TU WIEN Informatics

Personalized Advertisements: Current Practices and Student Awareness

BACHELORARBEIT

zur Erlangung des akademischen Grades

Bachelor of Science

im Rahmen des Studiums

Medieninformatik und Visual Computing

eingereicht von

Wassily Bartuska
Matrikelnummer 01427303

an der Fakultät für Informatik

der Technischen Universität Wien

Betreuung: Assoc. Prof. Dr. Dipl.-Ing. Hilda Tellioğlu

Wien, 5. Februar 2020

Wassily Bartuska	Hilda Tellioğlu

TU WIEN Informatics

Personalized Advertisements: Current Practices and Student Awareness

BACHELOR'S THESIS

submitted in partial fulfillment of the requirements for the degree of

Bachelor of Science

in

Media Informatics and Visual Computing

by

Wassily Bartuska
Registration Number 01427303

to the Faculty of Informatics

at the TU Wien

Advisor: Assoc. Prof. Dr. Dipl.-Ing. Hilda Tellioğlu

Vienna, 5th February, 2020

Wassily Bartuska	Hilda Tellioğlu

Erklärung zur Verfassung der Arbeit

Wassily Bartuska

Hiermit erkläre ich, dass ich diese Arbeit selbständig verfasst habe, dass ich die verwendeten Quellen und Hilfsmittel vollständig angegeben habe und dass ich die Stellen der Arbeit – einschließlich Tabellen, Karten und Abbildungen –, die anderen Werken oder dem Internet im Wortlaut oder dem Sinn nach entnommen sind, auf jeden Fall unter Angabe der Quelle als Entlehnung kenntlich gemacht habe.

Wien, 5. Februar 2020

Wassily Bartuska

Danksagung

Ich danke allen TeilnehmerInnen, die interessiert mit der Studie umgegangen sind und sogar nach der Durchführung noch interessante Fragen gestellt haben. Außerdem danke ich meinen KollegInnen, Freunden und Familie, die mich durch Feedback in der Entwicklung der Webseiten unterstützt haben. Weiters danke ich meiner Betreuerin für Feedback-Termine, die in Person abgehalten wurden, da mir diese Termine sehr hilfreich waren.

Acknowledgements

I thank all participants, who were curious about the study and even asked intriguing questions after the procedure was over. Additionally, I thank my colleagues, friends and family for supporting me through feedback while I developed the websites. I thank my supervisor as well, for organizing meetings for feedback, that were quite helpful.

Kurzfassung

Diese Bachelorarbeit befasst sich mit dem Thema der personalisierten Werbung im Internet, das ein Unterthema des Bereichs der Algorithmic Transparency darstellt. Zuerst wird allgemein beschrieben, warum sich mit Algorithmic Transparency befasst wird, dann werden wichtige Begriffe aus diesem Bereich erklärt. Anschließend werden Implementierung möglicher Systeme für personalisierte Werbung beschrieben, zwei ausgewählte Studien der personalisierten Werbung zusammengefasst und zuletzt eine eigene Studie, durchgeführt mit Studierenden dreier Universitäten in Wien, präsentiert.

Abstract

This bachelor's thesis is about personalized advertisements on the Internet. This field poses a subarea of the research related to algorithmic transparency. At first, a general overview of reasons research on algorithmic transparency is conducted for is given, then important terms of this field of research are explained. Subsequently, implementations for possible personalized advertisement systems are described, two selected studies of personalized advertisements are outlined, and finally, a study conducted with students from three Viennese universities is presented.

Contents

CHAPTER 1

Introduction

1.1 State of the Art

Nowadays complex algorithms are omnipresent in our daily lives. Almost everyone constantly has a smart-phone in their pocket. Even people without smart-phones use complex technology without genuinely understanding all factors and implications of the algorithms that are at work. The term "algorithmic transparency" describes if systems that use complex algorithms are understandable to average users.

Technology influences a huge variety of different aspects of the modern society. That is why a variety of disciplines are researching algorithmic transparency. This introduction serves as an overview of the research in a selection of disciplines. This serves as the basis for the rest of this thesis that focuses on a subtopic of algorithmic transparency: transparent personalized advertisement algorithms.

1.1.1 Communications

In regards to algorithms, the field of communications inspects the repercussions of search algorithms on political discourse. Many modern search engines and social media filter seemingly uninteresting content out of the information that the user gets. This system is always inherently biased by decision making in the development of the algorithms that are used.

As people see only one point of view for many important topics, discourse with other people becomes complicated because the different viewpoint is not accessible and therefore not comprehensible.

Similar to the works of Mittelstadt et al. [Mit16], The goal of research on algorithmic transparency in communications is to find a way to make filter algorithms transparent, so that users know what type of content they would not see.

1.1.2 Economics

The economical aspect of algorithm transparency includes data that is used commercially. Businesses such as Google or Facebook use a big amount of personal data every day. As described by Kim et al. [TWK17], these businesses often try to stay nontransparent by stating that the ways their algorithms function are business secrets that, if they were made public, would be used by the competition.

Recently the European Union agreed on introducing the General Data Protection Regulation (GDPR). This regulation has the goal to make the use of personal data more transparent.

1.2 Main Questions

Two main questions are presented in this thesis. Both are related to the transparency of complex algorithms used by big companies.

The first question is how big data companies (such as Alphabet, Facebook, Amazon etc.) use personal data of their users to create personalized advertisements that are then shown to the users. This thesis explores different implementations of generating personalized advertisements and how the mechanics underlying those implementations can be made to be more transparent as well as customized by the user.

The second question focuses on awareness. This thesis uses a simple website simulating personalized advertisements in order to assess the awareness regarding personalized advertisements of university students in Vienna. Furthermore, if the aforementioned students can be surprised by such a website is explored in this thesis.

1.3 Expected Result

The results of the theoretical part are expected to be a concise overview of different personalized advertisements implementations that serve to show how web services might be generating advertisements at the moment. Some uncertainty exists however, as web companies make it their priority to keep their actual implementations a secret. The results of the practical part is believed to show that students of technological universities are aware of personalized advertisements on the Internet and can not be surprised by simple websites but that other students might not be aware or can be surprised more easily.

1.4 Methods

The first main question is answered in this thesis by thorough research. Implementations of personalized advertisement solutions are shown. The transparency and data security of those implementations will increase from subsection to subsection.

The second main question is answered with the design of an experiment. This experiment is split into two parts on two different websites: In part one, participants will click through a website selecting products that they prefer over another and will then get an advertisement that is supposedly fitting to them as a person. In part two, participants will answer questions about how they feel regarding this advertisement. Including but not limited to if they are surprised by the recommendation and if they know how other websites generate personalized advertisements.

