TALLINN UNIVERSITY OF TECHNOLOGY
School of Information Technologies

Anıl Yarkın Yücel 184082IVSB

Optimizing User Identification Algorithms Using Anonymous Data Tracking

Bachelor's Thesis

Supervisor: Kaido Kikkas
PhD
Co-supervisor: Domjan Barić
MSc

Tallinn 2021

TALLINNA TEHNIKAÜLIKOOL
Infotehnoloogia teaduskond

Anıl Yarkın Yücel 184082IVSB

Kasutajatuvastuse algoritmide optimeerimine anonüümse andmejälgimise abil

bakalaureusetöö

Juhendaja: Kaido Kikkas
PhD
Kaasjuhendaja: Domjan Barić
MSc

Tallinn 2021

Author’s declaration of originality

I hereby certify that I am the sole author of this thesis. All the used materials, references to the literature and the work of others have been referred to. This thesis has not been presented for examination anywhere else.

Author: Anıl Yarkın Yücel

29.04.2021

Abstract

The goal of this thesis is developing a user identification algorithm, optimized for increased user privacy by following anonymous data tracking practices, as well as suggesting policies for achieving better anonymity for end-user entities, while creating a basic representation as a pseudo-code implementation.

Furthermore, this thesis will be ready for the future readers as a useful research study that supports better production-ready implementations of these algorithms in terms of privacy, as the related technologies advance.

This thesis describes how the most commonly used implementations of user identification algorithms in the information technology industry work, and briefly mentions possible direct and indirect effects are of these algorithms on those technology companies and their end-users.

This thesis is written in English and is 46 pages long, including 6 chapters, 5 figures and 3 tables.

Annotatsioon

Kasutajatuvastuse Algoritmide Optimeerimine Anonüümse Andmejälgimise Abil

Käesoleva bakalaureusetöö eesmärgiks on arendada välja suuremale privaatsusele suunatud kasutajatuvastuse algoritm, kasutades anonüümset andmejälgimist ning pakkudes välja eeskirja lõppkasutaja anonüümsuse suurendamiseks (luuakse pseudokoodi kujul baasprototüüp). Bakalaureusetöö võiks olla aluseks tulevastele uuringutele ja praktilisse kasutusse suunatud edasiarendustele vastavalt tehnolooogia arengule. Töös kirjeldatakse ka, kuidas kasutajatuvastuse enimlevinud algoritme IT-valdkonnas kasutatakse ning tutvustatakse lühidalt selle otseseid ja kaudseid mõjusid ettevõtetele ja lõppkasutajatele.

Lõputöö on kirjutatud Inglise keeles ning sisaldab teksti 46 leheküljel, 6 peatükki, 5 joonist, 3 tabelit.

List of abbreviations and terms

Algorithm	A process or set of rules to be followed in calculations or other problem-solving operations, especially by a computer
API	Application Program Interface
Browser Fingerprint	A device fingerprint (or browser fingerprint) is collected data of a remote computing device with its software and hardware specifications to achieve identification
Cookie	A piece of information which is stored on a computer device and consists of information related to what a web browser is required to remember
Data Cluster	A sub-group of data which shares similar characteristics and is significantly different to other clusters in a database, usually defined by the statistical technique of cluster analysis.
Dataset	A collection of related sets of information that is composed of separate elements but can be manipulated as a unit by a computer
De-anonymization	The process of matching anonymous information (or de-identified information) with data open for public access, or auxiliary data, for the purpose of identifying the original owner of the data
EDPB	European Data Protection Board
EU	European Union
GDPR	General Data Protection Regulation
GPS	Global Positioning System
HTML	Hypertext Markup Language
IT	Information Technology
Pseudo-code	In computer science, pseudocode is a plain language description of the steps in an algorithm or another system
UI	User Interface
UX	User Experience
Web	World Wide Web, which is also known as a Web, is a collection of websites or web pages stored in web servers and connected to local computers through the internet

Table of Contents

List of Figures

List of Tables

1 Introduction

Every individual who uses any generic IT environment (software, system, website...) which provides personalized service, needs to create their account to be able to access and use this personalized environment and its service(s). This process can be named as “user authentication”, where the users are expected to provide some of their information to the platform for their identification.

Apart from this process, some of the internet platforms also have another process where they gather user data, process and use it to provide personalized service and experience to those users. Meanwhile, this process requires those internet platforms to selectively target the intended users to generate as much revenue as possible over several forms of revenue methods. The algorithms, that are used to compare and identify possible user targets and differentiate from each other, are called “user identification algorithms” or “user matching algorithms” [1].

The main functions of those algorithms are differentiating the virtual entities of the end-users from each other and trying to understand which real and physical entity as an end user they can be. This achieves at least two things for the internet platforms and the efficiency of their revenue streams:

- Understanding the connectivity between multiple user accounts to provide personalized service to both accounts, depending on the gathered data from both accounts [2], [3]
- Accessing the identified end-users’ calculated and possible real-life social environment to provide this personalized service or similar services to them, as well [1, p 47115], [2, p 1], [3, p 1]

The traditional and simpler implementations of user identification algorithms do not have the desired precision due to their absence of user-generated data usage [1]. Therefore, more complex user identification algorithms were needed for a better precision to identify users among and across such platforms by, at least but not limited to, forging user-generated data.

There are already more methods of achieving better precision for the results of these user identification algorithms [1], and as the related technologies advance, there could be even more ways of implementations discovered for the same purpose. For instance, one of the previously mentioned methods require the use of elements such as but not limited to usernames, real names, birthday registrations and location data [2, p 1260,1266]. Thus, whilst the development of such existing and prospective user identification algorithms, it is necessary to pursue the end-user privacy and not overlook any possible vulnerabilities of end-user anonymity.

This study contains numerous special terms and abbreviations that are listed and defined in the dictionary, at the beginning.

1.1 Description of the Problem

The personalized services which many of the most commonly used social network platforms[1] provide are achieved through implementations of user identification algorithms that are believed and documented to be penetrating the anonymity of end-users [2], [3]. That breach of user anonymity leads the compromised user data to be exposed to public especially on cases where the sensitive data is leaked by Cybersecurity attacks or accessed in an unauthorized or unsupervised way.

For instance, it is possible to come across related advertisements on *X* website, about the information that is used or processed on *Y* website, in view of the fact that even if those platforms may not be subsidiary bodies of the same parental organization or do not show any relations to each other. Therefore, the commercial usage of this type of data sharing method is questionable in terms of privacy ethics.

1 https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/

Furthermore, the tracked data which is used by user identification algorithms usually includes the sensitive information that users register while they are creating those accounts in another online product or service of the company that controls a larger ecosystem of products [3].

This means that the users' data are processed and even shared in between multiple online products or services to provide more personalized experiences among different kinds of connected and independent platforms for achieving expected results in revenue.

Additionally, meanwhile the processing phase of the tracked data, there might be occurrences where a user identification algorithm or a part of it fails and outputs an error message into an unprotected virtual place which can be accessed publicly and possibly containing sensitive end-user data. Not to overlook, the excessive use of sensitive data from behavioral, statistical and interconnection based data tracking methods might lead to false identifications as well.

1.2 Methodology

In order to achieve the most efficient result for the intended outcomes of this thesis, the theoretical research is conducted predominantly in a qualitative manner where "interconnection based user identification", "statistical user identification", "behavioral user identification", some other similar types of these algorithms which discernibly carry different characteristics [1], and their combined usage is carefully analyzed.

Moreover, the practical part of the research is mainly supported by externally (other individuals) conducted experiments[1] and discussions[2], that are publicly available on GitHub and online software forums, respectively. However, the practical contribution of this thesis solely depends on the genuine design and build of the flowchart type optimized user identification algorithm, following a pseudo-code implementation which originates from the same method of development.

Mathematical equations and illustrations deriving from those equations, which were both formed and found on external sources, are used, analyzed and taken into account to

1 https://stackoverflow.com/questions/15966812/user-recognition-without-cookies-or-local-storage
https://github.com/erhant/profile-matching

2 https://datascience.stackexchange.com/questions/11204/user-identification-using-machine-learning

corroborate the efficiency and feasibility of such optimized user identification algorithm, as well as to describe the theoretical limits of this implementation.

