= 100

Instructions, queries only: Websites like Twitter offer their users the option to search for tweets using keyword queries (= sets of words). Upon entering the keywords, you receive a ranked list of tweets that somehow match these keywords as a response.

Assume for now you are a Twitter user with a number of tweets in your profile. Even though these tweets are public, they are usually exposed just to your followers. But if it turns out that one of your tweets is returned as a top-ranked result in response to these keyword queries, your tweet (and profile) will be exposed to whoever issues these keywords. There is a variety of people who might be looking for user profiles or tweets that match certain keywords. Examples include journalists looking for examples to their stories, companies building databases of people for different purposes, or even criminals looking for victims.

[1] https://www.mturk.com/

In this survey, we will ask for your opinions regarding sensitivity of different keyword queries in the above context. More specifically, we will show you a number of keyword queries and ask whether you would feel concerned (e.g., uncomfortable, embarrassed, privacy-violated, or threatened) if your tweet was returned as one of the top answers to these search terms.

Question, queries only: Would you feel concerned (uncomfortable, embarrassed, privacy-violated, or threatened) if your tweet was returned as one of the top answers to these search terms? Please, try to choose 'Yes' for at least 10-12 search terms you would feel most uncomfortable with, although you are also free to choose more.

Search terms: XXX

(Yes/No)

Instructions, queries+tweets: Websites like Twitter offer their users the option to search for tweets using keyword queries (= sets of words). Upon entering the keywords, you receive a ranked list of tweets that somehow match these keywords as a response.

Assume for now you are a Twitter user with a number of tweets in your profile. Even though these tweets are public, they are usually exposed just to your followers. But if it turns out that one of your tweets is returned as a top-ranked result in response to these keyword queries, your tweet (and profile) will be exposed to whoever issues these keywords. There is a variety of people who might be looking for user profiles or tweets that match certain keywords. Examples include journalists looking for examples to their stories, companies building databases of people for different purposes, or even criminals looking for victims.

In this survey, we will ask for your opinions regarding sensitivity of different keyword queries in the above context. More specifically, we will show you a number of **search terms** together with the **tweets that are returned as top results** for these terms. The question is: **in your opinion, should a user be concerned** (uncomfortable, embarrassed, privacy-violated, or threatened) if the tweet was returned as one of the top answers to the search terms?

Question, queries only: In your opinion, should a user be concerned (uncomfortable, embarrassed, privacy-violated, or threatened) if their tweet below was returned as one of the top answers to these search terms? Please, try to choose 'Yes' for at least 10-12 search terms you would feel most uncomfortable with, although you are also free to choose more.

Search terms: XXX

(Yes/No)

Bibliography

Ahmed Abbasi and Hsinchun Chen. Writeprints: A stylometric approach to identity-level identification and similarity detection in cyberspace. ACM Transactions on Information Systems, 26(2):7:1–7:29, 2008. 13

Rediet Abebe, Jon M. Kleinberg, and David C. Parkes. Fair division via social comparison. In Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems, AAMAS 2017, São Paulo, Brazil, May 8-12, 2017, pages 281–289. ACM, 2017. 29, 54

Alessandro Acquisti. Nudging privacy: The behavioral economics of personal information. IEEE Security & Privacy, 7(6):82–85, 2009. 20

Alessandro Acquisti, Curtis Taylor, and Liad Wagman. The economics of privacy. Journal of Economic Literature, 54(2):442–92, 2016. 20

Lada A. Adamic, Jun Zhang, Eytan Bakshy, and Mark S. Ackerman. Knowledge sharing and yahoo answers: everyone knows something. In Proceedings of the 17th International Conference on World Wide Web, WWW 2008, Beijing, China, April 21-25, 2008, pages 665–674. ACM, 2008. 96

Eytan Adar. User 4xxxxx9: Anonymizing query logs. In Proceedings of the Query Log Analysis Workshop, International Conference on World Wide Web, 2007. 18, 94

Philip Adler, Casey Falk, Sorelle A. Friedler, Tionney Nix, Gabriel Rybeck, Carlos Scheidegger, Brandon Smith, and Suresh Venkatasubramanian. Auditing black-box models for indirect influence. Knowledge and Information Systems, 54(1):95–122, 2018. 25, 55

Lalit Agarwal, Nisheeth Shrivastava, Sharad Jaiswal, and Saurabh Panjwani. Do not embarrass: re-examining user concerns for online tracking and advertising. In Symposium On Usable Privacy and Security, SOUPS '13, Newcastle, United Kingdom, July 24-26, 2013, pages 8:1–8:13. ACM, 2013. 20, 101

Charu C. Aggarwal. Data Mining - The Textbook. Springer, 2015. ISBN 978-3-319-14141-1. 112

Istemi Ekin Akkus, Ruichuan Chen, Michaela Hardt, Paul Francis, and Johannes Gehrke. Non-tracking web analytics. In the ACM Conference on Computer and Communications Security, CCS'12, Raleigh, NC, USA, October 16-18, 2012, pages 687–698, 2012. 16, 112

Mishari Almishari, Ekin Oguz, and Gene Tsudik. Fighting authorship linkability with crowdsourcing. In Proceedings of the second ACM conference on Online social networks, COSN 2014, Dublin, Ireland, October 1-2, 2014, pages 69–82, 2014. 15

Mor Armony and Amy R. Ward. Fair dynamic routing in large-scale heterogeneous-server systems. Operations Research, 58(3):624–637, 2010. 55

Ricardo A. Baeza-Yates. Bias on the web. Communications of the ACM, 61(6):54–61, 2018. 26, 27

Ero Balsa, Carmela Troncoso, and Claudia Díaz. OB-PWS: obfuscation-based private web search. In IEEE Symposium on Security and Privacy, SP 2012, 21-23 May 2012, San Francisco, California, USA, pages 491–505. IEEE Computer Society, 2012. 112

Solon Barocas and Andrew D Selbst. Big data's disparate impact. California Law Review, 104(3):671, 2016. 26, 27

Solon Barocas, Moritz Hardt, and Arvind Narayanan. Fairness and Machine Learning. fairmlbook.org, 2018. http://www.fairmlbook.org. 23

Fuat Basik, Bugra Gedik, Hakan Ferhatosmanoglu, and Mert Emin Kalender. S^3-tm: scalable streaming short text matching. VLDB Journal, 24(6):849–866, 2015. 73

Richard Berk, Hoda Heidari, Shahin Jabbari, Matthew Joseph, Michael J. Kearns, Jamie Morgenstern, Seth Neel, and Aaron Roth. A convex framework for fair regression. CoRR, abs/1706.02409, 2017. 25

Michael S. Bernstein, Eytan Bakshy, Moira Burke, and Brian Karrer. Quantifying the invisible audience in social networks. In 2013 ACM SIGCHI Conference on Human Factors in Computing Systems, CHI '13, Paris, France, April 27 - May 2, 2013, pages 21–30. ACM, 2013. 17, 58

