a

Thesis Dissertation

PRIVACY ISSUES IN FITNESS-BASED SOCIAL NETWORKS

Zoe Passiadou

UNIVERSITY OF CYPRUS

DEPARTMENT OF COMPUTER SCIENCE

May 2023

UNIVERSITY OF CYPRUS

DEPARTMENT OF COMPUTER SCIENCE

Exploring Privacy Issues in Fitness-Based Social Networks

Zoe Passiadou

Supervisor

Dr. Elias Athanasopoulos

Thesis submitted in partial fulfilment of the requirements for the award of degree of Bachelor's in Computer Science at University of Cyprus

Acknowledgments

First and foremost, I would like to express my gratitude to Dr. Elias Athanasopoulos, my dissertation supervisor. Throughout the process of my thesis, he remained patient and helpful, offering me any assistance I needed. I would also like to thank Pantelina Ioannou, who along with Dr. Athanasopoulos, advised me throughout the months to complete my thesis.

I would like to thank my university and specifically the Department of Computer Science, which provided me with everything I could ever hope for. Upon my acceptance to the university, I was uncertain about what to anticipate. Fortunately, I encountered the most highly skilled professors and teaching staff and the highest quality of study material and courses. For this reason, I feel a great sense of confidence to step into the world beyond academia.

Finally, I would like to thank all my friends and peers at the university. I could not be writing this today without the respect and support we have for one another. Over the past four years we have created memories I will never forget. With you, the university has transformed from just a campus with classrooms to a place that feels like home.

Abstract

This thesis examines the privacy risks in social fitness networks, specifically focusing on Strava a popular fitness network. The analysis of fitness networks shows how much personal information can be gathered from an athlete’s activity. The primary concern is how this data can be used to track an individual's location and reveal sensitive information.

As a solution, a privacy-preserving leader board was proposed which would allow athletes to compete anonymously while still maintaining the benefits of the leader board. The leader board algorithm used hashing and masking methods to hide sensitive information such as usernames and dates of the fitness activity, while still allowing comparison of performances between athletes.

The results of this project show that it is possible to maintain a user’s privacy through privacy-preserving features that also keep the core functionality of the platform. The proposed leader board algorithm provides a viable solution that could be adopted by other fitness platforms to protect user privacy. Overall, this thesis highlights the importance of privacy in social fitness networks and the need for innovative solutions to ensure user data is protected.

Contents

Table of figures

Chapter 1

Introduction

1.1 Motivation

The purpose of my thesis is to address the growing concers regarding privacy issues surrounding fitness-based social networks, while also suggesting ways to safeguard ourselves from these threats. With the increase of social networks in our lives, everyone needs to be alert about the data they are uploading and how it could be misused and abused. Fitness trackers have grown in popularity over the years, in fact, the number of users has doubled since 2018, with 224.27 million people now utilizing this technology [13]. With so many users, it is vital that we exam the implications associated with these networks, and the risks of sharing such data. The potential risks connected to these networks are significant, yet the awareness surrounding them is limited. Moreover, with millions of users, these networks have the potential to collect vast amounts of data regarding one's location and health. These very large amounts of sensitive information being gathered by these platforms, raises questions regarding who has access, how its stored and used, and what measures are taken by these platforms to protect their user's privacy.

My thesis sets to explore the privacy issues surrounding leader boards, specifically focusing on Strava. The goals are to understand not just the privacy issues but the extend of the problem. I hope through my thesis, to shed light on the risks of sharing data regarding your location and propose potential solution to mitigate these risks.

1.2 Thesis Structure

In this chapter I briefly touched on the key motivations and objectives of my thesis. I mentioned the growing popularity of fitness networks and concerns regarding them. Over the next few chapters I will dive into detail regarding the steps I took to complete my thesis. Each chapter will analyse a specific section to help understand everything from start to finish.

In Chapter 2 - Background, I provide the background knowledge of the research topic that is needed to follow the next chapters of the thesis. This chapter aims to give an overview of the key concepts associated with this topic. I delve into the meaning, background and the significance of privacy, I explain the meaning and use of fitness networks and Strava while also explaining the leader board features.

In Chapter 3 - Exploring Privacy Issues, I present a brief overview of the privacy risks associated with sharing location data in fitness networks such as Strava. I hope to give a better understanding of the rationale behind my thesis to help transition into the next chapters smoothly.

In Chapter 4 - Architecture, I aim to help readers understand the underlying architecture of the work I have done. I vaguely discuss the steps I took to gather and analyse data, and the steps taken to implement a possible solution.