Usability and reliability analysis of these types of algorithms might require the benefits of machine learning for results with high accuracy [4, p 98]. As a wider and more in-depth research for collecting large datasets are required to achieve a more concrete, production-ready implementation for this algorithm as a software tool, previously conducted experiments by external sources and their results are observed.

1.3 Contribution

During this thesis, the author suggests eliminating the problem by identifying the key aspects and causes of the end-user privacy vulnerabilities and isolating these findings from how the optimized user identification algorithm should be performed. Fulfilling this suggestion relies on the aims of this study, as listed below.

The aims of this thesis are:

- to analyze several forms of user identification algorithms and how they are interfering with sensitive user data, and then providing the outcomes of these analyses for revealing the most efficient method of optimization
- to develop a genuine and optimized version of user identification algorithm by a flowchart and a pseudo-code implementation deriving from this optimized algorithm
- to constitute the required policies that set specific guidelines for this solution to be effective

In addition to this, the author aims to pioneer the anticipated further research about how to achieve better anonymity while more personalized services or targeted marketing solutions are becoming prevalent in the IT industry related projects or companies during the recent years.

2 Analysis

2.1 Analysis of Interconnection Based User Identification

There have been administrative cases where some accounts were tested against linkage among multiple social networks, sourcing from the information gathered by a structural de-anonymization[1] attack. These types of attacks can be conducted by any individual with positive or negative motivations and the access to necessary tools, but more importantly depending on the size of the dataset(s) to be used. Any dataset containing private data (*e.g.* religious or political beliefs) related to actual end-users and which is eligible for purchase from a social network provider, can be used among other social networks or the branch products of the same social network ecosystem to expose related identities [4, p 3].

A social network service has responsibility[2] to protect the sensitive information of its users when storing and processing their data for providing personally tailored content or advertisements, as there is connectivity between different social network providers in terms of sharing tracked user data. Accordingly, keeping this sensitive data open for public access or even only for commercial access for any purpose violates the fundamental principles of user privacy [5, p 378], looking beyond the legal aspects by these users agreeing the respective Privacy Policies on the services.

The work of *Esfandyari et al.* (2016) describes that some APIs which were officially provided by Google Plus, Facebook and Twitter exposed information of identification properties about end-user data with respect to common and uncommon fields of information in between these service providers.

On top of that, it is possible to compare these common variables by specific methods for using machine learning and extract the theoretical similarity (in numerical values) of

1 De-anonymization (*a.k.a. re-identification*): The process of matching anonymous information (or de-identified information) with data open for public access, or auxiliary data, for the purpose of identifying the original owner of the data [6].

2 https://gdpr-info.eu/art-24-gdpr/

multiple virtual end-user identities among different social network providers by using several methods of calculation [5, p 382].

It is evident that forming such algorithms that detect relational user identities among multiple social network platforms is possible, supported by actual end-user data which can be found and publicly accessed through published APIs. Do those commercial organizations provide this potentiality publicly on purpose? There might not be a clear answer to this question, as it is difficult to prove the motivations behind making this type of sensitive data openly accessible. However, it is obvious that relying on social network platforms for end-user privacy is not logical, since those organizations are pertinent to benefiting from these violations financially [7].

Additionally, it is possible to analyze the design of how those APIs are constructed with their data inputs and outputs, and then arrive to a subjective conclusion about why the access to this sensitive data could be kept open for inter-commercial activities and public interest. These software tools rely heavily on the input data as usernames, gender and age group for comparison [5]. Nonetheless, previous study by *Rossi et al.* (2015) shows that the location data of the users of some social network providers was also used by several data tracking methods to collect and store mobile GPS signals and categorically applying the results of this tracking to re-identify end-users across platforms [8, p 5,10]. Consequently, these software solutions and their implementation methods point to possible violations of end-user privacy; resulting in beneficial use of this violation by spreading sensitive information across commercial organizations.

For instance, Google publicly demonstrates a patent[1] of a user identification algorithm which relies on the browser fingerprint[2] difference data (by hashing the browser fingerprint features for secure storage) and benefits from the statistical data originating from the users who pay a return-visit [9]. This means that the algorithm supplies the information of either a page is being revisited by a user or being visited by a fresh user. Needless to say, this algorithm is limited by how the collected browser fingerprint data is accurate. If the data is somehow manipulated or protected from being collected by the end-user, the outcomes of this algorithm would not be useful for user identification.

1 https://patents.google.com/patent/CN106529233A/en

2 Browser Fingerprint (*a.k.a. Device Fingerprint*): A device fingerprint is collected data of a remote computing device with its software and hardware specifications to achieve identification [10, p 878].

This implementation clearly uses data that is collected from multiple software sources that might be originating from the same hardware source. Possible replications of such algorithm (for experimental, educational, harmful or commercial purposes) could be produced to implement identification of revisiting users across social network platforms, as well as being able to filter what those target users view or visit on their browsing route in between platforms. Meanly, third party integration is the main aspect that separates this type of user identification method from others, through several methods, heavily relying on cookie usage.

Likewise, it is documented that Facebook has been actively tracking their users' activities by injecting several types of cookies[1] into many websites on the internet by partnering with them for shared user information [11]. This process has been mainly done by *Pixel Trackers*[2] and *Like Buttons* [12], collecting user data by non-anonymous data tracking with registered user relational data. A demonstration about these cookies can be seen on Figure 1.

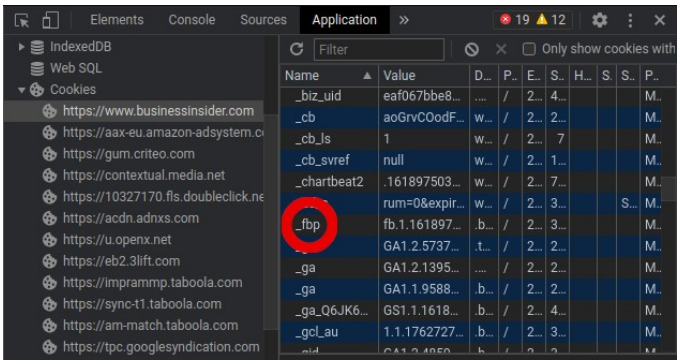

Figure 1. List of cookies on "businessinsider.com" including "_fbp" of Facebook, used for storing and tracking visits across websites [13].

During the analysis of this case, it is discovered that a social network provider has been using its tool(s) across other social network providers, including the ones that belong to the same parent commercial organization (*e.g.* Instagram) and different types of websites such as but not limited to news websites, forums, real estate rental platforms. Further usage of this interconnection based user identification, which is correlatively also classified as statistical user identification, is explained in the next section of this work.

1 Cookie (*a.k.a. HTTP Cookie*): A piece of information which is stored on a computer device and consists of information related to what a web browser is required to remember [14].

2 https://www.facebook.com/business/help/742478679120153?id=1205376682832142

2.2 Analysis of Statistical User Identification

There are many methods for identifying user demographics and culture background with very high accuracy after processing collected information such as natural language identification from natural person names and post content, internet traffic logs and network usage analysis. Concurrently, gender and age statistics can be also calculated by advanced visual analysis of user generated content such as photographs or profile pictures, which provide age estimations with an average error of 2.7 years, while surpassing human prediction with an average error rate of more than 6 years. Collecting and processing such big amounts of data and converting these into large datasets to push into deep neural networks provide results with 83% to 95% accuracy [15, p 171].

Moreover, it is also tested to classify users or user groups into specific socioeconomic status categories for achieving results to identify person importance in terms of potentiality of social impact [15, p 173]. These outcomes are accomplished through advanced usage of conventional and convolutional neural networks for different purposes respectively.

The substantial utilization of this type of statistical user identification algorithm is used by Instagram with the support of big data management [16], as it is publicly described on Instagram's Data Policy[1], user data consisting of page and account views and engagements, types of features used on the application, the usage time and frequency of each action, and more is collected to be logged (stored) [17] with statistical purposes. In addition to this, this policy further describes variety of other methods and attributes of data collection is being done, including caught device signals, network information with details, shared API data with statistical qualities that originates from partner organizations and more.