Elisa Bertino, Dan Lin, and Wei Jiang. A survey of quantification of privacy preserving data mining algorithms. In Privacy-Preserving Data Mining - Models and Algorithms, volume 34 of Advances in Database Systems, pages 183–205. Springer, 2008. 94, 98, 112

Bin Bi, Milad Shokouhi, Michal Kosinski, and Thore Graepel. Inferring the demographics of search users: social data meets search queries. In 22nd International World Wide Web Conference, WWW '13, Rio de Janeiro, Brazil, May 13-17, 2013, pages 131–140, 2013. 3

Asia J. Biega, Azin Ghazimatin, Hakan Ferhatosmanoglu, Krishna P. Gummadi, and Gerhard Weikum. Learning to un-rank: Quantifying search exposure for users in online communities. In Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, CIKM 2017, Singapore, November 06 - 10, 2017, pages 267–276, 2017a. 6, 60, 65

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, Shinjuku, Tokyo, Japan, August 7-11, 2017, pages 675–684, 2017b. 7, 8, 13, 18, 41, 113

Asia J. Biega, Krishna P. Gummadi, and Gerhard Weikum. Equity of attention: Amortizing individual fairness in rankings. In Proceedings of the 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR 2018, Ann Arbor, MI, USA, July 08-12, 2018, pages 405–414, 2018. 5, 6, 25, 28

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@CIKM 2014, Shanghai, China, November 7, 2014, pages 29–36, 2014. 82, 94

Joanna Asia Biega, Krishna P. Gummadi, Ida Mele, Dragan Milchevski, Christos Tryfonopoulos, and Gerhard Weikum. R-susceptibility: An ir-centric approach to assessing privacy risks for users in online communities. In Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval, SIGIR 2016, Pisa, Italy, July 17-21, 2016, pages 365–374, 2016. 7, 8, 58, 62, 70, 71, 112

Sarah Bird, Solon Barocas, Kate Crawford, Fernando Diaz, and Hanna Wallach. Exploring or exploiting? social and ethical implications of autonomous experimentation in ai. Available at SSRN: https://ssrn.com/abstract=2846909, 2016. 29

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. Journal of Machine Learning Research, 3:993–1022, 2003. 77

Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam Tauman Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain, pages 4349–4357, 2016. 27

Joseph Bonneau, Elie Bursztein, Ilan Caron, Rob Jackson, and Mike Williamson. Secrets, lies, and account recovery: Lessons from the use of personal knowledge questions at google. In Proceedings of the 24th International Conference on World Wide Web, WWW 2015, Florence, Italy, May 18-22, 2015, pages 141–150. ACM, 2015. 72

Danah Boyd. Facebook's privacy trainwreck: Exposure, invasion, and social convergence. Convergence, 14(1):13–20, 2008. 17, 58

Justin Brickell and Vitaly Shmatikov. The cost of privacy: destruction of data-mining utility in anonymized data publishing. In Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Las Vegas, Nevada, USA, August 24-27, 2008, pages 70–78, 2008. 18

Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. Semantics derived automatically from language corpora contain human-like biases. Science, 356(6334):183–186, 2017. 26

Ryan Calo and Alex Rosenblat. The taking economy: Uber, infrmation, and power. Columbia Law Review, 117:1623, 2017. 29, 32

Claudio Carpineto and Giovanni Romano. K_Θ-affinity privacy: Releasing infrequent query refinements safely. Information Processing & Management, 51(2):74–88, 2015. 77

Ben Carterette and Rosie Jones. Evaluating search engines by modeling the relationship between relevance and clicks. In Advances in Neural Information Processing Systems 20:

Proceedings of the Twenty-First Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December 3-6, 2007, pages 217–224. Curran Associates, Inc., 2007. 64

L. Elisa Celis, Damian Straszak, and Nisheeth K. Vishnoi. Ranking with fairness constraints. In 45th International Colloquium on Automata, Languages, and Programming, ICALP 2018, July 9-13, 2018, Prague, Czech Republic, pages 28:1–28:15, 2018. 23, 27, 33, 36, 54

Abhijnan Chakraborty, Asia J. Biega, Aniko Hannak, and Krishna P. Gummadi. Fair sharing for sharing economy platforms. In Proceedings of the FATREC@RecSys Workshop, 2017. 7, 29, 55

Bee-Chung Chen, Daniel Kifer, Kristen LeFevre, and Ashwin Machanavajjhala. Privacy-preserving data publishing. Foundations and Trends in Databases, 2(1-2):1–167, 2009. 98

Gang Chen, He Bai, Lidan Shou, Ke Chen, and Yunjun Gao. UPS: efficient privacy protection in personalized web search. In Proceeding of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2011, Beijing, China, July 25-29, 2011, pages 615–624, 2011. 16, 18, 72, 98

Le Chen, Alan Mislove, and Christo Wilson. Peeking beneath the hood of uber. In Proceedings of the 2015 ACM Internet Measurement Conference, IMC 2015, Tokyo, Japan, October 28-30, 2015, pages 495–508, 2015. 27

Le Chen, Ruijun Ma, Anikó Hannák, and Christo Wilson. Investigating the impact of gender on rank in resume search engines. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, CHI 2018, Montreal, QC, Canada, April 21-26, 2018, page 651, 2018. 27

Lisi Chen and Gao Cong. Diversity-aware top-k publish/subscribe for text stream. In Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data, Melbourne, Victoria, Australia, May 31 - June 4, 2015, pages 347–362. ACM, 2015. 73

Flavio Chierichetti, Ravi Kumar, Silvio Lattanzi, and Sergei Vassilvitskii. Fair clustering through fairlets. In Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA, pages 5036–5044, 2017. 23

Benny Chor, Oded Goldreich, Eyal Kushilevitz, and Madhu Sudan. Private information retrieval. In 36th Annual Symposium on Foundations of Computer Science, Milwaukee, Wisconsin, USA, 23-25 October 1995, pages 41–50. IEEE Computer Society, 1995. 19

Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. Big data, 5(2):153–163, 2017. 23

Aleksandr Chuklin, Ilya Markov, and Maarten de Rijke. Click Models for Web Search. Synthesis Lectures on Information Concepts, Retrieval, and Services. Morgan & Claypool Publishers, 2015. 22, 32, 55

Alissa Cooper. A survey of query log privacy-enhancing techniques from a policy perspective. TWEB, 2(4):19:1–19:27, 2008. 94, 112

Sam Corbett-Davies and Sharad Goel. The measure and mismeasure of fairness: A critical review of fair machine learning. CoRR, abs/1808.00023, 2018. 23

Denzil Correa, Leandro Araújo Silva, Mainack Mondal, Fabrício Benevenuto, and Krishna P. Gummadi. The many shades of anonymity: Characterizing anonymous social media content. In Proceedings of the Ninth International Conference on Web and Social Media, ICWSM 2015, University of Oxford, Oxford, UK, May 26-29, 2015, pages 71–80, 2015. 19, 64, 66

Nick Craswell, Onno Zoeter, Michael J. Taylor, and Bill Ramsey. An experimental comparison of click position-bias models. In Proceedings of the International Conference on Web Search and Web Data Mining, WSDM 2008, Palo Alto, California, USA, February 11-12, 2008, pages 87–94. ACM, 2008. 32, 42, 55