1.5 Timetable

Task	Hours
Research	100
Meetings	20
Developing Website	100
Conducting Study	35
Writing Experiment Part of Thesis	50

CHAPTER 2

Definitions

The topic of algorithmic transparency combines aspects of a wide variety of different disciplines. This chapter serves as a glossary for some of the terms that are used while researching algorithmic transparency.

2.1 Algorithmic

Theoretical computer scientists have tried to get an accurate definition of an algorithm for decades. That definition is rooted in state machines that can calculate any value of a defined grammar using input values. These state machines are known as Turing-Machines. A narrower definition of an algorithm was proposed in the nineteen-fifties by Gurevich et al. [Gur03]. This definition consists of five parts:

- "An algorithmic process splits into steps whose complexity is bounded in advance, i.e., the bound is independent of the input and the current state of the computation.
- Each step consists of a direct and immediate transformation of the current state.
- This transformation applies only to the active part of the state and does not alter the remainder of the state.
- The size of the active part is bounded in advance.
- The process runs until either the next step is impossible or a signal says the solution has been reached." [Gur03]

Gillespie provided a general definition for algorithms. He wrote "Algorithms need not be software: in the broadest sense, they are encoded procedures for transforming input data into a desired output, based on specified calculations." [Gil13].

2.2 Transparency

Transparency is closely related to openness. Both terms describe making information accessible.

Heald et al. [Hea06] discern the two terms as follows. Openness means a general accessibility of information about a system's operating principle. This information does not have to be made understandable for all stakeholders. Transparency contrasts openness by describing information about a system that is available as well as understandable to all users.

For the purpose of this thesis only transparency will be used. The term will describe information that is actually understood by the audience that is using the technology. Transparent implementations for algorithms have a guiding aspect to them, that lets the user understand how they work and what data is used. Meanwhile, openness was not deemed relevant to this particular thesis, due to the inherent usability and transparency focus of this thesis.

2.3 Awareness

The research of awareness is a part of the field of perceptual psychology. It is related to how humans perceive their environment.

In their paper "Toward a definition of Awareness" Merikle et al. [Mer84] describe awareness by checking if an objective stimulus is recognized by a person. Meaning if the person is consciously aware that the stimulus exists.

To be more specific to the topic of this thesis, the stimuli will be defined as the practices of big data companies. This puts awareness into a context of users recognizing certain uses of their data and being able to understand risks or dangers that emerge.

2.4 GDPR

The General Data Protection Regulation (GDPR) is a regulation that was put in place by the European parliament and council in 2016. It aims to protect users' personal data. The methods that are used include forcing companies to inform their users about how their data is used, which data can be linked to a specific user and to provide a way of deleting personal data.

The GDPR has had global repercussions due to the fact that if a company provides their services to citizens of the European Union, they must abide by the GDPR. Some companies have chosen not to provide further services in EU-territory.

2.5 Cookies

Kristol et al. [Kri01] offer a detailed description of cookies. The World Wide Web uses the Hypertext Transfer Protocol (HTTP) to exchange data between computers that function as servers and computers that function as clients. HTTP uses stateless servers. This means that the server has no persistent information about the client that is accessing the service. Some web services need persistent information (e.g., the content of a digital shopping basket). To combat the problem of non-persistent information, cookies were introduced in the nineteen-nineties.

A server sends information about a client's state to the client, so it can be saved. The client can then send the information back to the server when accessing the same service at a later time. Some non-transparency can occur if the client has to send the cookie information to more than one server at a time. This hides the propagation of, in some cases, personal information from the user and makes it difficult to comprehend which server holds what data related to a user.

An often used application of cookies is to store login information (e.g., on Facebook) to have user automatically log into their account at the next session. Cookies can also be used to identify users regardless of their IP-address by linking together pieces of personal information.

CHAPTER 3

Personalized Advertising Implementations

Personalized advertisements are the most obvious way how the use of nontransparent algorithms can surprise and sometimes even frighten users. Many people that use the Internet frequently have at least once experienced an advertisement on a website, that was eerily fitting to a product or website that the user has visited before.

3.1 Behavioral Targeting

Behavorial Targeting (BT) utilizes previous user behaviour in statistical machine learning applications.

Chen et al. [CPC09] describe the method of BT as follows. BT uses three deciding factors to determine if an advertisement could be interesting to a user. The first factor is previously visited sites. This shows a general interest in a topic by the user. For example: If a user visits a site that sells consumer electronics, BT notes that they are interested in the general field of consumer electronics. The second factor is terms that the user queried. If a user searches for cameras on the consumer electronics store website, BT knows that the user is interested in specifically cameras in the general field of consumer electronics. The third factor is advertisements that were clicked by the user. If a user has previously clicked on an advertisement for a camera company, BT would know that the user is interested in that particular company. The three factors are often stored in cookies and used to calculate a metric that determines if a user could be interested in an advertisement that is served to them. This creates a system that lets advertisers automatically choose users that have a higher chance of clicking their advertisement and therefore being more likely to buy their product.

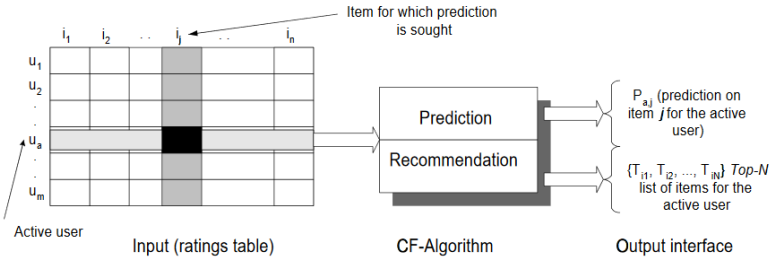

Figure 3.1: Collaborative Filtering Processing Steps as Explained by Sarwar et al. [SKKR01].

BT often looks at time periods that are rather short. If it looked at user behaviour over the course of a week, the relevant information would be sparse. That is why small windows of several minutes are often used by BT-systems.

A major challenge of BT is that the data used can get exceptionally large. Therefore, BT-systems put an emphasis on selecting only relevant data and linking similar data to reduce the amount of data that is being processed.

3.2 Collaborative Filtering

Collaborative Filtering techniques are based on user made ratings for items, that are used to calculate a score for how another user would like the specific item. This method was described in detail by Sarwar et al. [SKKR01].

The created scores for items are called prediction scores. The items with the highest prediction scores are labeled as recommendations that are then presented to the user, as shown by Figure 3.1. Recommended items can not already have been purchased by the user before, as the algorithm aims to get users to buy items. Collaborative Filtering approaches are split into two main categories.
The first is the memory-based, or sometimes called user-based, approach. This technique utilizes the database containing users and their respective ratings of items. Users that have similar ratings, also called neighbors, are found to calculate a recommendation score for a new item.
The second approach is using model-based, or sometimes called item-based, algorithms. These techniques focus on the ratings a user gave to items in the past. They deploy probabilistic algorithms to find the rating for a not yet rated and not purchased item.