It is important to enucleate what types of end-user data (sensitive or insensitive), which are statistically collected, are exposed to Instagram's user identification algorithm implementation, as it is claimed to support better practices of UI and UX [17].

Sensitive data categorization differs in between legal and commercial[2] points of view.

1 https://help.instagram.com/519522125107875

2 https://support.google.com/googleplay/android-developer/answer/10144311?hl=en#:~:text=Personal%20and%20sensitive%20user%20data,sensitive%20device%20or%20usage%20data.

According to EU Commission, regarding GDPR, sensitive data is considered as [18]:

- data exposing beliefs considering religion or philosophy
- ethnic origin data
- biometric data for identifying a natural person, and genetic data
- trade-union membership data
- health-related data
- data exposing a natural person's sexual orientation or sex life

If the statistically collected end-user data [17] on Instagram and the descriptions of sensitive data are compared, the common values would include possible violations of user privacy by the usage of this type of user identification algorithm. The results, which only include the collected sensitive data, can be seen on the Table 1.

Table 1: Types of sensitive data collected by Instagram.

Type or Name of the Data	**Classification Under Data Policy**
Health	User-Provided Content (with special protections)
Philosophical beliefs	User-Provided Content (with special protections)
Racial or ethnic origin	User-Provided Content (with special protections)
Religious and political views	User-Provided Content (with special protections)
Trade union membership	User-Provided Content (with special protections)

Although it is claimed to be user-provided content, these types of data are described to be collected through user input on account creation or inserting information on user profile after the creation, as well as creating or sharing content and most importantly while messaging and communicating with other users [17].

It is unclear on the reference document how these types of data are collected from the communications in between users, as sending a message to another user might not mean submitting this information to the social network provider.

It is also addressed that this collected data is regularly used to prevent missing people or help finding them, to provide mass aid under circumstances like natural disasters, and also to detect malicious in real-life and online activity [17].

Furthermore, even though these types of data or insensitive data are changed by the users for the purpose of protecting their privacy, it might still be conceivable to develop a method for identifying these users by comparing those statistical data between the time before and after the data has changed [19, p 1]. Therefore, it is fair to state that preserving and exchanging unmasked sensitive end-user data among publicly available APIs might lead to user privacy violations as well, since there could be links and relations formed for the time-based changed versions of the data.

Likewise, the targeted marketing services among many web services employ similar methods of user identification. Besides only sharing the raw collected data with each other, partnered platforms seem to share beneficial and value added data which carries a potential of usage for personalized advertisement.

Also taking into account that those statistically collected data points might be stored or logged in an encrypted shape [3, p 377]. However, this does not mean that this data remains encrypted or publicly inaccessible meanwhile the sharing of this data through APIs among different platforms or processing of this data through user identification algorithms.

Overall analysis of user identification based on statistical data shows that the sensitive end-user data might be a subject to vulnerabilities, especially considering that commercial organizations share this cluster of data in between themselves, without being open to full-supervision for their algorithms or functions which fulfill this process, due to *e.g.* copyright concerns or Cybersecurity precautions.

2.3 Analysis of Behavioral User Identification

Behavioral user identification algorithms require a bigger variety of meta-data or attributes as input for achieving a better accuracy related to the intended results, compared to other types of algorithms which do not analyze behaviors of the user(s) [1], [20], [21, p 6].

As behavioral identification requires an aspect of comparison, digging deeper into the provided type of datasets with several different types of data is needed.

Factually, it becomes a questionable way for user identification because of the heavy need for many types of data at the same time. Also, these types of data cannot be insensitive data due to the fact that the desired accuracy can only be achieved with specific, user-related content.

The methods of this type of user identification include analyzing typing styles, mapping location visits and processing shared online content [1, p 47114], resulting in an accurate outcome for efficient targeted marketing.

A paid platform for product analytics, "Indicative", describes that the behavioral data is generated with user input (user-generated content), but corresponds and can be referred as to "events". The main elements of user behavior data are classified as "actions" which can be exemplified as "login", "logout", "site visit". Simply, those events should expose the information of "who", "when", "where" for the data to be able to called "behavioral data" [22].

According to the experimental study of *Naini et al.* the behavioral patterns, which can be extracted from histograms of tracked user data, might be a tool for identifying and tracking user entities solely by behavior observation and classification [19, p 1]. This makes it highly possible and viable to produce a pseudo-algorithm for experimental purposes, after accessing user data or by reproducing sample data, and ending up with identified user entities.

It is further explained that the randomness in the user behavior causes corruption or decrease in the expected accuracy for the identification process [19, p 1].

To conduct such experiments with behavioral user data, it is possible to find publicly available datasets, that are collected or created, on data science networking websites such as *Kaggle*[1].

Various dataset results for a search filtered with term "user behavior" are shown on Figure 2.

1 https://www.kaggle.com/search?q=user+behavior+in%3Adatasets

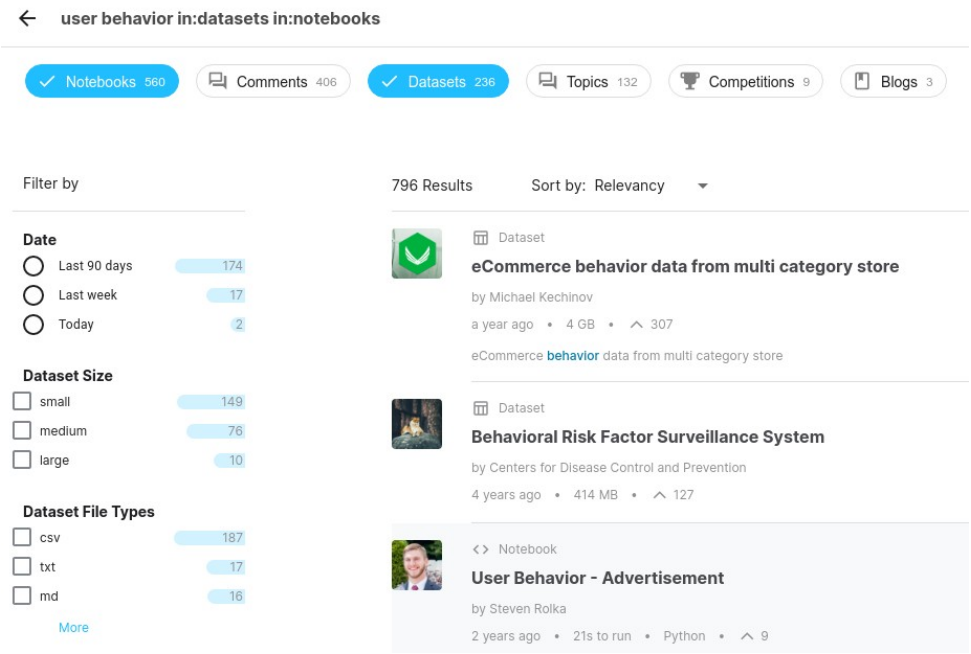

Figure 2: Search results for "user behavior" as datasets and notebooks on Kaggle.

Moreover, Facebook for Business promotes their targeted marketing feature for targeting users over calculated consumer behavior [23]. As previously discussed in Section 2.1 of this paper about interconnection based user identification, Facebook does this with distributed network of user data trackers across the world wide web, including billions of web pages. *Facebook Pixel* tracker specifically targets the users who express potential consumer-like behavior [23], [24].

Statistics show that, the users arrived or passed through pages of Parse.ly platform were referred by Google search with a rate of more than 55%, by Facebook (posts and content links) with a rate of more than 22% [25]. The remaining part, less than 23%, is referred by mostly other types of websites which share smaller chunks of referrer rate.

This big network of data management provides big potential revenue for those types of commercial organizations and their partners. However, meanwhile this is happening, the control of this data is not always in the hands of the user. Behavioral data extraction and tracking is a very controversial topic, being discussed in various forums with ethical aspects in question [24].