W. Bruce Croft, Donald Metzler, and Trevor Strohman. Search Engines - Information Retrieval in Practice. Pearson Education, 2009. ISBN 978-0-13-136489-9. 106, 107

J. Shane Culpepper, Fernando Diaz, and Mark D. Smucker. Research frontiers in information retrieval: Report from the third strategic workshop on information retrieval in lorne (SWIRL 2018). SIGIR Forum, 52(1):34–90, 2018. 1

Amit Datta, Michael Carl Tschantz, and Anupam Datta. Automated experiments on ad privacy settings. PoPETs, 2015(1):92–112, 2015. 25

Wei-Yen Day and Ninghui Li. Differentially private publishing of high-dimensional data using sensitivity control. In Feng Bao, Steven Miller, Jianying Zhou, and Gail-Joon Ahn, editors, Proceedings of the 10th ACM Symposium on Information, Computer and Communications Security, ASIA CCS '15, Singapore, April 14-17, 2015, pages 451–462. ACM, 2015. 77

Georges Dupret and Benjamin Piwowarski. A user browsing model to predict search engine click data from past observations. In Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2008, Singapore, July 20-24, 2008, pages 331–338. ACM, 2008. 55

Cynthia Dwork. Differential privacy: A survey of results. In Theory and Applications of Models of Computation, 5th International Conference, TAMC 2008, Xi'an, China, April 25-29, 2008. Proceedings, pages 1–19, 2008. 15, 76, 94, 98

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard S. Zemel. Fairness through awareness. In Innovations in Theoretical Computer Science 2012, Cambridge, MA, USA, January 8-10, 2012, pages 214–226, 2012. 24, 33, 35, 54

Peter Eckersley. How unique is your web browser? In Privacy Enhancing Technologies, 10th International Symposium, PETS 2010, Berlin, Germany, July 21-23, 2010. Proceedings, pages 1–18, 2010. 12

Benjamin Edelman, Michael Ostrovsky, and Michael Schwarz. Internet advertising and the generalized second-price auction: Selling billions of dollars worth of keywords. American economic review, 97(1):242–259, 2007. 55

Carsten Eickhoff. Cognitive biases in crowdsourcing. In Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining, WSDM 2018, Marina Del Rey, CA, USA, February 5-9, 2018, pages 162–170. ACM, 2018. 26

Michael D. Ekstrand, Mucun Tian, Ion Madrazo Azpiazu, Jennifer D. Ekstrand, Oghenemaro Anuyah, David McNeill, and Maria Soledad Pera. All the cool kids, how do they fit in?: Popularity and demographic biases in recommender evaluation and effectiveness. In Conference on Fairness, Accountability and Transparency, FAT 2018, 23-24 February 2018, New York, NY, USA, volume 81 of Proceedings of Machine Learning Research, pages 172–186. PMLR, 2018a. 23

Michael D. Ekstrand, Mucun Tian, Mohammed R. Imran Kazi, Hoda Mehrpouyan, and Daniel Kluver. Exploring author gender in book rating and recommendation. In Proceedings of the 12th ACM Conference on Recommender Systems, RecSys 2018, Vancouver, BC, Canada, October 2-7, 2018, pages 242–250, 2018b. 23

Tobias Emrich, Hans-Peter Kriegel, Peer Kröger, Johannes Niedermayer, Matthias Renz, and Andreas Züfle. On reverse-k-nearest-neighbor joins. GeoInformatica, 19(2):299–330, 2015. 60, 73

Danielle Ensign, Sorelle A. Friedler, Scott Neville, Carlos Scheidegger, and Suresh Venkatasubramanian. Runaway feedback loops in predictive policing. In Conference on Fairness, Accountability and Transparency, FAT 2018, 23-24 February 2018, New York, NY, USA, volume 81 of Proceedings of Machine Learning Research, pages 160–171. PMLR, 2018. 29

Sedigheh Eslami, Asia J. Biega, Rishiraj Saha Roy, and Gerhard Weikum. Privacy of hidden profiles: Utility-preserving profile removal in online forums. In Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, CIKM 2017, Singapore, November 06 - 10, 2017, pages 2063–2066, 2017. 8, 19

Liyue Fan, Luca Bonomi, Li Xiong, and Vaidy S. Sunderam. Monitoring web browsing behavior with differential privacy. In Proceedings of the 23rd International World Wide Web Conference, WWW '14, Seoul, Republic of Korea, April 7-11, 2014, pages 177–188. ACM, 2014. 94, 112

Lujun Fang and Kristen LeFevre. Privacy wizards for social networking sites. In Proceedings of the 19th International Conference on World Wide Web, WWW 2010, Raleigh, North Carolina, USA, April 26-30, 2010, pages 351–360, 2010. 17

Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Sydney, NSW, Australia, August 10-13, 2015, pages 259–268, 2015. 23, 24, 33, 54, 72

Benjamin Fish, Jeremy Kun, and Ádám Dániel Lelkes. A confidence-based approach for balancing fairness and accuracy. In Proceedings of the 2016 SIAM International Conference on Data Mining, Miami, Florida, USA, May 5-7, 2016, pages 144–152. SIAM, 2016. 25

Sorelle A. Friedler, Carlos Scheidegger, Suresh Venkatasubramanian, Sonam Choudhary, Evan P. Hamilton, and Derek Roth. A comparative study of fairness-enhancing interventions in machine learning. CoRR, abs/1802.04422, 2018. 24

Batya Friedman and Helen Nissenbaum. Bias in computer systems. ACM Transactions on Information Systems, 14(3):330–347, 1996. 27

Norbert Fuhr, Anastasia Giachanou, Gregory Grefenstette, Iryna Gurevych, Andreas Hanselowski, Kalervo Järvelin, Rosie Jones, Yiqun Liu, Josiane Mothe, Wolfgang Nejdl, Isabella Peters, and Benno Stein. An information nutritional label for online documents. SIGIR Forum, 51(3):46–66, 2017. 28

Benjamin C. M. Fung, Ke Wang, Rui Chen, and Philip S. Yu. Privacy-preserving data publishing: A survey of recent developments. ACM Computing Surveys, 42(4):14:1–14:53, 2010. 76, 94, 98

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna M. Wallach, Hal Daumé III, and Kate Crawford. Datasheets for datasets. CoRR, abs/1803.09010, 2018. 26

Arthur Gervais, Reza Shokri, Adish Singla, Srdjan Capkun, and Vincent Lenders. Quantifying web-search privacy. In Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security, Scottsdale, AZ, USA, November 3-7, 2014, pages 966–977, 2014. 19, 72, 98

Ali Ghodsi, Matei Zaharia, Benjamin Hindman, Andy Konwinski, Scott Shenker, and Ion Stoica. Dominant resource fairness: Fair allocation of multiple resource types. In Proceedings of the 8th USENIX Symposium on Networked Systems Design and Implementation, NSDI 2011, Boston, MA, USA, March 30 - April 1, 2011. USENIX Association, 2011. 55