3.3 Automated Analysis of Interests and Activities

Using the automated analysis of interests and activities to personalize search results was proposed by researchers at the MIT and at Microsoft Research [TDH05]. Although it serves as a way to rearrange website hits when using a query, it can also be used to

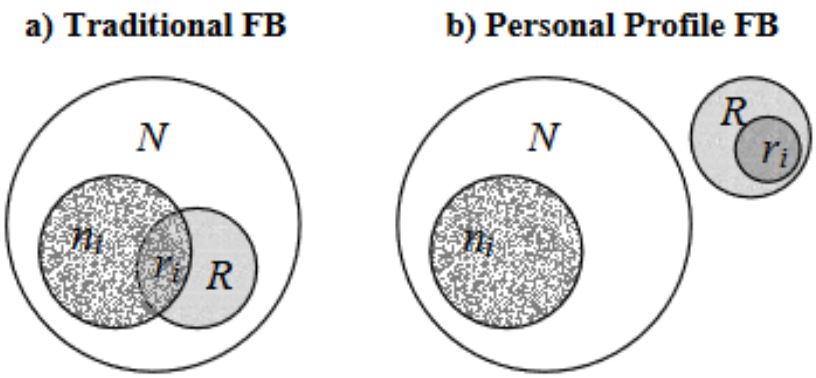

Figure 3.2: Seeing the user space as a separate entity from the corpus according to Teevan et al. [TDH05]. With N being the corpus, ni being items of the corpus and R being the collection of relevance information ri. The image on the left shows an implementation integrating the relevance information in the corpus, while the depiction on the right shows the separation highlighted in this thesis.

choose the right advertisement for a website to be shown to the user. Similar to how the Google search engine shows advertisements for websites when a user has entered a query.

The approach starts by categorizing the corpus, the user space and documents. The corpus contains documents that can be found on the Internet. This could include all documents available or just documents that are relevant to a given query. The user space is where relevant information is collected. This is also where most of the processing is done. Documents are websites that contain words and sentences that are relevant to queries or to rearranging websites according to interests and activities of the user.

Relevant information about the user's interests and activities is gathered by looking at previously accessed websites or read emails. The top websites related to the user query are pulled from the Internet to be processed locally. This results in the separation of the space of relevant information (R) from the domain (N) as shown by Figure 3.2. The aforementioned terms of interest are used to re-rank the websites containing the same terms or related ones.

This implementation of ranking websites focuses on privacy. The calculation of ranks for the websites is done on the user space (i.e. offline). Personal data containing information about previously viewed websites or other information does not have to be processed by a server, meaning that the risk of a breach of private data is minimal compared to other implementations. This can serve as a way to prevent private data to be misused without the user's knowledge.
The privacy aspect comes at a price however as only a limited number of websites can be looked at in this approach. The key is to find a number of websites that are all somewhat relevant to the query but also include terms corresponding to the interests and activities

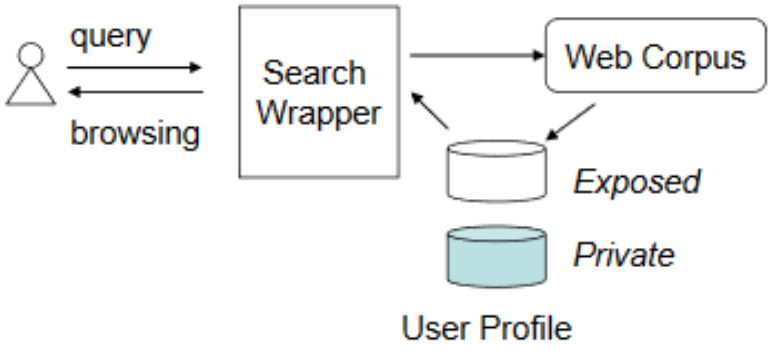

Figure 3.3: Differentiating between public and private parts of the profile as detailed by Xu et al. [XWZC07].

of the user.

3.4 Privacy Enhanced Web Search

To enhance the privacy of web users and transparency of data shared, this approach was suggested by researchers at Simon Fraser University and Microsoft Research [XWZC07]. It aims to let users select what kind of information is shared with web applications. Meaning that a private profile and a public subset of interests are generated as shown by Figure 3.3.

The first part of the algorithm is to generate a hierarchical representation of the user's interests. This assumes that some terms are related to others in a parent-child relationship, which means that the hierarchy is uni-directional. General terms of an interest are the parents of more specific terms (e.g., dramas as a child and movies as the parent). The used data is collected, similar to the previously discussed approach, from documents or websites that were read by the user.

Some terms can be merged into one, because of their similarities. The user chooses a threshold δ that is used for the similarity calculation using the Jaccard function. This threshold is also used to determine the child terms of parent interests that are being used.

Documents that include a term are attributed to the term. This mapping of documents to terms is called support. If a document includes more than one term of interest, the support value is split between the terms (e.g., if a document includes two terms the value would be 0.5 for each term).

The assumption that it is in the user's interest to hide more specific information is made. This means that child terms pose as more sensitive information than parent terms.

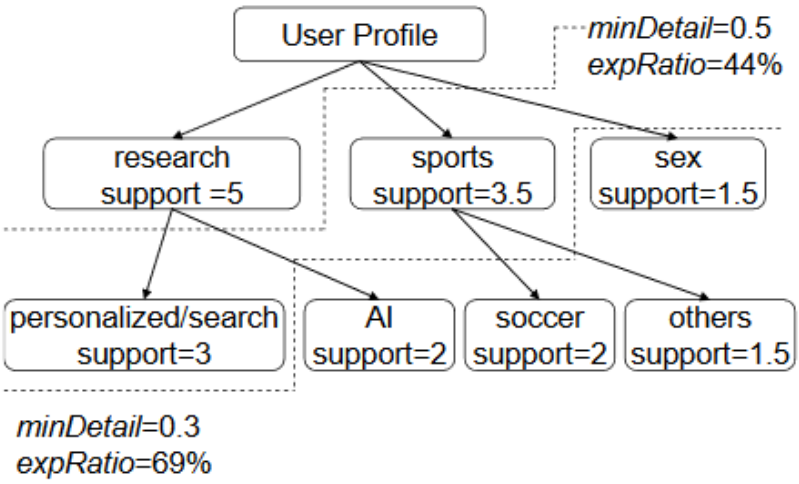

Figure 3.4: A user profile of interests taken from Xu et al. [XWZC07].