Another article published on 2018 shows very detailed results deriving from YouTube statistics that from in a year, “relaxing” content had been increased in watch time with a

rate of 70%, from January 2017 through June 2018 views per month of primitive technology, survival, and bushcraft videos had increased for 248%, and finally 70% of Generation-Z members (estimated to be people 25 and under) had claimed that they used YouTube to watch videos for connect with other people [26].

The behavioral user identification algorithm which was used by YouTube made this statistical data possible with this certainty of accuracy. The power and the potential of behavioral user identification, and what could be accomplished with these implementations were shown with numbers.

2.4 Further Types of User Identification

As discovered through the analyses of the focused three types of user identification algorithms, each type shows similar characteristics while serving different purposes and carrying small differences in terms of their structure and design. Combinations of these methods are also done for increasing the result accuracy even further, as seen on the Facebook and YouTube examples which used behavioral and statistical user identification side by side, along with implementing interconnection based user identification solutions in addition to the other two [23], [24], [26].

Consequently, on theoretical and practical context, those independent types of user identification algorithms could be implemented alongside each other or unified with compatibility into a whole (forming a different result), where many types of data are taken as input and passed through several types of data processing methods including neural network classifications, simulations with machine learning and more.

There are more smaller groups of user identification methods to consider [27]. Differentiating from others, these methods are less commonly implemented due to their direct interference with sensitive data, compared to indirect usages of sensitive data. The ones that are important for the focus of this study are listed below.

2.4.1 User Identification with Biometric Data

According to Odinlaw, biometric data points can be considered as any information considering identification of individuals, resulting in different metrics related to human beings [28].

Some examples according to these classifications are listed as:

- Fingerprint data
- Facial data points (for facial recognition)
- Voice data

Biometric data to identify or classify user identity across multiple platforms can be used by anonymization (disconnecting the data from personally identifiable information) of this data for preserving user privacy, while keeping (or exposing) specific data points not to lose accuracy beyond a feasible limit [3, p 378]. According to an EDPB guideline publication in 2020, misuse of facial recognition data points (*e.g.* possible errors in facial recognition software, biased implementations) and their accidental public distribution could lead to unforeseen outcomes in sociological aspects, including intensifying prejudices of society [29, p 5, 6].

Furthermore, usage of voice recognition patterns and algorithms is getting extensively popular for different kinds of features among IT organizations [30](*see: Related articles*), [31]. Identification of users across platforms is also possible through this technology, with support of third party integrations or interconnection based implementations, even though these are expensive technologies to maintain.

2.4.2 User Identification with Transaction Tracking

Transaction data is formed and dealt with events, so that, it should be classified as administrative data [32, p 556]. Thus, transaction data is naturally a separate field of information than being a subcategory of or connected to statistical data.

Also, since the transactional information consists of secret information such as credit card numbers, bank account numbers and sometimes even online banking passwords or passphrases, this information is behaved separately than regular sensitive information.

It is possible to classify or track user preferences of transaction, their frequencies of purchase tendency and other types of distinctive information for accurate user identification [33]. This processed data then can be used for targeted marketing appliances from perspectives of such as an investor, a company or a governmental body.

2.5 Outcomes

All of these methods of user identification are carrying potential for causing vulnerabilities of end-user privacy online, at different levels. Each types' most influential strengths and weaknesses in terms of privacy concerns, serviceableness at high accuracy, and more aspects are the results of this analyses. Classification for sensitive information follows previously described GDPR guideline [18]. Those results are illustrated on the Table 2.

Table 2: Strengths and weaknesses of analyzed types of algorithms.

Type of Algorithm	**Strengths**	**Weaknesses**
Interconnection Based	• Very high accuracy	• Exposed sensitive information
Statistical	• Very high accuracy • Effective usage for social responsibility [17] • Usage of advanced technology (neural networks)	• Exposed sensitive information • Complex implementation
Behavioral	• High accuracy • Usage of advanced technology (machine learning)	• Complex implementation • Possible exposure of sensitive information
Biometric Data Based	• High accuracy (except malfunction) • Usage of advanced technology (neural networks)	• Possible exposure of sensitive information • Expensive hardware
Transaction Tracking Based	• Wide use among e-commerce ecosystem	• Average accuracy • Possible exposure of sensitive information

It is crucial to achieve high accuracy while keeping the complexity of the implementation as low as possible.

Usage of advanced technology is a groundbreaking rule for IT industry improvement, causing new ideas and solutions to emerge.

In order to achieve the most efficient way for both preserving end-user privacy and providing commercial oriented personalization services across multiple platforms, centralized and decentralized perspectives of user identification algorithm development need to be experimented [34, p 6]. Correspondingly, this could be accomplished through centralization of the identification functions by isolating it from the decentralized and distributed end-user meta-data [34, p 28].

3 Solution

3.1 Technical Decisions

Pursuant to the outcomes of analyses performed, several crucial points have emerged to consider while designing such an algorithm for more accurate user identification results while annihilating possible violations of end-user anonymity. These aspects are categorized accordingly to the structure and main elements of this algorithm.

3.1.1 Input Data

The algorithm is designed from a point of view of a social network provider. Thus, the input data is assumed to provide all of the required data points without missing information, as well as corresponding amounts of data (equal amounts of statistical, behavioral and interconnection based data points), whereas for commercial implementations, the algorithm would be required to provide specific steps for reproducing and filling the missing fields with auxiliary data, *a.k.a.* data generalization process [1, p 47116]. Namely, these required data points are:

1. **Statistical user data**
    1.1. Keywords from user-generated content (*e.g.* post context)
2. **Behavioral user data**
    2.1. Keywords from search queries
    2.2. Timestamps of search queries
3. **Interconnection based user data**
    3.1. Keywords from page visits (*i.e.* HTML metadata “keywords”)
    3.2. Timestamps of page visits

Section 2 of this study includes the information for how and why these data points (first level list items) are separately categorized. Additionally, the reasons for excluding further types of algorithms (Section 2.4) are their low accuracy performance and impracticality for commercial implementation.

Furthermore, it is also assumed that the dataset is constructed of these three main data types carrying identifiable labels. For instance, key *X* data should carry value *Y*, as well as label *Statistical*. This way, the clarity of the information and its distribution are efficient for a satisfactory accuracy. These data points are provided as an input in a bigger package of data cluster, referred as "dataset".

3.1.2 Exposition of the Algorithm

The method for exposition of the algorithm is decided to be an illustration as a flowchart, ideally structured in a language that can be understood even by individuals with little to no knowledge of IT.

The flowchart diagram is created on Flowchart Maker & Online Diagram Software[1].

It starts with the data input "Dataset", and ends with the output of "User Identification" (further explained in Section 3.1.5).

Key elements of the flowchart are "Data Sanitization", "Data Classification", and further separation of main data types into smaller groups of input types, as described as second level list items in Section 3.1.1.

3.1.3 Pre-processing of the Data

One of the main goals of this study, is finding a method for accurate user identification using anonymous data tracking, concordantly achieving nearly full anonymity of the end-user while the processing of this data through user identification algorithms such as the solution discussed on this study. Accomplishing this goal is possible thanks to "data sanitization", or *a.k.a.* data anonymization, on the condition of taking the necessary precautions against de-anonymization attacks [19]. Afterwards, the sanitized data is separated into subcategories of main data types, the phase being referred as "data classification".

1 app.diagrams.net

3.1.3.1 Data Sanitization

When sensitive user data [18] is stored online, especially by large commercial organizations who carry out targeted marketing practices, it is *often* labeled as so (*see:* article item *2* under Section 3.4, Policies) [35]. Thanks to these classifications, it is possible to differentiate from what data is sensitive and what data is not.

Even though the categorization for the term "sensitive data" or "sensitive user data" differs for legal aspect and organizational aspect [18], [36], it is safer to pick the legal approach in terms of keeping the accuracy as high as possible and sticking to official terminations. Hence, the data to be sanitized for this implementation consists of what was described as sensitive data on GDPR guidelines, as addressed before on Section 2.2.