Ali Ghodsi, Vyas Sekar, Matei Zaharia, and Ion Stoica. Multi-resource fair queueing for packet processing. In ACM SIGCOMM 2012 Conference, SIGCOMM '12, Helsinki, Finland - August 13 - 17, 2012, pages 1–12. ACM, 2012. 55

Oana Goga, Howard Lei, Sree Hari Krishnan Parthasarathi, Gerald Friedland, Robin Sommer, and Renata Teixeira. Exploiting innocuous activity for correlating users across sites. In Proceedings of the 22nd International World Wide Web Conference, WWW '13, Rio de Janeiro, Brazil, May 13-17, 2013, pages 447–458. International World Wide Web Conferences Steering Committee / ACM, 2013. 94

Oana Goga, Patrick Loiseau, Robin Sommer, Renata Teixeira, and Krishna P. Gummadi. On the reliability of profile matching across large online social networks. In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Sydney, NSW, Australia, August 10-13, 2015, pages 1799–1808, 2015. 13

Michael T. Goodrich, Michael Mitzenmacher, Olga Ohrimenko, and Roberto Tamassia. Privacy-preserving group data access via stateless oblivious RAM simulation. In Proceedings of the Twenty-Third Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2012, Kyoto, Japan, January 17-19, 2012, pages 157–167. SIAM, 2012. 99

Michaela Götz, Ashwin Machanavajjhala, Guozhang Wang, Xiaokui Xiao, and Johannes Gehrke. Publishing search logs - A comparative study of privacy guarantees. IEEE Transactions on Knowledge and Data Engineering, 24(3):520–532, 2012. 18, 72, 94, 112

Jerald Greenberg. A taxonomy of organizational justice theories. Academy of Management review, 12(1):9–22, 1987. 33

Nina Grgic-Hlaca, Elissa M. Redmiles, Krishna P. Gummadi, and Adrian Weller. Human perceptions of fairness in algorithmic decision making: A case study of criminal risk prediction. In Proceedings of the 2018 World Wide Web Conference on World Wide Web, WWW 2018, Lyon, France, April 23-27, 2018, pages 903–912, 2018a. 28

Nina Grgic-Hlaca, Muhammad Bilal Zafar, Krishna P. Gummadi, and Adrian Weller. Beyond distributive fairness in algorithmic decision making: Feature selection for procedurally fair learning. In Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018, pages 51–60, 2018b. 28, 54

Rachid Guerraoui, Anne-Marie Kermarrec, Rhicheek Patra, and Mahsa Taziki. D2P: distance-based differential privacy in recommenders. PVLDB, 8(8):862–873, 2015. 113

Fan Guo, Chao Liu, and Yi Min Wang. Efficient multiple-click models in web search. In Proceedings of the Second International Conference on Web Search and Web Data Mining, WSDM 2009, Barcelona, Spain, February 9-11, 2009, pages 124–131. ACM, 2009. 55

Sara Hajian and Josep Domingo-Ferrer. A methodology for direct and indirect discrimination prevention in data mining. IEEE Transactions on Knowledge and Data Engineering, 25 (7):1445–1459, 2013. 24

Maria Halkidi and Iordanis Koutsopoulos. A game theoretic framework for data privacy preservation in recommender systems. In Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2011, Athens, Greece, September 5-9, 2011. Proceedings, Part I, volume 6911 of Lecture Notes in Computer Science, pages 629–644. Springer, 2011. 112

Aniko Hannak, Piotr Sapiezynski, Arash Molavi Kakhki, Balachander Krishnamurthy, David Lazer, Alan Mislove, and Christo Wilson. Measuring personalization of web search. In 22nd International World Wide Web Conference, WWW '13, Rio de Janeiro, Brazil, May 13-17, 2013, pages 527–538. International World Wide Web Conferences Steering Committee / ACM, 2013. 98

Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain, pages 3315–3323, 2016. 23, 25, 33, 54

Xi He, Ashwin Machanavajjhala, and Bolin Ding. Blowfish privacy: tuning privacy-utility trade-offs using policies. In International Conference on Management of Data, SIGMOD 2014, Snowbird, UT, USA, June 22-27, 2014, pages 1447–1458. ACM, 2014. 98

Yuan Hong, Jaideep Vaidya, Haibing Lu, and Mingrui Wu. Differentially private search log sanitization with optimal output utility. In 15th International Conference on Extending Database Technology, EDBT '12, Berlin, Germany, March 27-30, 2012, Proceedings, pages 50–61. ACM, 2012. 94

Bill Howe, Julia Stoyanovich, Haoyue Ping, Bernease Herman, and Matt Gee. Synthetic data for social good. CoRR, abs/1710.08874, 2017. 14

Daniel C Howe and Helen Nissenbaum. Trackmenot: Resisting surveillance in web search. Lessons from the Identity trail: Anonymity, privacy, and identity in a networked society, 23:417–436, 2009. 16, 18

Jian Hu, Hua-Jun Zeng, Hua Li, Cheng Niu, and Zheng Chen. Demographic prediction based on user's browsing behavior. In Proceedings of the 16th International Conference on World Wide Web, WWW 2007, Banff, Alberta, Canada, May 8-12, 2007, pages 151–160. ACM, 2007. 94

Danny Yuxing Huang, Doug Grundman, Kurt Thomas, Abhishek Kumar, Elie Bursztein, Kirill Levchenko, and Alex C. Snoeren. Pinning down abuse on google maps. In Proceedings of the 26th International Conference on World Wide Web, WWW 2017, Perth, Australia, April 3-7, 2017, pages 1471–1479. ACM, 2017. 72

Thorsten Joachims. Training linear svms in linear time. In Proceedings of the Twelfth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Philadelphia, PA, USA, August 20-23, 2006, pages 217–226. ACM, 2006. doi: 10.1145/1150402.1150429. 61, 67

Thorsten Joachims and Filip Radlinski. Search engines that learn from implicit feedback. IEEE Computer, 40(8):34–40, 2007. 32, 43, 55

Thorsten Joachims, Laura A. Granka, Bing Pan, Helene Hembrooke, and Geri Gay. Accurately interpreting clickthrough data as implicit feedback. In SIGIR 2005: Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Salvador, Brazil, August 15-19, 2005, pages 154–161, 2005. 26

Thorsten Joachims, Adith Swaminathan, and Tobias Schnabel. Unbiased learning-to-rank with biased feedback. In Proceedings of the Tenth ACM International Conference on Web Search and Data Mining, WSDM 2017, Cambridge, United Kingdom, February 6-10, 2017, pages 781–789. ACM, 2017. 55

Rosie Jones, Ravi Kumar, Bo Pang, and Andrew Tomkins. "i know what you did last summer": query logs and user privacy. In Proceedings of the Sixteenth ACM Conference on Information and Knowledge Management, CIKM 2007, Lisbon, Portugal, November 6-10, 2007, pages 909–914, 2007. 19, 94