Specific terms have generally less support than general terms, hence a parameter called *minDetail* is used to let the user regulate the information that is shared with the server.

The second important parameter is called *expRatio*. It measures the amount of data that is exposed to the server. It is calculated by dividing the exposed terms by the total number of terms. The *expRatio* is dependent on the *minDetail* as it decreases if only general terms are shared with the server and increases if more specific information is exposed. The *expRatio* serves as a way to make the amount of information that is shared more transparent to the user. An example for what information of a given user profile is hidden according to selected *minDetail* and *expRatio* values is shown in Figure 3.4.

In accordance to other approaches, the constructed user profile is sent to the web application. There, the web pages are reranked using the support scores of the terms included in the profile.

3.5 Using a Middleman to Serve Advertisements

This approach uses client based information to make personalizing advertisements more secure, like the approaches discussed before. It was described by Guha et al. in 2009 [GRT+09].

This model includes the regular stakeholders like users, publishers and advertisers, but it also introduces more. The broker is, similar to a search engine, responsible for listing content and advertisements for the user to see. Then there is also the dealer, who functions as a middleman. An overview of how the model could be implemented is provided in Figure 3.5.

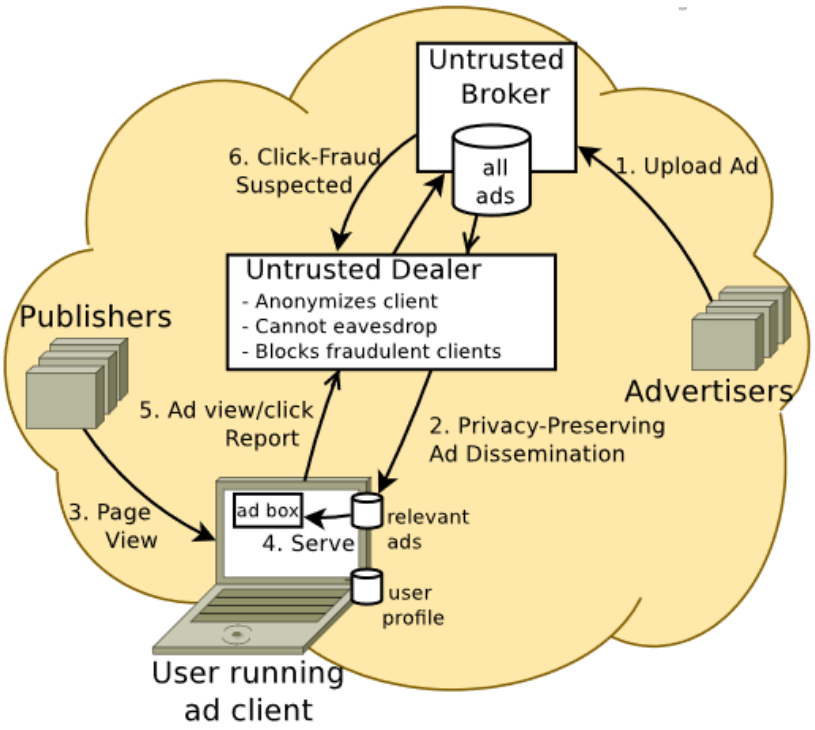

Figure 3.5: Using a dealer between a client side application and the broker from Guha et al. [GRT+09].

The dealer should ideally be a neutral trusted entity, as it can access private information of the users. Possible candidates for this role include non-governmental privacy groups or governmental agencies. The dealer is the only one who knows the identity of the user. It is thereby hidden from brokers. The dealer treats the advertisement clicks of the users as separate entities, preventing the linking of interests and subsequent identification of users. Hiding the identity is an advantage for privacy but can be a disadvantage to the broker, publisher and advertisers.

If the broker does not know the identity of the user, attacks such as designated denial of service attacks cannot be combated. The solution is to let the dealer keep track of identification of the users and their clicks. If an attack is being carried out, the broker sends the click ids that it got from the dealer back to the dealer. The dealer then maps the click ids to the user responsible and an attack can be stopped.

3.6 Selective Data Sharing

The middleman used in the previous approach would have to be trustworthy. There have been several scandals involving governmental or non governmental institutions that

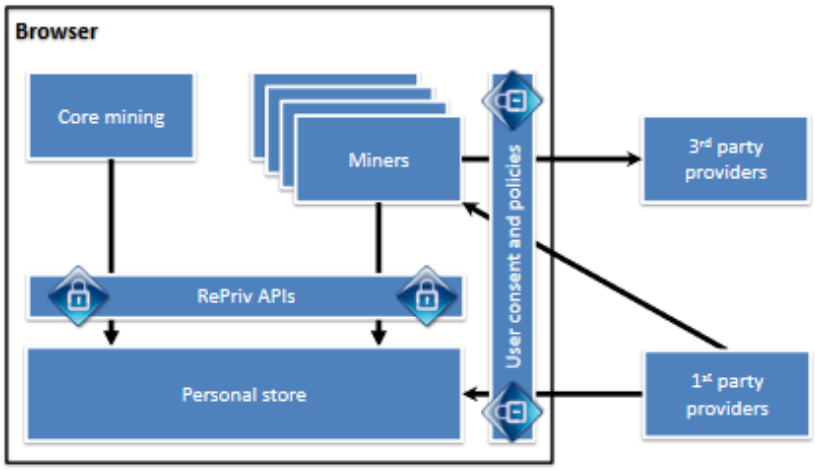

Figure 3.6: Using the Re-priv API to restrict access to personal data according to Fredrikon et al. [FL11].

used personal data in a way that was disadvantageous to the general public, such as reports about dealings at the National Security Agency. By providing the users with a way to customize the amount of data that is shared with web companies, personalized advertisements would become more transparent.

The implementation called Re-priv by Fredrikon et al. [FL11] is based on interest categories and software that asks the user for information. Every software implementation for each websites would ask for different information. For example: A website about movies would request information that belongs to the interest category of movies. Besides the category a number is used to signify the specific subset of the category, making more specific interests possible (e.g., action movies). The user then decides whether to accept the request for a specific level of private information or to deny it. The architecture of re-Priv using the re-Priv extensions that are called *miners* is shown in Figure 3.6.

CHAPTER 4

Related User Research

The "Related User Research" Chapter gives an overview of selected experiments that were conducted in the past. These experiments are related to personalized advertisements in the form of recommendation algorithms on Facebook. This chapter serves to establish a context for Chapter 5.