In this study and the flowchart diagram, "data sanitization" refers to isolating the sensitive labeled information from the actual data cluster which is about to be processed through the algorithm, as opposed to the more commonly known description [37] of data sanitization which means destroying or getting rid of data irreversibly.

When this type of data is captured with the respective labels among them, following the proposed methodology of this study, isolating these unwanted parts of the data from the rest of it would complete the step of sanitization and leading to achieving better anonymity for the end-user.

The main reason behind not selectively looking for possible sensitive information in the data by *e.g.* searching keywords that expressing religious or political context is that those keywords might just be fictional (*i.e.* user-generated) content, rather than factual (*i.e.* sensitive data which might lead to identification), that need to be preserved for the sake of keeping a satisfactory level of accuracy.

3.1.3.2 Data Classification

After the data is sanitized from sensitive information, it is now required to be categorized into three main groups of data types (as proposed in Section 3.1.1).

The method for classifying data follows the assumption of receiving labeled data with respect to which main category they belong to. In case of an error in classification, the

accuracy of the algorithm would be drastically decreased, causing a big impact of output corruption. Thus, the correct classification deriving from correct labels is a must.

3.1.4 Methods of Calculation

This study proposes to compare the tracked data from statistical, behavioral, and interconnection classifications by calculating the prevalence of each keyword input among the categorized sections of the dataset. Meaning that, the formed output would include different frequency results for each user or user group in virtue of the combined comparison of different data types. By this way, the accuracy is still preserved to be satisfactory for the possible experimental and commercial implementations even when abnegating better possible accuracy in case of no data sanitization, while granting a better end-user anonymity with processing of only the sanitized data.

These different data types include keyword and timestamp data (only keyword data for Statistical classification). The keyword data for each type get constructed into vectors for the purpose of increasing the program eligibility for recasting the scale of operation over the data clusters [38], and mainly for making this algorithm feasible.

Simultaneously, the timestamp information of those data values (keyword occurrences for behavioral and interconnection based classification) are extracted and compared for similarity. Timestamp data of statistical classification is not taken into account due to the fact that user searches and page visits tend to show connections in terms of occurrence [39]. If those occurrences show a certain level of similarity (later explained below), it means that the user data gathered from different classifications show connection and provide the information that they might belong to the identical user.

It is then needed to compare the results of these separate phases of calculation and analyze the proximity of their results in order to match the user data correlation. This is done by following the suggested methods of calculation, arising from the conventional arithmetic solutions (*see:* Section 3.1.4.2).

Several mathematical equations and explanations for frequency calculation [40], [41], [42] and similarity calculation [43] are carefully compared and understood with their practical advantages and disadvantages for implementing on solutions of this study. Having said that, the best possible results for comparison in terms of accurate

implementation would come out of running actual tests on separate machine learning implementations which derive from each calculation method or selected calculation methods. As this study focuses on better user anonymity and an algorithm constructed with anonymous data tracking, this decision for methods of calculation grants optimization for more anonymity while it does not grant optimization in terms of accuracy.

Those three key aspects of the mathematical fundamentals of this algorithm are listed below.

3.1.4.1 Frequency Calculation

This study proposes to use "tf-idf" (term frequency–inverse document frequency) [42] for calculating prevalence of each keyword in given data clusters. This method brings two separate calculation methods together, "term frequency" to calculate the number of times that a keyword occurs in a data cluster, and "inverse document frequency" to calculate how much information is provided by each keyword in relation to the items of the term frequency equation (*i.e.* checking if the keyword(s) appear commonly or rarely among the provided data cluster(s)). Term frequency and inverse document frequency methods are shown on Equations (1) and (2), respectively. Then, their items are described as well. Finally, "tf-idf" method as the combination of two, is demonstrated on Equation (3).

$$tf(k,d) = \frac{f_{k,d}}{\sum_{k' \in d} f_{k',d}} \tag{1}$$

Equation (1) results as the frequency of keyword $\boldsymbol{k}$, where $\boldsymbol{f_{k,d}}$ represents the raw count of a keyword in a data cluster $\boldsymbol{d}$.

$$idf(k,D) = \log\left(\frac{N}{|\{d \in D : k \in d\}|}\right) \tag{2}$$

Equation (2) results as the weight of keyword $\boldsymbol{k}$, using $\boldsymbol{N}$ for the total number of data clusters in the dataset (for this implementation, fixed to 3), and $|\{\ d \in D : k \in d\ \}|$ as the number of data clusters where the keyword $\boldsymbol{k}$ appears. Following how the algorithm is intended to be implemented, those keywords are selected from the dataset for the calculation. Thus, it would be impossible for the denominator to result in *0*.

$$tfidf(k,d,D) = tf(k,d) \cdot idf(k,D) \tag{3}$$

The result of these frequency calculations for each data classification is constructed as a three-dimensional array[1], each of which consisting of matching amounts of elements with other arrays. An example of multi-dimensional array is illustrated (inspired by an external illustration[2]) on Figure 3.

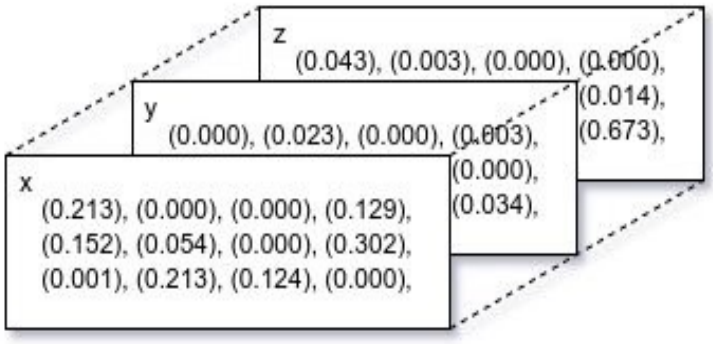

Figure 3: Example illustration of a three-dimensional array.

The results from this calculation (tf-idf) range between *0* and *1*, similar to Figure 3.

3.1.4.2 Similarity Calculation

Jaccard Coefficient [40], [43] is chosen to be the method for similarity calculation due to the fact that data types subject to comparison are the same and they are timestamps, so that they both can be treated as strings. There are examples of calculating similarity of a two-dimensional array filled with timestamps using this technique, implemented on programming languages which have built-in functions with Jaccard Coefficient [44]. The similarity calculation method using Jaccard Coefficient is shown on Equation (4).

$$SimCalc_{Jacard} = \frac{|s_a \cap s_b|}{|s_a \cup s_b|} = \frac{|s_a \cap s_b|}{|s_a| + |s_b| - |s_a \cap s_b|} \tag{4}$$

To clarify, timestamp strings vector, constructed with data from behavioral cluster, is labeled as "$\boldsymbol{s_a}$", while the vector consisting of data from interconnection cluster is labeled as "$\boldsymbol{s_b}$". The result from this calculation ranges between *0* and *1*.

1 https://www.mathworks.com/help/matlab/math/multidimensional-arrays.html
2 https://tex.stackexchange.com/questions/300109/simple-visualization-of-3d-matrix

3.1.4.3 Weighted and Arithmetic Mean

According to an article published by T. Gupta, the proximity analysis is a challenging task due to its lack of standards for implementation [40]. That is why, rather than trying to understand if the results of two different calculation methods are mathematically convergent, it is much simpler to compare means of these results, possibly sacrificing small amount of accuracy but profiting from the complexity of the implementation.

Assuming that the frequency calculations might output a vector including many zero values. To increase accuracy that was sacrificed before for better anonymity, it is possible to take weights of these vectors according to their ratios of the ***count of non-zero values*** to the ***count of all values***. The weight distribution is calculated with method shown on Equation (5).

$$WeightCalc = \frac{count\ of\ nonzero\ values}{count\ of\ all\ values} \tag{5}$$

After calculating weights of each three vectors, their weighted mean is also calculated using Equation (6), where $\boldsymbol{w_i}$ are weights to apply on the $\boldsymbol{X_i}$ values to be averaged.

$$WeightedMean = \frac{\sum_{i=1}^{3} w_i X_i}{\sum_{i=1}^{3} w_i} \tag{6}$$

This weighted mean can be calculated with this conventional formula, without needing to specify it for vectorial comparison [45].