Rosie Jones, Ravi Kumar, Bo Pang, and Andrew Tomkins. Vanity fair: privacy in querylog bundles. In Proceedings of the 17th ACM Conference on Information and Knowledge Management, CIKM 2008, Napa Valley, California, USA, October 26-30, 2008, pages 853–862. ACM, 2008. 111

Roberto J. Bayardo Jr. and Rakesh Agrawal. Data privacy through optimal k-anonymization. In Proceedings of the 21st International Conference on Data Engineering, ICDE 2005, 5-8 April 2005, Tokyo, Japan, pages 217–228, 2005. 14

Pawel Jurczyk and Eugene Agichtein. Discovering authorities in question answer communities by using link analysis. In Proceedings of the Sixteenth ACM Conference on Information and Knowledge Management, CIKM 2007, Lisbon, Portugal, November 6-10, 2007, pages 919–922. ACM, 2007. 96

Gary Kacmarcik and Michael Gamon. Obfuscating document stylometry to preserve author anonymity. In ACL 2006, 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, Sydney, Australia, 17-21 July 2006, 2006. 15

Faisal Kamiran, Toon Calders, and Mykola Pechenizkiy. Discrimination aware decision tree learning. In ICDM 2010, The 10th IEEE International Conference on Data Mining, Sydney, Australia, 14-17 December 2010, pages 869–874. IEEE Computer Society, 2010. doi: 10.1109/ICDM.2010.50. 25

Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. Fairness-aware classifier with prejudice remover regularizer. In Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2012, Bristol, UK, September 24-28, 2012. Proceedings, Part II, pages 35–50, 2012. 25, 33, 54

Sampath Kannan, Michael J. Kearns, Jamie Morgenstern, Mallesh M. Pai, Aaron Roth, Rakesh V. Vohra, and Zhiwei Steven Wu. Fairness incentives for myopic agents. In Proceedings of the 2017 ACM Conference on Economics and Computation, EC '17, Cambridge, MA, USA, June 26-30, 2017, pages 369–386. ACM, 2017. 29

Michael J. Kearns, Aaron Roth, and Zhiwei Steven Wu. Meritocratic fairness for cross-population selection. In Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017, pages 1828–1836, 2017. 24, 33, 54

Niki Kilbertus, Adrià Gascón, Matt J. Kusner, Michael Veale, Krishna P. Gummadi, and Adrian Weller. Blind justice: Fairness with encrypted sensitive attributes. In Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018, pages 2635–2644, 2018. 25

Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. Human decisions and machine predictions. The Quarterly Journal of Economics, 133(1): 237–293, 2017a. 28, 54

Jon M. Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. In 8th Innovations in Theoretical Computer Science Conference, ITCS 2017, January 9-11, 2017, Berkeley, CA, USA, volume 67 of LIPIcs, pages 43:1–43:23. Schloss Dagstuhl - Leibniz-Zentrum fuer Informatik, 2017b. 23

Aleksandra Korolova, Krishnaram Kenthapadi, Nina Mishra, and Alexandros Ntoulas. Releasing search queries and clicks privately. In Proceedings of the 18th International Conference on World Wide Web, WWW 2009, Madrid, Spain, April 20-24, 2009, pages 171–180. ACM, 2009. 94

Michal Kosinski, David Stillwell, and Thore Graepel. Private traits and attributes are predictable from digital records of human behavior. Proceedings of the National Academy of Sciences, 110(15):5802–5805, 2013. ISSN 0027-8424. doi: 10.1073/pnas.1218772110. 3, 13

Andreas Krause and Eric Horvitz. A utility-theoretic approach to privacy in online services. Journal of Artificial Intelligence Research, 39:633–662, 2010. 98

Joshua A Kroll, Solon Barocas, Edward W Felten, Joel R Reidenberg, David G Robinson, and Harlan Yu. Accountable algorithms. University of Pennsylvania Law Review, 165: 633, 2016. 25

Juhi Kulshrestha, Motahhare Eslami, Johnnatan Messias, Muhammad Bilal Zafar, Saptarshi Ghosh, Krishna P. Gummadi, and Karrie Karahalios. Quantifying search bias: Investigating sources of bias for political searches in social media. In Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing, CSCW 2017, Portland, OR, USA, February 25 - March 1, 2017, pages 417–432, 2017. 27

Ravi Kumar, Jasmine Novak, Bo Pang, and Andrew Tomkins. On anonymizing query logs via token-based hashing. In Proceedings of the 16th International Conference on World Wide Web, WWW 2007, Banff, Alberta, Canada, May 8-12, 2007, pages 629–638. ACM, 2007. 94

Preethi Lahoti, Gerhard Weikum, and Krishna P. Gummadi. ifair: Learning individually fair data representations for algorithmic decision making. CoRR, abs/1806.01059, 2018. 24

Alex Leavitt. "this is a throwaway account": Temporary technical identities and perceptions of anonymity in a massive online community. In Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing, CSCW 2015, Vancouver, BC, Canada, March 14 - 18, 2015, pages 317–327, 2015. 20

Kristen LeFevre, David J. DeWitt, and Raghu Ramakrishnan. Incognito: Efficient full-domain k-anonymity. In Proceedings of the ACM SIGMOD International Conference on Management of Data, Baltimore, Maryland, USA, June 14-16, 2005, pages 49–60, 2005. 14, 18

Kristen LeFevre, David J. DeWitt, and Raghu Ramakrishnan. Mondrian multidimensional k-anonymity. In Proceedings of the 22nd International Conference on Data Engineering, ICDE 2006, 3-8 April 2006, Atlanta, GA, USA, page 25, 2006. 14

Jurek Leonhardt, Avishek Anand, and Megha Khosla. User fairness in recommender systems. In Companion of the The Web Conference 2018 on The Web Conference 2018, WWW 2018, Lyon , France, April 23-27, 2018, pages 101–102. ACM, 2018. 25

Adam Lerner, Anna Kornfeld Simpson, Tadayoshi Kohno, and Franziska Roesner. Internet jones and the raiders of the lost trackers: An archaeological study of web tracking from 1996 to 2016. In 25th USENIX Security Symposium, USENIX Security 16, Austin, TX, USA, August 10-12, 2016. USENIX Association, 2016. 112

Karen Levy and Solon Barocas. Designing against discrimination in online markets. Berkeley Technology Law Journal, 32:1183, 2017. 32

Chao Li, Daniel Yang Li, Gerome Miklau, and Dan Suciu. A theory of pricing private data. ACM Transactions on Database Systems, 39(4):34:1–34:28, 2014. 20

Ninghui Li, Tiancheng Li, and Suresh Venkatasubramanian. t-closeness: Privacy beyond k-anonymity and l-diversity. In Proceedings of the 23rd International Conference on Data Engineering, ICDE 2007, The Marmara Hotel, Istanbul, Turkey, April 15-20, 2007, pages 106–115, 2007. 13, 14, 18, 76, 94, 98

Ninghui Li, Wahbeh H. Qardaji, Dong Su, Yi Wu, and Weining Yang. Membership privacy: a unifying framework for privacy definitions. In 2013 ACM SIGSAC Conference on Computer and Communications Security, CCS'13, Berlin, Germany, November 4-8, 2013, pages 889–900. ACM, 2013. 76, 94