In Chapter 5 – Implementation, I include the detailed process of all implementation aspects of my thesis. I dive into detail on the scripts I wrote to obtain and analyse data from Strava. Additionally, I explain all the steps I took towards a potential safeguard against privacy risks in fitness networks.

In Chapter 6 – Evaluation, I present all the findings I have gathered throughout the entirety of my thesis. In this section I hope to provide a comprehensive analysis and assessment of the problem and the proposed solutions.

In Chapter 7 – Related Work, I briefly discuss related work. Specifically, work that focuses on threats of fitness networks and users' understanding of privacy surrounding them.

Lastly, in Chapter 8 – Conclusion, I make a general conclusion regarding my thesis as well as ideas for improvement and future work.

Chapter 2
Background

2.1 Privacy

2.1.1 Definition and Importance

In today's digital age, almost everyone has an online presence. In some way or another, we are connected to the internet, and our personal information is available at the click of a button. It is without a doubt, that the level of accessibility of personal information due to today's digital age has become a significant concern. It is vital that we prioritize our privacy and take the necessary measures to protect our personal information.

There are many interpretations and definitions associated with privacy. However, no matter whom you ask, and what definition you are given, everyone will agree on one thing, its importance. Privacy is everyone's fundamental right, to be in control over their data, and the way it is collected, stored, and processed [18]. Privacy allows people to safeguard themselves and create boundaries from unwanted attention on the internet, keeping them safe from different forms of invasions in their personal lives.

It is crucial to emphasize the importance of protecting and preserving a user's personal information. All platforms strive to have a loyal and consistent user base that trusts to hold and protect their private data. Thus, prioritizing user privacy and safety allows for both user and provider satisfaction, as it allows for the platform to obtain a good reputation while also allowing users to feel at ease when engaging with it.

2.1.2 User Ignorance

Privacy is a hot topic among experts, but when it comes to your average person, this might not always be the case. We often see people that are unaware that the information they are uploading is available to anyone and can be misused and abused, and if they are aware, they often have this "it won't happen to me" mentality. Day by day millions of users are uploading their private and vulnerable information on apps without thinking about the consequences that it might bring.

For this reason, every platform that collects, stores, and uses user data should make it their utmost priority the enforcement of robust measures to ensure good user privacy and safety. A user's ability to control how, when, and where their data is presented should be provided.

2.2 Privacy under the European Union

2.2.1 Past Directives for data protection

Privacy is a growing concern for the European Union since the advance of technology. Specifically, it is a concern for the European Convention on Human Rights (ECHR). In fact, the right to privacy is protected under article 8 which states "Right to respect for private and family life". The European Union has created many data protection regulations and initiatives inspired by article 8.

On the 24th of October 1995, directive 95/46/EC also known as the Data Protection Directive was adopted, on the protection of individuals regarding the processing of personal data and on the free movement of such data [7].

The directive stated the following [7]:

- The object of the Directive

  1. In accordance with this Directive, Member States shall protect the fundamental rights and freedoms of natural persons, and in particular their right to privacy with respect to the processing of personal data.

2. Member States shall neither restrict nor prohibit the free flow of personal data between Member States for reasons connected with the protection afforded under paragraph 1.

- Member States shall provide that personal data must be:

a) processed fairly and lawfully;
b) collected for specified, explicit and legitimate purposes and not further processed in a way incompatible with those purposes. Further processing of data for historical, statistical or scientific purposes shall not be considered as incompatible provided that Member States provide appropriate safeguards;
c) adequate, relevant and not excessive in relation to the purposes for which they are collected and/or further processed;
d) accurate and, where necessary, kept up to date; every reasonable step must be taken to ensure that data which are inaccurate or incomplete, having regard to the purposes for which they were collected or for which they are further processed, are erased or rectified;
e) kept in a form which permits identification of data subjects for no longer than is necessary for the purposes for which the data were collected or for which they are further processed. Member States shall lay down appropriate safeguards for personal data stored for longer periods for historical, statistical, or scientific use.

- Member States shall provide that personal data may be processed only if:

a) the data subject has unambiguously given his consent; or
b) processing is necessary for the performance of a contract to which the data subject is party or in order to take steps at the request of the data subject prior to entering into a contract; or
c) processing is necessary for compliance with a legal obligation to which the controller is subject; or
d) processing is necessary in order to protect the vital interests of the data subject; or

e) processing is necessary for the performance of a task carried out in the public interest or in the exercise of official authority vested in the controller or in a third party to whom the data are disclosed; or
f) processing is necessary for the purposes of the legitimate interests pursued by the controller or by the third party or parties to whom the data are disclosed, except where such interests are overridden by the interests for fundamental rights and freedoms of the data subject which require protection under Article 1 (1).