The generic arithmetic mean calculation is followed for the final comparison of the results of Similarity (by Jacard method) and Weighted Mean calculations, as shown on Equation (7).

$$ArithmeticMean = \left(\frac{1}{2}\right) \cdot \left(SimCalc_{Jacard} + WeightedMean\right) \tag{7}$$

3.1.5 Output

After the frequency calculations are done, the results are calculated for their weight in terms of beneficial data, comparing their percentages of non-zero values. Then, those weights are distributed for calculating their weighted mean. Finally, the similarity

calculation output and this weighted mean are taken into account for an arithmetic mean, which finally results in between *0* and *1*.

For the decision of user identification, the threshold value is set to *0.6*. Below this threshold corresponds to *0*, meaning a different user; whereas *1* corresponds to the identical user. The threshold is directly proportional to the accuracy of the algorithm, therefore adjusting it would alter the possible variations for its accuracy.

Examples of the possible outputs and their intended results for user identification after this threshold are shown on the Table 3.

Table 3: Possible results after user identification threshold.

Arithmetic Mean Output	**Result After Threshold**
0.000000	0
0.279834	0
0.599999	0
0.600000	1
0.923489	1
1.000000	1

3.2 Optimized User Identification Algorithm

Elements of the formed algorithm and how it is designed to work are explained in Section 3.1. This algorithm is supported by the analysis of observed types of user identification implementations, theoretical and practical research done previously, and scientific (mathematical) solutions. The accuracy of this algorithm also is directly proportional to the size of the input “Dataset”.

The flowchart diagram of the optimized user identification algorithm is shown on Figure 4.

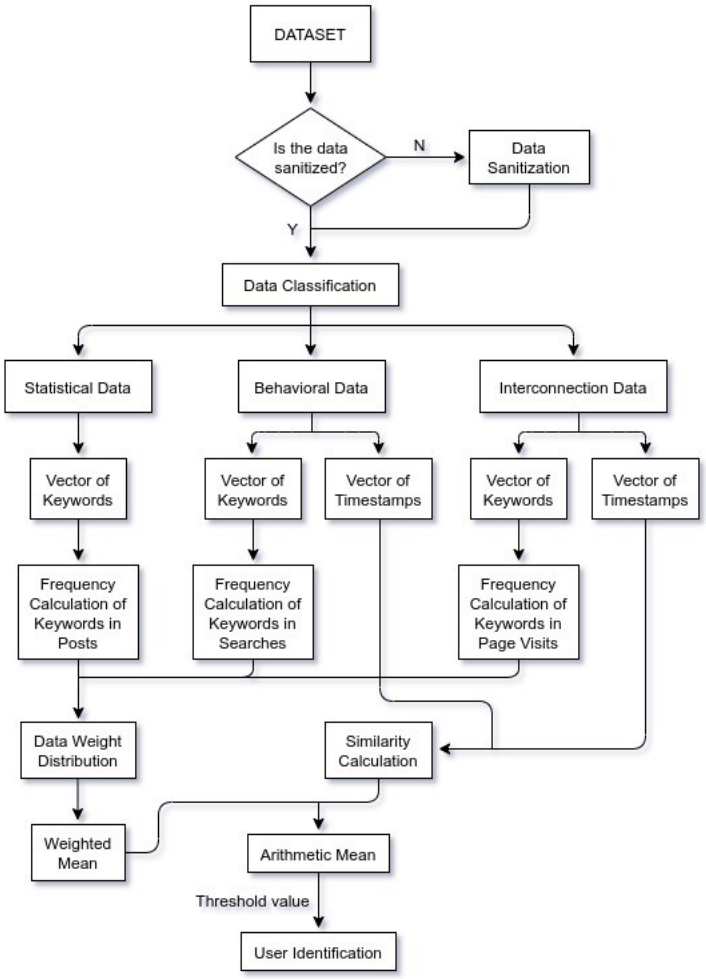

Figure 4: Algorithm flowchart diagram

3.3 Pseudo-code Implementation

The pseudo-code explains how the algorithm can be implemented in any language, focusing on the necessary steps. It is not implemented in a programming language, but it's rather a guideline for applying the algorithm in practice. Vectorization of the data [46], frequency calculation [47], and the weighted mean calculation [48] would be too long and detailed to exhibit in a pseudo-code example, and there are already provided libraries and plugins in many scripting or programming languages to achieve those functionalities. Hence, this implementation does not describe those details.

The pseudo-code is shown on Figure 5, included in the Appendix 2.

The pseudo-code follows the details included in Technical Decisions (*see:* Section 3.1)

3.4 Policies

The Data Policies of the commercial organizations do not always provide clear information about how the tracked (or collected) data is used among their services and if so, shared with third parties.

As the analysis outcomes dictate, certain aspects of data protection policies need to be added and updated in order to increase the quality and the rate of preservation of end-user anonymity among online services, especially commercial organizations which provide targeted marketing practices.

Those policy recommendations are listed below.

1. Sensitive user data classifications by EU (GDPR) [18] need to be updated to include and regulate also these wider categories of personally identifiable information such as:

    1.1. Legal identification number (or code) - [should not be stored]

    1.2. Birth date information - [can be stored but not processed]

    1.3. Debit/credit/virtual card information - [should not be stored]

    1.4. Phone number - [can be stored but not processed]

    1.5. Geolocation (*e.g.* travels, home address) - [can be stored but not processed]

2. In case of storage of sensitive information, service providers should *always* label this kind of information as so, *especially for this algorithm to be effective*.

3. Deterrence of GDPR [49] should be strengthened by following suggested methods:

    3.1. Restriction of targeted marketing services should apply to violators

    3.2. Active regular and irregular inspections, and auditing should be performed

3.5 Ethical Aspects

The solution of this study suggests changes both in regulatory terms and practical implementation that could pioneer practices respecting personal privacy in terms of Ethical Aspects of IT.

For a final analysis of how the suggested solutions conform with Ethical Aspects of IT [50], various points of view were considered. With respect to this, as the optimized user identification algorithm works over the input of anonymous data tracking, it focuses to isolate any information that could breach Personal Privacy aspect, while trying to preserve a satisfactory accuracy of user identification.

Upwards contribution to personal privacy respecting practices, the suggested additional specifications of sensitive user data could be extended.

Moreover, continuing to achieve equilibrium of respect for privacy and further development of user identification technology might be questionable from an ethical perspective due to the fact that this type of technology motivates from the exposure of sensitive user data [51]. Thus, different methods of user identification technologies where they do not interfere with sensitive user data at all would match the ethical expectations in a fitter way. Suppressing or wiping these technologies out of the IT industry will not provide a solution either, since the advancement of unprecedented technologies relies on developing more, rather than restricting.

4 Reliability

The evaluation of reliability of the contribution parts for this study are done by general assumptions of technical and operational behavior. As there is no testable outcome, reliability analysis depends on these aspects.

4.1 Theoretical Behavior

As this study explains design aspects with reasoning (see: Section 3.1 & sub-sections), supported by analyses of references and previously conducted experiments, theoretically its implementation is feasible and efficient for granting user privacy. The optimized user identification algorithm mainly focuses to achieve the best possible anonymity while keeping the abnegation of reliability and accuracy. Although, it is expected to see an uncertain amount of accuracy loss, compared to how it would have been implemented without the isolation of sensitive user data.

It is possible to collect a desired type of dataset following the assumptions and recommendations explained on Section 3.1.1, and building a working implementation with any programming language (ideally related to machine learning or big data, *e.g.* Python, Prolog, R, Java) following the design of the algorithm.

4.2 Usability Analysis

The flowchart diagram of the algorithm is limited for use cases such as:

- Trying to match the possibilities of where several types of data might originate from
- Identifying potential user interest out of several types of data, comparably

Consequently, this algorithm can be used for targeted marketing implementations for experimental or commercial purposes.