Tiancheng Li and Ninghui Li. On the tradeoff between privacy and utility in data publishing. In Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Paris, France, June 28 - July 1, 2009, pages 517–526, 2009. 18

Lydia T. Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. Delayed impact of fair machine learning. In Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018, volume 80 of JMLR Workshop and Conference Proceedings, pages 3156–3164. JMLR.org, 2018. 29

Kristian Lum and William Isaac. To predict and serve? Significance, 13(5):14–19, 2016. 29

Ashwin Machanavajjhala, Daniel Kifer, Johannes Gehrke, and Muthuramakrishnan Venkitasubramaniam. L-diversity: Privacy beyond k-anonymity. TKDD, 1(1):3, 2007. 12, 14, 18, 76, 94, 98

Huina Mao, Xin Shuai, and Apu Kapadia. Loose tweets: an analysis of privacy leaks on twitter. In Proceedings of the 10th annual ACM workshop on Privacy in the electronic society, WPES 2011, Chicago, IL, USA, October 17, 2011, pages 1–12, 2011. 13

Rahat Masood, Dinusha Vatsalan, Muhammad Ikram, and Mohamed Ali Kâafar. Incognito: A method for obfuscating web data. In Proceedings of the 2018 World Wide Web Conference on World Wide Web, WWW 2018, Lyon, France, April 23-27, 2018, pages 267–276. ACM, 2018. 16

Alessandra Mazzia, Kristen LeFevre, and Eytan Adar. The pviz comprehension tool for social network privacy settings. In Symposium On Usable Privacy and Security, SOUPS '12, Washington, DC, USA - July 11 - 13, 2012, page 13, 2012. 17

Andrew Kachites McCallum. Mallet: A machine learning for language toolkit. 2002. 89, 117

Andrew W. E. McDonald, Sadia Afroz, Aylin Caliskan, Ariel Stolerman, and Rachel Greenstadt. Use fewer instances of the letter "i": Toward writing style anonymization. In Privacy Enhancing Technologies - 12th International Symposium, PETS 2012, Vigo, Spain, July 11-13, 2012. Proceedings, pages 299–318, 2012. 15

Frank McSherry and Ilya Mironov. Differentially private recommender systems: Building privacy into the netflix prize contenders. In Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Paris, France, June 28 - July 1, 2009, pages 627–636. ACM, 2009. 112

Rishabh Mehrotra, Ashton Anderson, Fernando Diaz, Amit Sharma, Hanna M. Wallach, and Emine Yilmaz. Auditing search engines for differential satisfaction across demographics. In Proceedings of the 26th International Conference on World Wide Web Companion, Perth, Australia, April 3-7, 2017, pages 626–633. ACM, 2017. 4, 23, 28, 55, 72

Rishabh Mehrotra, James McInerney, Hugues Bouchard, Mounia Lalmas, and Fernando Diaz. Towards a fair marketplace: Counterfactual evaluation of the trade-off between relevance, fairness & satisfaction in recommendation systems. In Proceedings of the 27th ACM International Conference on Information and Knowledge Management, CIKM 2018, Torino, Italy, October 22-26, 2018, pages 2243–2251, 2018. 23, 25

Wei Meng, Ren Ding, Simon P. Chung, Steven Han, and Wenke Lee. The price of free: Privacy leakage in personalized mobile in-apps ads. In 23rd Annual Network and Distributed System Security Symposium, NDSS 2016, San Diego, California, USA, February 21-24, 2016. The Internet Society, 2016a. 112

Wei Meng, Byoungyoung Lee, Xinyu Xing, and Wenke Lee. Trackmeornot: Enabling flexible control on web tracking. In Proceedings of the 25th International Conference on World Wide Web, WWW 2016, Montreal, Canada, April 11 - 15, 2016, pages 99–109, 2016b. 13, 16, 112

Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. In Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States., pages 3111–3119, 2013. 77, 82, 89

Mainack Mondal, Peter Druschel, Krishna P Gummadi, and Alan Mislove. Beyond access control: Managing online privacy via exposure. In Proceedings of the Workshop on Useable Security, pages 1–6, 2014. 58, 71

Mainack Mondal, Johnnatan Messias, Saptarshi Ghosh, Krishna P. Gummadi, and Aniket Kate. Forgetting in social media: Understanding and controlling longitudinal exposure of socially shared data. In Twelfth Symposium on Usable Privacy and Security, SOUPS 2016, Denver, CO, USA, June 22-24, 2016, pages 287–299, 2016. 13, 17, 19, 58, 64, 65, 66, 71

Arvind Narayanan and Vitaly Shmatikov. Obfuscated databases and group privacy. In Proceedings of the 12th ACM Conference on Computer and Communications Security, CCS 2005, Alexandria, VA, USA, November 7-11, 2005, pages 102–111. ACM, 2005. 111

Arvind Narayanan and Vitaly Shmatikov. Robust de-anonymization of large sparse datasets. In 2008 IEEE Symposium on Security and Privacy (S&P 2008), 18-21 May 2008, Oakland, California, USA, pages 111–125, 2008. 3, 12

Arvind Narayanan and Vitaly Shmatikov. De-anonymizing social networks. In 30th IEEE Symposium on Security and Privacy (S&P 2009), 17-20 May 2009, Oakland, California, USA, pages 173–187, 2009. 12, 94

Arvind Narayanan, Hristo S. Paskov, Neil Zhenqiang Gong, John Bethencourt, Emil Stefanov, Eui Chul Richard Shin, and Dawn Song. On the feasibility of internet-scale author identification. In IEEE Symposium on Security and Privacy, SP 2012, 21-23 May 2012, San Francisco, California, USA, pages 300–314, 2012. 13, 94

Guillermo Navarro-Arribas, Vicenç Torra, Arnau Erola, and Jordi Castellà-Roca. User k-anonymity for privacy preserving data mining of query logs. Information Processing & Management, 48(3):476–487, 2012. 77

Valeria Nikolaenko, Stratis Ioannidis, Udi Weinsberg, Marc Joye, Nina Taft, and Dan Boneh. Privacy-preserving matrix factorization. In 2013 ACM SIGSAC Conference on Computer and Communications Security, CCS'13, Berlin, Germany, November 4-8, 2013, pages 801–812. ACM, 2013. 112

Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Emre Kiciman. Social data: Biases, methodological pitfalls, and ethical boundaries. Available at SSRN: https://ssrn.com/abstract=2886526, 2016. 26, 27

HweeHwa Pang, Xiaokui Xiao, and Jialie Shen. Obfuscating the topical intention in enterprise text search. In IEEE 28th International Conference on Data Engineering (ICDE 2012),

Washington, DC, USA (Arlington, Virginia), 1-5 April, 2012, pages 1168–1179, 2012. 16, 18, 94

Sai Teja Peddinti and Nitesh Saxena. On the privacy of web search based on query obfuscation: A case study of trackmenot. In Privacy Enhancing Technologies, 10th International Symposium, PETS 2010, Berlin, Germany, July 21-23, 2010. Proceedings, pages 19–37, 2010. 19