2.2.2 GDPR

With the purpose of strengthening data protection laws, directive 95/46/EC was replaced by the General Data Protection Regulation (GDPR) in May 2018. Due to the rapid technological advances, we face today, the European Union acknowledged that it was necessary to create new legislation to ensure data protection against new privacy challenges.

The General Data Protection Regulation stated the following:

- Subject-matter and objectives [2]:

a) This Regulation lays down rules relating to the protection of natural persons with regard to the processing of personal data and rules relating to the free movement of personal data.
b) This Regulation protects fundamental rights and freedoms of natural persons and in particular their right to the protection of personal data.
c) The free movement of personal data within the Union shall be neither restricted nor prohibited for reasons connected with the protection of natural persons with regard to the processing of personal data.

- Ten main requirements of GDPR [4]:

a) Lawful, fair and transparent processing
b) Limitation of purpose, data and storage
c) Data subject rights

d) Consent
e) Personal data breaches
f) Privacy by Design
g) Data Protection Impact Assessment
h) Data transfers
i) Data Protection Officer
j) Awareness and training

Since the enforcement of GDPR, all platforms that collected, stored, and processed data were and are required to follow it. This means all platforms had to evaluate and modify their technologies to meet GDPR requirements [10].

2.3 Fitness Networks

2.3.1 What are they?

Fitness tracking networks are software applications designed to help individuals track their fitness journey. They can help users track and monitor their fitness, sleep, diet, health, etc. Often these applications are used in combination with wearables to collect and upload the data to the application.

Some of the common uses for these types of networks are:

1. Exercise tracking: allows users to track the type, duration, and intensity of their workouts while giving them feedback about their performance.
2. Diet tracking: allows users to log their meals and keep up with their diet, calories, and nutrition.
3. Sleep tracking: allows users to monitor their sleep patterns, by monitoring their hours of sleep along with things like REM sleep.
4. Wellness tracking: allows users to track their wellness through the monitoring of their heart rate and stress levels.

Majority of people when they think of data privacy, social media apps such as Facebook, Instagram, and TikTok come to mind. However, fitness networks have grown in immense popularity over the last decade within the fitness community. They allow

users, to connect with others, share, and track their progress and achievements. They have since become a staple not only in athletes' lives but also in people's everyday routines for basic functions such as step counting, sleep and heart rate tracking, and many more.

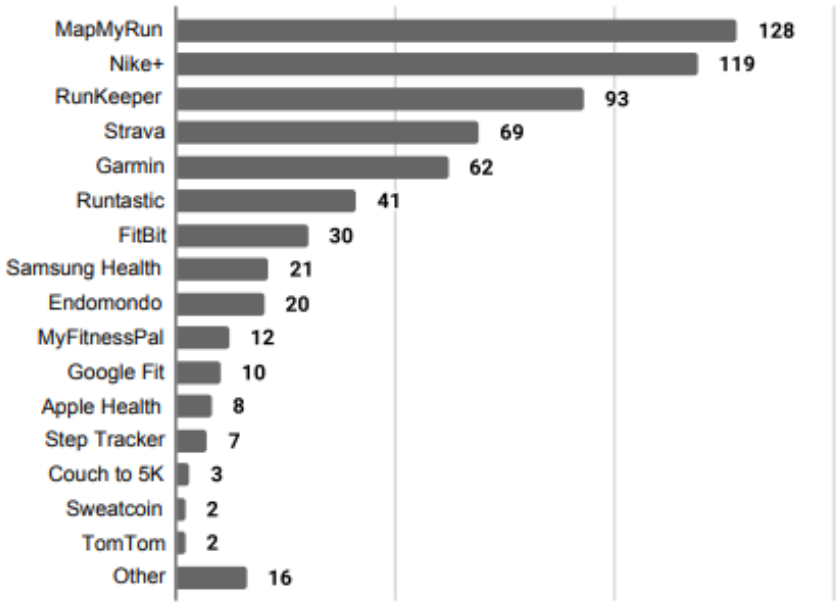

Figure 2. 1: Prevalence of different fitness applications [8].

With the rise in popularity of these platforms, a concern for privacy has also risen. These applications allow users to upload personal data such as age, weight, health issues, and location. If personal data about someone is exposed, they could fall victim to identity theft, their data being used for targeted advertising without their consent, and unwanted exposure of a person's health issues to others such as their employers. Lastly, access to one's location raises concerns for someone's safety as they could easily fall victim to stalking. These are just a few privacy concerns that come with these networks. Users should consider these issues and review their setting options and the data they are uploading with great caution.