5 Limitations and Imminent Studies

Even though the theoretical background of this study follows detailed solution methods, it has not been possible to build a working example due to the limited amount of time as opposed to what would be needed to do it, as well as conducting real-data-driven reliability tests both against privacy preservation rate and accuracy of the results. Collecting such a large dataset that is needed for an actual implementation is expected to take extensive amounts of time and detailed testing practices for a satisfactory outcome.

Furthermore, although this study acknowledges the fact that including geo-location data in the algorithm and the dataset input would only increase the accuracy of the results, the outcomes of this study does not include any possible way for using this type of input to increase accuracy while preventing the violation of user privacy. Yet, there could be methods of achieving this through imminent studies.

Last but not least, it is highly encouraged to do research for and comparably analyze the best possible ways of implementing this algorithm. Since the tests are not conducted due to the absence of a program written with machine learning practices, the pseudo-code only supports the flowchart diagram (algorithm). Because they are made compatible with each other on purpose, modifying one of them would also alter possible implementations and results of the other. Paying attention to these facts would shape a better path for improvement of this study.

6 Conclusion

The most common practices of user identification algorithms were analyzed and compared. During this process, each of them exposed their strengths and weaknesses related to each other. Acquired outcomes of these analyses constructed a path for the formation of the key components of the contributions done by this study.

The goals of this thesis were to develop an optimized user identification algorithm by minimizing or possibly annihilating the use and process of sensitive user data under favor of anonymous data tracking utilization, deriving a pseudo-code implementation from this designed algorithm, and suggesting updates in data protection policies in the light of the solution for the problem explained. Each of those goals were achieved in this study.

For the most efficient accomplishment of these goals, many articles from academic and online sources were observed and considered for research before the development phase of each part of those goals. However, further research is still essential for improving the suggested solutions of this study.

References

[1] Deng, K., Xing, L., Zheng, L., Wu, H., Xie, P., & Gao, F. (2019). A User Identification Algorithm Based on User Behavior Analysis in Social Networks. IEEE Access, 7, 47114-47123. https://doi.org/10.1109/access.2019.2909089

[2] Jain, P., Kumaraguru, P., & Joshi, A. (2013). @i seek 'fb.me': Identifying users across multiple online social networks. Proceedings Of The 22Nd International Conference On World Wide Web - WWW '13 Companion. https://doi.org/10.1145/2487788.2488160

[3] Yi, X., Bertino, E., Rao, F., & Bouguettaya, A. (2016). Practical privacy-preserving user profile matching in social networks. 2016 IEEE 32Nd International Conference On Data Engineering (ICDE), 373-384. https://doi.org/10.1109/icde.2016.7498255

[4] Gulyás, G. (2015). Protecting Privacy Against Structural De-anonymization Attacks in Social Networks. https://doi.org/10.13140/rg.2.1.1471.6963

[5] Esfandyari, A., Zignani, M., Gaito, S., & Rossi, G. (2016). User identification across online social networks in practice: Pitfalls and solutions. Journal Of Information Science, 44(3), 377-391. https://doi.org/10.1177/0165551516673480

[6] "De-Anonymization" (Online Source). Available: https://www.investopedia.com/terms/d/deanonymization.asp (Accessed: 3 April 2021).

[7] "NSA files decoded: Edward Snowden's surveillance revelations explained" (Online Source). Available: https://www.theguardian.com/world/interactive/2013/nov/01/snowden-nsa-files-surveillance-revelations-decoded (Accessed: 3 April 2021).

[8] Rossi, L., Williams, M., Stich, C., & Musolesi, M. (2015). Privacy and the City: User Identification and Location Semantics in Location-Based Social Networks. ICWSM.

[9] "A return visit-paying user identification algorithm based on browser fingerprint differences" (Online Source). Available: https://patents.google.com/patent/CN106529233A/en (Accessed: 5 April 2021).

[10] Laperdrix, P., Rudametkin, W., & Baudry, B. (2016). Beauty and the Beast: Diverting Modern Web Browsers to Build Unique Browser Fingerprints. 2016 IEEE Symposium On Security And Privacy (SP). https://doi.org/10.1109/sp.2016.57

[11] "Facebook knows what you're doing on other sites and in real life." (Online Source). Available: https://www.businessinsider.com/facebook-clear-history-offline-activity-tracker-tool-how-to-use-2020-1 (Accessed: 8 April 2021).

[12] "How Facebook Tracks You, Even When You're Not on Facebook" (Online Source). Available: https://www.consumerreports.org/privacy/how-facebook-tracks-you-even-when-youre-not-on-facebook/ (Accessed: 8 April 2021).

[13] "CookieDatabase: _fbp" (Online Source). Available: https://cookiedatabase.org/cookie/facebook/_fbp/ (Accessed: 8 April 2021).

[14] “What are Cookies?” (Online Source). Available: https://www.kaspersky.com/resource-center/definitions/cookies (Accessed: 8 April 2021).

[15] Brandt, J., Buckingham, K., Buntain, C., Anderson, W., Ray, S., Pool, J., & Ferrari, N. (2020). Identifying social media user demographics and topic diversity with computational social science: a case study of a major international policy forum. Journal Of Computational Social Science, 3(1), 167-188. https://doi.org/10.1007/s42001-019-00061-9

[16] “How Instagram Uses AI and Big Data Technology?” (Online Source). Available: https://www.analyticssteps.com/blogs/how-instagram-uses-ai-and-big-data-technology (Accessed: 9 April 2021).

[17] “Data Policy - Instagram” (Online Source). Available: https://help.instagram.com/519522125107875 (Accessed: 9 April 2021).

[18] “What personal data is considered sensitive?” (Online Source). Available: https://ec.europa.eu/info/law/law-topic/data-protection/reform/rules-business-and-organisations/legal-grounds-processing-data/sensitive-data/what-personal-data-considered-sensitive_en (Accessed: 10 April 2021).

[19] Naini, F., Unnikrishnan, J., Thiran, P., & Vetterli, M. (2016). Where You Are Is Who You Are: User Identification by Matching Statistics. IEEE Transactions On Information Forensics And Security, 11(2), 358-372. https://doi.org/10.1109/tifs.2015.2498131

[20] Halfaker, A., Keyes, O., Kluver, D., Thebault-Spieker, J., Nguyen, T., & Shores, K. et al. (2015). User Session Identification Based on Strong Regularities in Inter-activity Time. Proceedings Of The 24Th International Conference On World Wide Web. https://doi.org/10.1145/2736277.2741117

[21] Fan, X., Chow, K., & Xu, F. (2014). Web User Profiling Based on Browsing Behavior Analysis. Progress In Pattern Recognition, Image Analysis, Computer Vision, And Applications, 57-71. https://doi.org/10.1007/978-3-662-44952-3_5

[22] “What Is Behavioral Data and Behavioral Analytics?” (Online Source). Available: https://www.indicative.com/indicative-blog/what-is-behavioral-data-and-behavioral-analytics/ (Accessed: 10 April 2021).

[23] “Facebook Advertising Targeting Options” (Online Source). Available: https://www.facebook.com/business/ads/ad-targeting (Accessed: 11 April 2021).

[24] “How Facebook and Google Track Your Online Behavior” (Online Source). Available: https://medium.com/@ConnorFinnegan/how-facebook-and-google-track-your-online-behavior-26f161d370ab (Accessed: 11 April 2021).

[25] “Parse.ly's Network Referrer Dashboard” (Online Source). Available: https://www.parse.ly/resources/data-studies/referrer-dashboard (Accessed: 11 April 2021).

[26] “YouTube Analyzed Trillions of Data Points in 2018, Revealing 5 Eye-Opening Behavioral Statistics” (Online Source). Available: https://www.inc.com/tom-popomaronis/youtube-analyzed-trillions-of-data-points-in-2018-revealing-5-eye-opening-behavioral-statistics.html (Accessed: 11 April 2021).

[27] Goga, O. (2014). Matching user accounts across online social networks : methods and applications. (Corrélation des profils d'utilisateurs dans les réseaux sociaux : méthodes et applications).

[28] “What is biometric data and how is it legally protected?” (Online Source). Available: https://odinlaw.com/what-is-biometric-data-and-how-is-it-legally-protected/ (Accessed: 11 April 2021).