Sai Teja Peddinti and Nitesh Saxena. Web search query privacy: Evaluating query obfuscation and anonymizing networks. Journal of Computer Security, 22(1):155–199, 2014. 98, 112

Sai Teja Peddinti, Aleksandra Korolova, Elie Bursztein, and Geetanjali Sampemane. Cloak and swagger: Understanding data sensitivity through the lens of user anonymity. In 2014 IEEE Symposium on Security and Privacy, SP 2014, Berkeley, CA, USA, May 18-21, 2014, pages 493–508, 2014. 19, 77, 78, 94

Dino Pedreschi, Salvatore Ruggieri, and Franco Turini. Discrimination-aware data mining. In Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Las Vegas, Nevada, USA, August 24-27, 2008, pages 560–568, 2008. 24, 33, 54

Marco Pennacchiotti and Ana-Maria Popescu. A machine learning approach to twitter user classification. In Proceedings of the Fifth International Conference on Weblogs and Social Media, Barcelona, Catalonia, Spain, July 17-21, 2011. The AAAI Press, 2011. 96

Forough Poursabzi-Sangdeh, Daniel G. Goldstein, Jake M. Hofman, Jennifer Wortman Vaughan, and Hanna M. Wallach. Manipulating and measuring model interpretability. CoRR, abs/1802.07810, 2018. 28

David Rebollo-Monedero, Jordi Forné, and Josep Domingo-Ferrer. Query profile obfuscation by means of optimal query exchange between users. IEEE Transactions on Dependable and Secure Computing, 9(5):641–654, 2012. 99, 112

Michael K. Reiter and Aviel D. Rubin. Crowds: Anonymity for web transactions. ACM Transactions on Information and System Security, 1(1):66–92, 1998. 16, 99, 111

Marian-Andrei Rizoiu, Lexing Xie, Tibério S. Caetano, and Manuel Cebrián. Evolution of privacy loss in wikipedia. In Proceedings of the Ninth ACM International Conference on Web Search and Data Mining, San Francisco, CA, USA, February 22-25, 2016, pages 215–224, 2016. 19

Ronald E. Robertson, David Lazer, and Christo Wilson. Auditing the personalization and composition of politically-related search engine results pages. In Proceedings of the 2018 World Wide Web Conference on World Wide Web, WWW 2018, Lyon, France, April 23-27, 2018, pages 955–965. ACM, 2018. 27

Andrea Romei and Salvatore Ruggieri. A multidisciplinary survey on discrimination analysis. Knowledge Eng. Review, 29(5):582–638, 2014. 54

Alex Rosenblat and Luke Stark. Algorithmic labor and information asymmetries: A case study of uber's drivers. International Journal of Communication, 10:27, 2016. 29

Ravi S. Sandhu, Edward J. Coyne, Hal L. Feinstein, and Charles E. Youman. Role-based access control models. IEEE Computer, 29(2):38–47, 1996. 13, 16

Nuno Santos, Alan Mislove, Marcel Dischinger, and Krishna Gummadi. Anonymity in the personalized web. In NSDI Posters '08, 2008. 99

Roman Schlegel, Apu Kapadia, and Adam J. Lee. Eyeing your exposure: quantifying and controlling information sharing for improved privacy. In Symposium On Usable Privacy and Security, SOUPS '11, Pittsburgh, PA, USA - July 20 - 22, 2011, page 14. ACM, 2011. 17, 58, 71

Xuehua Shen, Bin Tan, and ChengXiang Zhai. Privacy protection in personalized search. SIGIR Forum, 41(1):4–17, 2007. 72, 112

Yilin Shen and Hongxia Jin. Epicrec: Towards practical differentially private framework for personalized recommendation. In Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, Vienna, Austria, October 24-28, 2016, pages 180–191. ACM, 2016. 112

Reza Shokri, George Theodorakopoulos, George Danezis, Jean-Pierre Hubaux, and Jean-Yves Le Boudec. Quantifying location privacy: The case of sporadic location exposure. In Privacy Enhancing Technologies - 11th International Symposium, PETS 2011, Waterloo, ON, Canada, July 27-29, 2011. Proceedings, volume 6794 of Lecture Notes in Computer Science, pages 57–76. Springer, 2011. 58, 71

Ashudeep Singh and Thorsten Joachims. Fairness of exposure in rankings. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2018, London, UK, August 19-23, 2018, pages 2219–2228, 2018. 5, 23, 25, 28, 33, 36, 54

Adish Singla, Eric Horvitz, Ece Kamar, and Ryen White. Stochastic privacy. In Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence, July 27 -31, 2014, Québec City, Québec, Canada., pages 152–158, 2014. 13, 15, 18, 94, 112

Trevor Strohman, Donald Metzler, Howard Turtle, and W Bruce Croft. Indri: A language model-based search engine for complex queries. In Proceedings of the International Conference on Intelligent Analysis, volume 2, pages 2–6, 2005. 107

Latanya Sweeney. Achieving k-anonymity privacy protection using generalization and suppression. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 10(5):571–588, 2002a. 14

Latanya Sweeney. k-anonymity: A model for protecting privacy. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 10(5):557–570, 2002b. 12, 14, 76, 94, 98

Latanya Sweeney. Discrimination in online ad delivery. Communications of the ACM, 56 (5):44–54, 2013. 25

Yla R Tausczik and James W Pennebaker. The psychological meaning of words: Liwc and computerized text analysis methods. Journal of language and social psychology, 29(1): 24–54, 2010. 62

Jaime Teevan, Susan T. Dumais, and Eric Horvitz. Personalizing search via automated analysis of interests and activities. In In Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2005, Salvador, Brazil, August 15-19, 2005, pages 449–456. ACM, 2005. 98

Vincent Toubiana, Arvind Narayanan, Dan Boneh, Helen Nissenbaum, and Solon Barocas. Adnostic: Privacy preserving targeted advertising. In Proceedings of the Network and Distributed System Security Symposium, NDSS 2010, San Diego, California, USA, 28th February - 3rd March 2010, 2010. 16

Blase Ur, Pedro Giovanni Leon, Lorrie Faith Cranor, Richard Shay, and Yang Wang. Smart, useful, scary, creepy: perceptions of online behavioral advertising. In Symposium On Usable Privacy and Security, SOUPS '12, Washington, DC, USA - July 11 - 13, 2012, page 4. ACM, 2012. 20

Isabel Valera, Adish Singla, and Manuel Gomez Rodriguez. Enhancing the accuracy and fairness of human decision making. In Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montréal, Canada., pages 1774–1783, 2018. 28

Dinusha Vatsalan, Peter Christen, and Vassilios S. Verykios. A taxonomy of privacy-preserving record linkage techniques. Information Systems, 38(6):946–969, 2013. 94

Akrivi Vlachou, Christos Doulkeridis, Yannis Kotidis, and Kjetil Nørvåg. Monochromatic and bichromatic reverse top-k queries. IEEE Transactions on Knowledge and Data Engineering, 23(8):1215–1229, 2011. 73