4.1 Interviews Regarding Facebook Algorithm

In a study conducted by Bucher in 2016 [Buc16] participants were questioned regarding their feelings towards the recommendation algorithm deployed by the social media website Facebook. Participants for the study were selected by observing posts on Twitter. Posts on Twitter were searched for by using combinations of the words Facebook, algorithm and adjectives such as "weird", "creepy" and "great" [Buc16]. Participants that wrote a statement about their perception of the Facebook recommendation algorithm were then asked to take part in interviews, that would further explore their feelings.
One participant talked about how the Facebook algorithm makes assumptions about their demographic group and what they should be interested in. That participant felt offended by the negative reminders they got when the algorithm recommended a product or a website to them. They talked about how they were not in a relationship and did not possess an adequate amount of wealth. Consequently they felt reminded of their own shortcomings when the algorithm recommended websites such as dating sites.
Another participant shared their frustration regarding the recommendation of their own posts to other people. They felt as if they normally would have known how the algorithm would behave and at which times what type of post will be highlighted to other people and therefore felt as if the algorithm was broken or behaving incorrectly when their post did not garner enough attention.
A different participant explained how the algorithm showed another person a picture of that person's deceased daughter. This picture was connoted by the algorithm with

a positive note as part of a yearly review post. The participant felt that the algorithm does not possess human traits that would prevent it from showing that person a picture of their daughter when she had deceased not long before that.
The final participant that was interviewed explained how they went out of their way to behave in certain ways to help their friends to benefit from the inner workings of the Facebook algorithm. This led to the researcher believing that the participants and most Facebook users have certain ideas how the Facebook algorithm works and based on these ideas, exhibit certain behaviors to gain the advantages that the algorithm can give them.
This experiment showed not only how users react to the Facebook algorithm but what the users think about the algorithm and what emotions or affections the algorithm causes as well. Similarly, Chapter 5 explores how participants use terms such as being cautious or showing their indifference toward personalized advertisement algorithms.

4.2 News Feed Algorithm Study

In 2015 Eslami et al. conducted an experiment involving participants' awareness regarding a news feed algorithm such as the one used by Facebook [ERV+15]. This study consisted of assessing the awareness of how posts on a user's news feed are filtered by the algorithm, actively showing the difference of how a news feed is structured using no algorithm to hide seemingly uninteresting posts and subsequent interviews if the participants' behavior regarding their news feed had changed.
The results of the first part of the study (i.e., the questions regarding awareness) showed that as much as 62.5% were not aware that an algorithm sorts and hides certain posts that are deemed to be not interesting to the user. 37.5% of the participants were aware that an algorithm is used to filter the Facebook news feed.
The second part served to show the unaware participants that in fact an algorithm is used to filter their news feed. This helped to eliminate the notion that unaware participants had done something wrong when they did not see certain posts, that the unaware participants expressed before. Furthermore, some participants exhibited affections such as anger and frustration about the algorithm hiding posts with seemingly no advantage to the user.
The follow up conducted several months later showed that participants' behavior regarding their news feeds had changed. They made use of settings given to them by Facebook to control what is shown on their news feeds.
This study showed the importance of awareness and how an experiment regarding awareness can lead to a significant change in behavior when dealing with algorithms.

CHAPTER 5

Awareness Experiment

In January of the year 2020 an experiment was conducted at three different universities in Vienna. This experiment had the goal to assess the awareness of Viennese students regarding their personal data on the Internet. It consisted of a local version and an online version. Participants were asked to click through a website, that was set up on a tablet in a public space without the researcher being present, before answering interview questions in the local version. The interview was replaced by simple text boxes that participants had to fill out in the online version. The questions asked in the local and the online parts were designed to be identical.

5.1 Participants

Participants were limited to being only students of one of three Viennese universities. The universities in question were the University of Vienna, the TU Wien and the University of Natural Resources and Life Sciences, Vienna. A total of 56 people participated in the study. After filtering out unusable answers, the answers of 42 participants were selected to be used for this thesis. Of these participants 10 were studying at the University of Vienna, 20 at the TU Wien and 12 at the University of Natural Resources and Life Sciences as shown by Figure 5.1. the fact that students of the TU Wien dominate this statistic can be explained by their general interest in algorithm studies and therefore their increased willingness to participate in local or online studies regarding algorithms. 26 of the participants stated to be male while 16 identified themselves as female. The majority of the interviews was conducted in German and the questions on the website were asked in German as well. Two participants chose to be interviewed in English. For the purpose of this thesis, answers that were originally given in German, but had to be used in the "Results" Section, were translated to English.

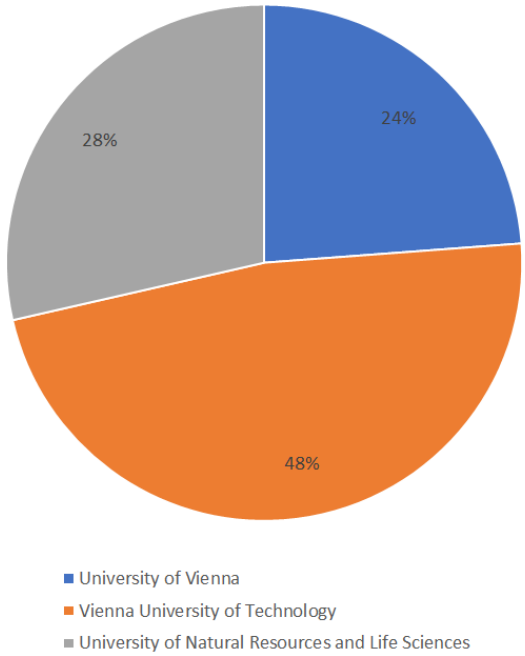

Figure 5.1: Diagram showing the distribution of students enrolled in one of the three Universities.

5.2 Procedure

Two different websites were created. One to be used in an environment that was akin to an experiment and one that was to be used by participants on their own devices at whatever time they please. The frontend (i.e., the part of the application that was visible to the participants) was created using HTML, CSS and the JavaScript-framework Vue, while the backend (i.e., the part handling data) was realised with the python framework flask. The application consisted of three choices the participants had to make by touching images on-screen and thereby answering a question.
The experiment environment was setup using an iPad that accessed the website via a browser. The tablet was set up at a table in the common rooms of the respective universities. The researcher was hiding nearby, as a part of the experiment was for participants to become curious of their own accord. From a distance, participants could see that the start screen of the website was flashing in alternating shades of green and red. It read "Erfahren Sie etwas über sich selbst" which translates to "Find out something about yourself". If the participants touched a button beneath the text, the application would start.
The first choice the participants had to make is whether they prefer products by Apple or Microsoft, as shown by Figure 5.2. This choice would increase a variable called "purchase power" and show the next page if Apple was selected or simply show the next page if Microsoft was selected.

Figure 5.2: Depiction of the first choice users had to make. Selecting Apple would increase the "purchase power" variable.

Figure 5.3: Depiction of the second choice users had to make. Selecting the item containing "der Standard" and "die Presse" resulted in an increase of the "education" variable.