[29] “Guidelines 3/2019 on processing of personal data through video devices” (Online Source). Available: https://edpb.europa.eu/sites/edpb/files/files/file1/edpb_guidelines_201903_video_devices_en_0.pdf (Accessed: 12 April 2021).

[30] “Access Google Assistant with your voice” (Online Source). Available: https://support.google.com/assistant/answer/7394306?co=GENIE.Platform%3DAndroid&hl=en (Accessed: 12 April 2021).

[31] “Voice Recognition (Speaker Recognition)” (Online Source). Available: https://searchcustomerexperience.techtarget.com/definition/voice-recognition-speaker-recognition (Accessed: 12 April 2021).

[32] Hand, D. (2018). Statistical challenges of administrative and transaction data. Journal Of The Royal Statistical Society: Series A (Statistics In Society), 181(3), 555-605. https://doi.org/10.1111/rssa.12315

[33] “User-ID: Measuring Real Users Instead Of Devices” (Online Source). Available: https://www.bounteous.com/insights/2015/08/13/user-id-measuring-real-users-instead-devices/ (Accessed: 12 April 2021).

[34] Carmagnola, F., & Cena, F. (2009). User identification for cross-system personalisation. Inf. Sci., 179, 16-32.

[35] “How To Store And Secure Sensitive Data In Web Applications” (Online Source). Available: https://beaglesecurity.com/blog/article/how-to-store-and-secure-sensitive-data-in-web-applications.html (Accessed: 14 April 2021).

[36] “User Data” (Online Source). Available: https://support.google.com/googleplay/android-developer/answer/10144311?hl=en#:~:text=Personal%20and%20sensitive%20user%20data,sensitive%20device%20or%20usage%20data. (Accessed: 15 April 2021).

[37] “Data Sanitization” (Online Source). Available: https://uit.stanford.edu/security/data-sanitization (Accessed: 15 April 2021).

[38] “Vector Data Models” (Online Source). Available: https://saylordotorg.github.io/text_essentials-of-geographic-information-systems/s08-02-vector-data-models.html (Accessed: 17 April 2021).

[39] “6 Types of User Behavior to Track on Your Website & the Tools to Do It” (Online Source). Available: https://www.searchenginejournal.com/user-behavior-tracking-tools/343057/#close (Accessed: 17 April 2021).

[40] “Measures of Proximity in Data Mining & Machine Learning” (Online Source). Available: https://towardsdatascience.com/measures-of-proximity-in-data-mining-machine-learning-e9baaed1aafb#:~:text=Proximity%20measures%20refer%20to%20the,neighbour%20classification%2C%20and%20anomaly%20detection. (Accessed: 15 April 2021).

[41] “Descriptive Statistics and Frequency Distributions” (Online Source). Available: https://opentextbc.ca/introductorybusinessstatistics/chapter/descriptive-statistics-and-frequency-distributions-2/ (Accessed: 14 April 2021).

[42] “Calculate Similarity — the most relevant Metrics in a Nutshell” (Online Source). Available: https://towardsdatascience.com/calculate-similarity-the-most-relevant-metrics-in-a-nutshell-9a43564f533e (Accessed: 15 April 2021).

[43] “TFIDF Statistics” (Online Source). Available: https://jmotif.github.io/sax-vsm_site/morea/algorithm/TFIDF.html (Accessed: 28 April 2021).

[44] “Jaccard Index between timestamps in Python” (Online Source). Available: https://stackoverflow.com/questions/63235482/jaccard-index-between-timestamps-in-python (Accessed: 17 April 2021).

[45] “How to calculate the weighted mean?” (Online Source). Available: https://stats.stackexchange.com/questions/86461/how-to-calculate-the-weighted-mean (Accessed: 18 April 2021).

[46] “Vectorization in Python” (Online Source). Available: https://www.geeksforgeeks.org/vectorization-in-python/ (Accessed: 18 April 2021).

[47] “Processing textual data using TF-IDF in Python” (Online Source). Available: https://www.freecodecamp.org/news/how-to-process-textual-data-using-tf-idf-in-python-cd2bbc0a94a3/ (Accessed: 28 April 2021).

[48] “Weighted average using numpy.average” (Online Source). Available: https://stackoverflow.com/questions/38241174/weighted-average-using-numpy-average (Accessed: 18 April 2021).

[49] “GDPRFines / Penalties” (Online Source). Available: https://gdpr-info.eu/issues/fines-penalties/ (Accessed: 19 April 2021).

[50] “The Ethical and Legal Implications of Information Systems” (Online Source). Available: https://bus206.pressbooks.com/chapter/chapter-12-the-ethical-and-legal-implications-of-information-systems/ (Accessed: 26 April 2021).

[51] “Can marketing personalisation be unethical?” (Online Source). Available: https://annaloverus.medium.com/when-can-marketing-personalisation-be-unethical-dd3bdb1aaa35 (Accessed: 26 April 2021).

Appendix 1 – Non-exclusive licence for reproduction and publication of a graduation thesis[1]

I Anıl Yarkın Yücel

1 Grant Tallinn University of Technology free licence (non-exclusive licence) for my thesis “Optimizing User Identification Algorithms Using Anonymous Data Tracking”, supervised by Kaido Kikkas and Domjan Barić

1.1 to be reproduced for the purposes of preservation and electronic publication of the graduation thesis, incl. to be entered in the digital collection of the library of Tallinn University of Technology until expiry of the term of copyright;

1.2 to be published via the web of Tallinn University of Technology, incl. to be entered in the digital collection of the library of Tallinn University of Technology until expiry of the term of copyright.

2 I am aware that the author also retains the rights specified in clause 1 of the non-exclusive licence.

3 I confirm that granting the non-exclusive licence does not infringe other persons' intellectual property rights, the rights arising from the Personal Data Protection Act or rights arising from other legislation.

29.04.2021

1 The non-exclusive licence is not valid during the validity of access restriction indicated in the student's application for restriction on access to the graduation thesis that has been signed by the school's dean, except in case of the university's right to reproduce the thesis for preservation purposes only. If a graduation thesis is based on the joint creative activity of two or more persons and the co-author(s) has/have not granted, by the set deadline, the student defending his/her graduation thesis consent to reproduce and publish the graduation thesis in compliance with clauses 1.1 and 1.2 of the non-exclusive licence, the non-exclusive license shall not be valid for the period.

Appendix 2 – Pseudo-code Implementation

```
Input dataset

Function sanitize_data(dataset)
      While true
            If (find value “label” equals “sensitive”) true
                  dataset equals dataset minus sensitive
            else
                  break

Function classify_data(dataset)
      For counter i equals 0, i is smaller than 3, increment by 1
            If value in dataset includes label “statistical”
                  s_d equals dataset[value]
            else if value in dataset includes label “behavioral”
                  b_d equals dataset[value]
            else if value in dataset includes label “interconnection”
                  i_d equals dataset[value]
            vectorize_data(dataset[value])
      vectorize_data(classified_data[] equals [s_d, b_d, i_d])

Function vectorize_data(classified_data[])
      // built-in functions
      frequency_calculate(vector[0,1,3])
      similarity_calculate(vector[2], vector[4])

Function frequency_calculate(vector)
      // built-in functions
      data_weight_distribute(frequencies → vector)

Function data_weight_distribute(vector)
      Loop for count_nonzero in vector
      Loop for count_all in vector
      weighted_mean(count_nonzero / count_all, vector)

Function similarity_calculate(vA, vB)
      return (count(difference(vA,vB)) / count(union(vA,vB)))

Function weighted_mean(weights[], vector)
      // built-in functions
      return weighted_mean

Function arithmetic_mean()
      threshold_check((similarity_calculate + weighted_mean) / 2)

Function threshold_check(arithmetic_mean)
      return 1 if arithmetic_mean equals or larger than 6 else 0

If dataset includes “label” equals “sensitive”
      sanitize_data(dataset)
else
      classify_data(dataset)
```

Figure 5: Pseudo-code deriving from the algorithm, simple explanation