Jilles Vreeken, Matthijs van Leeuwen, and Arno Siebes. Preserving privacy through data generation. In Proceedings of the 7th IEEE International Conference on Data Mining (ICDM 2007), October 28-31, 2007, Omaha, Nebraska, USA, pages 685–690, 2007. 14

Elaine Walster, Ellen Berscheid, and G William Walster. New directions in equity research. Journal of personality and social psychology, 25(2):151, 1973. 33, 35

Peng Wang and Chinya V. Ravishankar. On masking topical intent in keyword search. In IEEE 30th International Conference on Data Engineering, Chicago, ICDE 2014, IL, USA, March 31 - April 4, 2014, pages 256–267, 2014. 16, 18, 72, 112

Xuanhui Wang, Michael Bendersky, Donald Metzler, and Marc Najork. Learning to rank with selection bias in personal search. In Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval, SIGIR 2016, Pisa, Italy, July 17-21, 2016, pages 115–124. ACM, 2016. 55

Ingmar Weber and Carlos Castillo. The demographics of web search. In Proceeding of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2010, Geneva, Switzerland, July 19-23, 2010, pages 523–530. ACM, 2010. 94

Yongkai Wu, Lu Zhang, and Xintao Wu. On discrimination discovery and removal in ranked data using causal graph. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2018, London, UK, August 19-23, 2018, pages 2536–2544, 2018. 28

Yabo Xu, Ke Wang, Benyu Zhang, and Zheng Chen. Privacy-enhancing personalized web search. In Proceedings of the 16th International Conference on World Wide Web, WWW 2007, Banff, Alberta, Canada, May 8-12, 2007, pages 591–600, 2007. 16, 18, 72, 112

Yabo Xu, Ke Wang, Guoliang Yang, and Ada Wai-Chee Fu. Online anonymity for personalized web services. In Proceedings of the 18th ACM Conference on Information and Knowledge Management, CIKM 2009, Hong Kong, China, November 2-6, 2009, pages 1497–1500. ACM, 2009. 99

Menahem E Yaari and Maya Bar-Hillel. On dividing justly. Social choice and welfare, 1(1): 1–24, 1984. 35

Grace Hui Yang, Ian Soboroff, Li Xiong, Charles L. A. Clarke, and Simson L. Garfinkel. Privacy-preserving IR 2016: Differential privacy, search, and social media. In Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval, SIGIR 2016, Pisa, Italy, July 17-21, 2016, pages 1247–1248. ACM, 2016. 112

Ke Yang and Julia Stoyanovich. Measuring fairness in ranked outputs. In Proceedings of the 29th International Conference on Scientific and Statistical Database Management, Chicago, IL, USA, June 27-29, 2017, pages 22:1–22:6. ACM, 2017. 23, 28, 33, 36, 54

Ke Yang, Julia Stoyanovich, Abolfazl Asudeh, Bill Howe, H. V. Jagadish, and Gerome Miklau. A nutritional label for rankings. In Proceedings of the 2018 International Conference on Management of Data, SIGMOD Conference 2018, Houston, TX, USA, June 10-15, 2018, pages 1773–1776. ACM, 2018. 28

Sirui Yao and Bert Huang. Beyond parity: Fairness objectives for collaborative filtering. In Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA, pages 2925–2934, 2017. 23, 25

Sergey Yekhanin. Private information retrieval. Commununications of the ACM, 53(4): 68–73, 2010. 94

Zhonghao Yu, Sam Macbeth, Konark Modi, and Josep M. Pujol. Tracking the trackers. In Proceedings of the 25th International Conference on World Wide Web, WWW 2016, Montreal, Canada, April 11 - 15, 2016, pages 121–132, 2016. 13, 16, 112

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez-Rodriguez, and Krishna P. Gummadi. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In Proceedings of the 26th International Conference on World Wide Web, WWW 2017, Perth, Australia, April 3-7, 2017, pages 1171–1180, 2017. 23, 25, 33, 54

Meike Zehlike and Carlos Castillo. Reducing disparate exposure in ranking: A learning to rank approach. CoRR, abs/1805.08716, 2018. 28

Meike Zehlike, Francesco Bonchi, Carlos Castillo, Sara Hajian, Mohamed Megahed, and Ricardo A. Baeza-Yates. Fa*ir: A fair top-k ranking algorithm. In Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, CIKM 2017, Singapore, November 06 - 10, 2017, pages 1569–1578. ACM, 2017. 5, 23, 25, 27, 33, 36, 54

Richard S. Zemel, Yu Wu, Kevin Swersky, Toniann Pitassi, and Cynthia Dwork. Learning fair representations. In Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013, volume 28 of JMLR Workshop and Conference Proceedings, pages 325–333. JMLR.org, 2013. 24, 33, 54

Aston Zhang, Xing Xie, Kevin Chen-Chuan Chang, Carl A. Gunter, Jiawei Han, and XiaoFeng Wang. Privacy risk in anonymized heterogeneous information networks. In Proceedings of the 17th International Conference on Extending Database Technology, EDBT 2014, Athens, Greece, March 24-28, 2014., pages 595–606. OpenProceedings.org, 2014. 94

Jun Zhang, Mark S. Ackerman, and Lada A. Adamic. Expertise networks in online communities: structure and algorithms. In Proceedings of the 16th International Conference on World Wide Web, WWW 2007, Banff, Alberta, Canada, May 8-12, 2007, pages 221–230. ACM, 2007. 96

Sicong Zhang and Grace Hui Yang. Deriving differentially private session logs for query suggestion. In Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval, ICTIR 2017, Amsterdam, The Netherlands, October 1-4, 2017, pages 51–58, 2017. 18

Sicong Zhang, Grace 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 2016, Pisa, Italy, July 17-21, 2016, pages 753–756. ACM, 2016a. 72, 112

Sicong Zhang, Grace Hui Yang, Lisa Singh, and Li Xiong. Safelog: Supporting web search and mining by differentially-private query logs. In 2016 AAAI Fall Symposia, Arlington, Virginia, USA, November 17-19, 2016, 2016b. 18

Yang Zhang, Mathias Humbert, Tahleen Rahman, Cheng-Te Li, Jun Pang, and Michael Backes. Tagvisor: A privacy advisor for sharing hashtags. In Proceedings of the 2018 Conference on World Wide Web, WWW 2018, Lyon, France, April 23-27, 2018, pages 287–296, 2018. 13, 15

Elena Zheleva and Lise Getoor. To join or not to join: the illusion of privacy in social networks with mixed public and private user profiles. In Proceedings of the 18th International Conference on World Wide Web, WWW 2009, Madrid, Spain, April 20-24, 2009, pages 531–540, 2009. 15

Yun Zhu, Li Xiong, and Christopher Verdery. Anonymizing user profiles for personalized web search. In Proceedings of the 19th International Conference on World Wide Web, WWW 2010, Raleigh, North Carolina, USA, April 26-30, 2010, pages 1225–1226, 2010. 18, 72, 111

List of Figures

List of Tables