The second page asked the participants about the Austrian newspaper they favor. The selection consisted of der Standard, die Presse, die Kronen Zeitung, Heute and Österreich, as shown by Figure 5.3. Selecting der Standard or die Presse would result in an increase of a variable called "education" and the next page being shown. Touching the images of Kronen Zeitung, Heute or Österreich would only show the next page without increasing the variable.

The final selection participants had to make is where they lived at the time of participating in the study. An image representing the countryside and a different image representing the city were available, as seen in Figure 5.4. Selecting the image associated with the countryside resulted in the "car" variable to be increased, before showing the recommendation page, while selecting the city resulted in no increase but only the change of pages.

The recommendation page showed a mobility related product (e.g., an electric car) depending on the selections made on the previous pages, depicted in Figure 5.5. Participants could then touch a green arrow below the recommendation to answer short questions. The questions were "Does the recommendation fit to you as a person?", "Are

Figure 5.4: The third choice. Selecting the icon for the countryside increased the "car" variable.

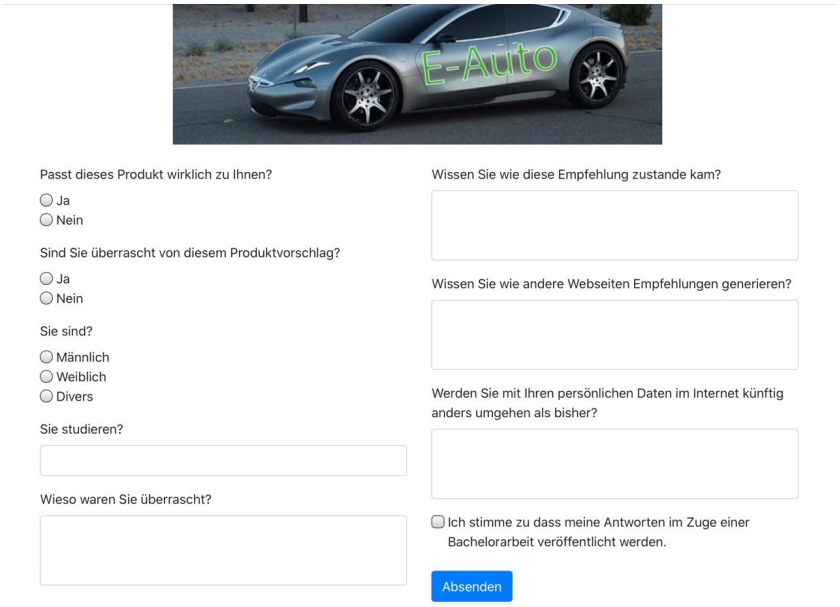

Figure 5.5: The recommendation page with an exemplary recommendation. This depicts the online version of the application including the interview questions.

you surprised by the recommendation?", a question about their gender identity and a question about their field of study.

After completing the experiment, participants were then asked to answer four questions by the researcher, who had at that time made himself known and explained what the experiment was for. These open questions were recorded and consisted of "Why are you surprised/not surprised by this recommendation?" depending on what the participant answered before, "Can you imagine how this recommendation came to be?", "Do you know how other websites generate recommendations?" and "Are you going to handle

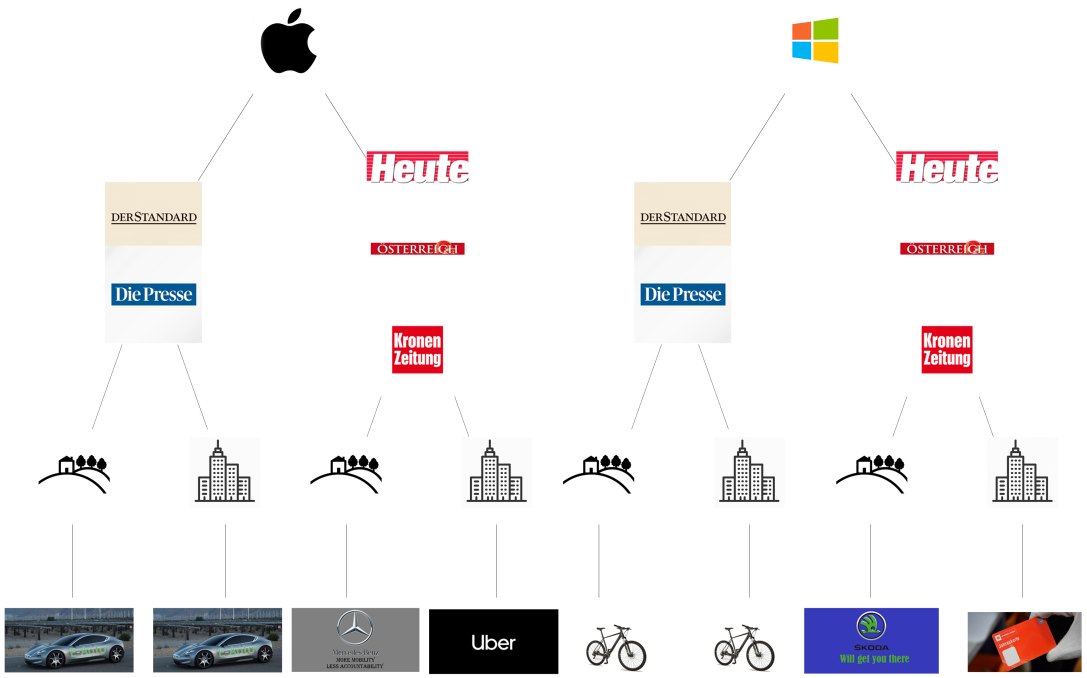

Figure 5.6: Graph detailing the paths a user can take in the application.

private data on the Internet differently than before?".
The website that was created to be used by other participants at home had notable differences from the website used in the experiment. The title page used a neutral white background instead of the flashing eye-catching background that was made to get participants' attention in the local version and the recommendation page had no arrow that had to be touched to show the questions, but included the open questions that were asked by the researcher in the experiment.

5.3 Application in Detail

This section serves as a detailed overview of the assumptions and design decisions that went into creating the web applications that were used in the experiment and by participants at home. The choices a user could make are binary in nature. Even though there is one stage of the application where more than two items are presented to be chosen from, the logic of the application only uses a system where a variable per stage is either incremented by one or not incremented depending on the choice the user made. The paths a user can take while using the application can be described as a function that depends on three variables. The first variable is called purchasing power. Its value

f(0,0,0)	Viennese public transport ticket
f(0,0,1)	Skoda
f(0,1,0)	bicycle
f(0,1,1)	bicycle
f(1,0,0)	Uber
f(1,0,1)	Mercedes
f(1,1,0)	electric car
f(1,1,1)	electric car

Table 5.1: Variable values and their corresponding results.

is decided by clicking on the Apple icon (value = 1) or on the Microsoft icon (value = 0). Similarly the newspaper icons change the value of the education variable and the city or countryside distinction sets the cars variable. Figure 5.6 shows all of the possible outcomes depending on what decisions were made by using the application and thereby describes the values of the three variables in relation to the result. Additionally, the results can be described by a function using the three variables. Table 5.1 shows the resulting recommendations of the eight possible value combinations.