2.3.2 Strava

Strava currently has approximately 95 million active users and is one of the world's leading fitness-tracking applications [15]. It was originally launched in 2009 by Mark Gainey and Michael Horvath in San Francisco, California, and has been updated several times over the years.

Similarly, too many other fitness networks, Strava allows users to track their workouts, such as running, cycling, swimming, and other activities. Through Strava, and GPS data, users can compete with friends, enter virtual challenges, and compete for top spots on leader boards. Strava offers a segment feature that allows users to compete in specific routes, popular routes include Central Park in New York, Alpe d'Huez, located in the French Alps, and Beach Road in Melbourne, Australia.

Overall, Strava has created a community of like-minded people, that use the app to connect and compete. When used with caution it can be a great tool to encourage and motivate people to meet their fitness goals. The app has also been used to organize charity events, challenge events, and other virtual fitness events. However, with millions of people uploading their personal information over the years, issues and concerns are bound to arise.

Over the years Strava has had a few data breaches [12]:

1. Global Heatmap: Strava heatmap feature is an interactive map that shows the most popular routes taken by Strava users around the world. Through this feature, the location of US military bases was exposed.
2. Leader boards: Although many Strava users have private accounts, users that participate in segments have their names and other information uploaded to these leader boards, exposing their data.
3. Bicycle Theft: Since users are recording their cycling segments, and exposing their location, they can and have become victims of bicycle theft.
4. Attack on Privacy Zone: To not reveal the exact location of users, Strava uses a privacy zone to mask the exact starting and finishing point of a route. However, in 2018 it was proven that the exact location could be exposed, leaving individuals vulnerable.

With such information being exposed by Strava, the question of what else is available to the public arises.

2.3.3 Leader boards

One of the most popular features of Strava is its leader boards. When an athlete runs in a relatively popular segment, they are added to a ranking that allows them to compare their effort. This leader board ranks these athletes by their effort, with the best effort at the top and the worst at the bottom. For instance, one of the most popular segments, Central Park in New York City, has over 40 thousand efforts. If someone has a Strava subscription, they can view all 40 thousand of those efforts, otherwise, they can only view the top ten athletes. Strava's Leader boards are a crucial feature to keep the community alive and competitive.

After completing a segment, your effort is uploaded to the leader boards. When added to these boards your information now becomes public. An athlete's username, which is often their full name, the date they ran, the pace and time it took to complete the segment, and of course, the route is now publicly displayed on these boards.

Figure 2. 2: Example of a segment map in Strava.

It is important to note that many of the accounts appearing on these boards are private. If people with private accounts have utilized this setting because they care about their privacy, their personal information appearing on these boards can be controversial. Even if athletes with private accounts are aware that they are appearing on these boards, are they aware of the safety risks? For instance, their location and the date they were at that location are available and could potentially be dangerous.

Chapter 3
Exploring Privacy Issues

3.1 Comprehending the significance of protecting data.

It is no coincidence that there are so many laws in place to protect users' data. In today's digital age, information is constantly at risk of being exposed or misused. Thus, it is important to prioritize safeguarding our infromation from any unauthorised access, misuse, and abuse. Understanding the significance of protecting your data on a personal level is crucial. Controlling who can see your information gives you a sense of privacy and safety and a greater sense of control over your digital life. Furthermore, by being aware of the online risks that come with data sharing, one can make more informed decisions about their online behaviour and take steps to protect themselves from potential harm.

3.2 Privacy Issues in Fitness Based Social Networks

Fitness based social networks have grown in popularity over the years. Allowing users to connect with others with similar interests, track their fitness journey and compete in a friendly environment. However, with the increase in popularity more and more people are talking about the privacy concerns associated with them. subject Due to the nature of the platforms, a user's data, including GPS routes and exercise logs, can be easily

accessed by others. This means that anyone with a public account can fall victim to cyber-attacks or stalking.

As mentioned previously, Strava is no stranger to data breaches and privacy concerns. After the heatmap incident which disclosed the location of secret military bases and other sensitive areas, many people have expressed concerns about what other privacy issues could be associated with the app. Specifically, which is what I will be analysing in the sections that are to follow, are privacy concerns surrounding leader boards. When someone's account is public, it is easy for anyone to simply view their account, and draw conclusion about their life, their patterns, and habits and even where they live, in fact 95.1% of regular Strava users are considered to be at risk of having sensitive information such as their homes and other locations exposed [17]. When a user is consistently running, at the same time on the same day of every week, it is easy for people to track them. For anyone with a sense of awareness surrounding privacy, this wouldn't come to them as a surprise, as the accounts are public and available to everyone. But even with private accounts, your privacy is not guaranteed. After completing any segment, your effort is uploaded to a leader board, whether your account is public or private. This would mean that even if you care about protecting your privacy, the public can still see your information through these leader boards. While for some, this might not be to alarming, for other who take their privacy with utmost seriousness it may discourage them from using platforms such as Strava. For this reason, I believe making the appropriate changes to ensure a user's privacy includes the development of a privacy preserving leader board.