5.4 Results

This section shows the results regarding the research questions as explained in Chapter 1. To summarize, the research questions were "Are university students aware of current practices in personalized advertisements?" and "Can students be surprised by a simple website that does not use a sophisticated algorithm or artificial intelligence to recommended products?"

5.4.1 Research Question 1

All participants except for three showed an interest or previous knowledge regarding personalized advertisement. Most answers regarding implementations of big data companies used the term "cookies" or user profile construction techniques such as using previous product searches by the user to generate recommendations. Two participants even went so far as to assume that Internet services on their phone recorded what they talked about with their friends to use that information for product recommendations.
The three outliers stated that they "know that personalized advertisements exist, but they do not know how recommendations are generated", "know that data companies generate personalized advertisements but do not know how they are done" and "have never researched the topic at hand and have no interest to do so".
Related to awareness regarding personalized advertisements practices, the last survey question "Will you change your behavior regarding personal data on the Internet?" asked for awareness of personal data misuse that could possibly be created by the website created for this study. Most of the participants responded that they will not change

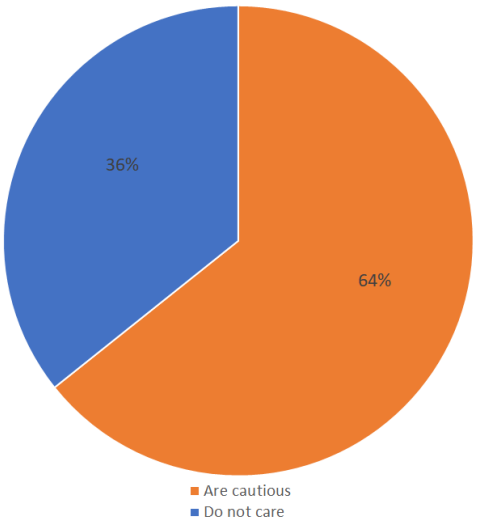

Figure 5.7: Divide in opinion regarding personal data use on the Internet.

their behavior. The ways they handle their personal data on the Internet varied however. 64% of the participants claimed that they were careful in regards to their personal data on the Internet. Some of the participants felt as if they should be more careful but they do not have the time necessary to prevent misuse or targeted manipulation of their interests. The remaining 36% of participants did not care about use of their personal data on the Internet or felt personalized advertisements had only advantages. This divide is shown in Figure 5.7. Four participants claimed that they liked personalized advertisements, because they simply "work better" than other advertisements that are not targeted and tailored to user interests. One notable utterance by a participant was that they felt "flattered" when a big data company wanted to use their data. A notably missing factor regarding personalized advertisements and content recommendation was the possibility of opinions being influenced by content that is shown to users to create a certain opinion on a topic, as mentioned by Mittelstadt et al. [Mit16]. Only one participant (= 2.4%) mentioned how they tried to not be influenced by recommendations in forming an opinion, but they still believed that they would consciously be influenced by personalized advertisements. One participant even disregarded the risk of opinion shaping through personalized recommendation systems by stating that "as long as it is only about the money" they do not care how they are influenced. Furthermore, one participant highlighted the usability aspect. They feared that too much data security or openness with data would lead to web services that are less usable.

5.4.2 Research Question 2

On the subject of participants being surprised, 42.8% of the participants stated that they were surprised by the study. The reasons for this feeling of surprise varied however. Due to the participants not having any information what the subject matter of the study

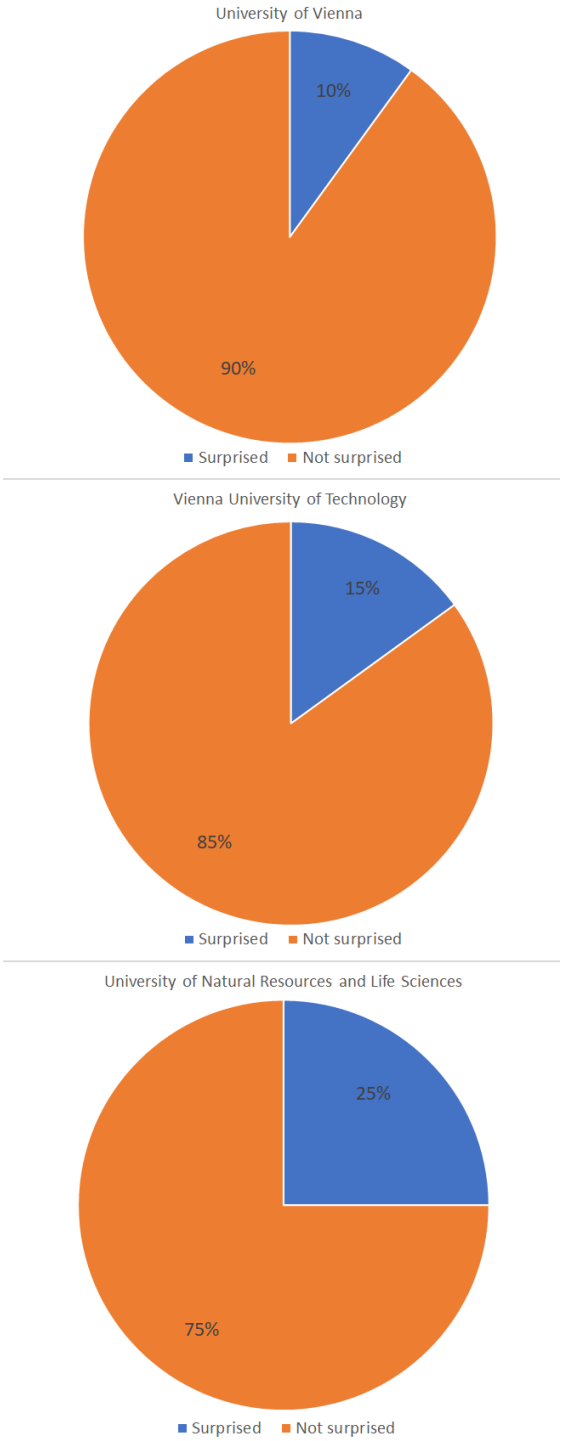

Figure 5.8: Percentage of surprised participants matched to universities.

was and the researcher not being present for the first part of the study, 14.3% of the participants were surprised that they got any recommendation of a product at all, not because the recommendation was generally fitting or unfitting. 7.1% of the participants were surprised that such a simple website was used to give a recommendation that they did not feel as being fitting. 21.4% were genuinely surprised by a recommendation that they felt fitting. 54.8% of the participants stated that they were not surprised by their recommendation, because the website was simple enough for them to assume an accurate way of how the recommendation was created. Of the 21.4% of the participants that could be surprised by the website used in the study, 14.3% at the University of Vienna, 42.9% studied at the TU Wien and 42.9% at the University of Natural Resources and Life Sciences. Compared to the total number of participants enrolled at each university, the percentage of genuinely surprised students is 10% for the University of Vienna, 15% for the TU Wien and 25% for the University of Natural Resources and Life Sciences.

5.5 Discussion

Comparing to the study presented by Eslami et al. [ERV+15] as shown in Section 4.2, significantly more participants were aware of personalized advertisement practices than about the Facebook news feed algorithm in their study. This would suggest that students are more aware of algorithms being at place in web services. Topics that could be interesting for further studies regard creating awareness among the insignificant amount of students that do not care about their personal data about potential risks on the Internet, testing different transparent algorithms on user acceptance or further educating participants about different ways how their personal data is used. As this study has shown, a more complex system would be needed to possibly surprise students with accurate recommendations. This system would have exceeded the scope of this thesis.

5.6 Conclusion

The study at hand has shown that students are in fact aware of current practices, but only limited to speculative descriptions of online data profile techniques. This was expected as seen in Chapter 1. What was not expected was the difference in opinion regarding personal data usage. This study has shown that a significant number of students do not care about their personal data related to personalized advertisements. Furthermore, most students can not be surprised by a simple advertisement recommendation website, yet they can be surprised by unorthodox study methods that include setting up a tablet at a public place and a researcher hiding nearby only showing themselves to ask the interviewee questions.

CHAPTER 6

Analysis

This chapter serves as the conclusion to this thesis. First, the theoretical descriptions of Chapter 3 and the experiment detailed in Chapter 5 are linked. Then, the possibilities for further research are outlined.

6.1 Analysis Regarding the Research Questions

To summarize, the main groups of questions of this thesis were "How do big data companies use personal data of their users to create personalized advertisements?" and "Are students aware of personalized advertisement algorithms? Can they be surprised by a simple website, that emulates an advertisement algorithm?".
Regarding the first question, Chapter 3 gave an overview over possible implementations that may be in use by big data companies. However, as those companies wish to keep their actually used implementations a business secret, this thesis could not claim what algorithms are currently being used with certainty.
Chapter 5 showed how Viennese students think about personalized algorithms. Almost all of the students participating in the study were aware that personalized advertisement algorithms exist and a majority of them could provide at least a basic explanation of how they might work.
The study in Chapter 5 showed that educating users about algorithms that are used for personalized advertisements is important as students used emotionally loaded terms such as "being cautious" around algorithms. That might be possible by using research such as in Chapter 3, where algorithms are explained in a concise overview. Furthermore, Chapter 3's focus on transparency and usability can also be linked to the information obtained from the study in Chapter 5. If students are emotional or even afraid enough so that they have to be cautious when dealing with algorithms, then transparency and usability could help weaken these negative emotions.

6.2 Future Work

Due to the limited scope of this thesis and the large field of personalized advertisements, many topics are still left for future research. First of all, if big data companies were ever to release their algorithms to the public, studies regarding their transparency, usability and feelings of users towards the algorithm could be conducted. At the time of writing this paper, this possibility seems slim. Even if the algorithms are not made public, a study could be designed around a more sophisticated implementation of an emulated algorithm. This could then be used to educate the percentage of people that are not aware of the dangers that personalized advertisements might bring. Additionally those very dangers could be researched as they were not a part of this thesis at all. Another field of interest for a study could be to research why some students do not care about their personal data being used. The last possibility for future work that will be highlighted here is what exactly the 64% of students know about personalized algorithms and how they think they are implemented. The study in this thesis only asked if they were aware of the algorithms and if they were cautious or indifferent to them, but not how exactly they think that the algorithms work.

List of Figures

List of Tables

Bibliography

[Buc16] Taina Bucher. The algorithmic imaginary: Exploring the ordinary affects of facebook algorithms. *Information, Communication Society*, 20:1–15, 02 2016.

[CPC09] Ye Chen, Dmitry Pavlov, and John F. Canny. Large-scale behavioral targeting. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD'09, pages 209–218, New York, NY, USA, 2009. ACM.

[ERV+15] Motahhare Eslami, Aimee Rickman, Kristen Vaccaro, Amirhossein Aleyasen, Andy Vuong, Karrie Karahalios, Kevin Hamilton, and Christian Sandvig. "I always assumed that I wasn't really that close to [her]": Reasoning about invisible algorithms in news feeds. 04 2015.

[FL11] Matthew Fredrikson and Ben Livshits. Repriv: Re-envisioning in-browser privacy. IEEE Computer Society, May 2011.

[Gil13] Tarleton Gillespie. *The Relevance of Algorithms*. 01 2013.

[GRT+09] Saikat Guha, Alexey Reznichenko, Kevin Tang, Hamed Haddadi, and Paul Francis. Serving ads from localhost for performance, privacy, and profit. 01 2009.

[Gur03] Yuri Gurevich. Algorithms: A quest for absolute definitions. October 2003.

[Hea06] David Heald. Varieties of transparency. In *Transparency: The Key to Better Governance?* Oxford University Press for The British Academy, 2006.

[Kri01] David M. Kristol. Http cookies: Standards, privacy, and politics. *ACM Trans. Internet Technol.*, 1(2):151–198, November 2001.

[Mer84] Philip M. Merikle. Toward a definition of awareness. *Bulletin of the Psychonomic Society*, 22(5):449–450, Nov 1984.

[Mit16] Brent Mittelstadt. Automation, algorithms, and politics: Auditing for transparency in content personalization systems. *International Journal of Communication*, 10, 2016.

[SKKR01] Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl. Item-based collaborative filtering recommendation algorithms. In *Proceedings of the 10th International Conference on World Wide Web*, WWW '01, pages 285–295, New York, NY, USA, 2001. ACM.

[TDH05] Jaime Teevan, Susan T. Dumais, and Eric Horvitz. Personalizing search via automated analysis of interests and activities. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '05, pages 449–456, New York, NY, USA, 2005. ACM.

[TWK17] Bryan Routledge Tae Wan Kim. Algorithmic transparency, a right to explanation and trust. 2017.

[XWZC07] 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 '07, pages 591–600, New York, NY, USA, 2007. ACM.