Chapter 4
Architecture

Fitness platforms like every other, have privacy regulations they must follow and implement. However, even when following all regulations and guidelines every platform comes with privacy concerns. Even seemingly harmless networks like Strava have features that could raise some concerns for someone looking a little deeper. A closer inspection of the private account feature in Strava is necessary to understand its true significance. Despite Strava's leader boards being a popular and frequently used feature, a closer examination reveals a flaw. The appearance of private accounts (which indicates a concern for one’s privacy and data protection) on these very public leader boards raises questions about how this feature aligns with Strava's commitment to user privacy. Given the privacy concerns surrounding these leader boards, we undertake an in-depth investigation into this issue and consider strategies for resolving it. As part of this effort, we analyse the data obtained from the leader boards and develop a script to mask user information, with the goal of preserving the privacy and security of Strava's users.

4.1 Finding data.

To better understand the privacy concerns associated with Strava’s leader board, we conduct a comprehensive investigation of this issue. To thoroughly investigate and develop an effective solution, it is vital that we have full access to the data within these

leader boards. Without it, our ability to accurately analyse the data and develop an accurate solution may be limited. To do this, downloading these leader boards, is essential for future steps.

As a first attempt to download data related to Strava's leader boards, I aim to use Strava's very own API. It quickly became apparent that the API has limitations, in fact, the API only returns the information owned by the authenticated athlete, which made it difficult to collect the data needed for the analysis. Since using the API was unsuccessful, alternative methods had to be sought, leading to the development of a script to achieve this.

Viewing Strava's leader boards is available to anyone even if they don't have an account. However, only the top ten athletes are available, only with a premium account is it possible to view the entire leader board, and as a result, I upgraded my account to premium. With the full leader board now accessible, the next step was to develop the script that would allow me to download it, and extracted the necessary data displayed within the leader boards.

With the help of these scripts, I was able to overcome the initial obstacles and further my research. Additionally, with now having access to the data, I would hopefully be able to draw more meaningful conclusions and insights about the issue at hand. Thus, allowing me later to formulate a set of recommendations and solutions that could help address the privacy concerns related to the Strava leader boards.

4.2 Gathering data.

4.2.1 Number of private accounts

As mentioned in previous parts, it is important to understand Strava's dedication to its user's privacy. For this reason, I found it necessary to investigate the number of private accounts within these leader boards. This step was necessary to see the extent of the problem.

Keeping score of the number of private accounts within these leader boards did not come without its difficulties. To achieve this, it is necessary to visit all user's accounts

and note their account status. It is important to note that these leader boards can have thousands of athletes, and this is where issues arise. While running a script to access all these profiles, I soon realised that Strava has limitations on the number of requests someone can make, which significantly slowed down the process. Despite the issue, I stayed persistent and was able to find the number of private accounts within the leader boards I had.

This finding is an important step in continuing my research while also developing a better understanding of the issue. With such a big number of private accounts on these leader boards, one could argue that this information can help platforms identify areas where they need to improve their privacy features and policies to ensure that users feel comfortable sharing their fitness data on the platform.

4.2.2 Seeking Correlations.

Sometimes looking for correlations between variables can reveal insights that might not be immediately obvious. For this reason, I chose the next step in my research, to be just that, to look for any possible correlation between the athletes and their privacy preferences. By doing this I hoped to gain a better understanding of influences users to choose a private account and why that might be.

I decided to use three variables to test for any correlation between the account preferences. Firstly, I decided to check the ranking of a user on a leader board. If a user is higher up on a leader board, the higher the chances are that they take their fitness very seriously. Therefore, perhaps it could show that users with higher interest will most likely prefer for their efforts to be publicly available. In contrast to users that are lower on the leader boards, who possibly might just be your average person running to stay healthy and do not care if people see their efforts.

The second variable I decided to investigate was the date on which the effort was made. As the years go by and technology advances so do privacy concerns, meaning that they were not as prevalent early on, and most people were not as concerned. Thus, users who have been on Strava for a longer time may have signed up before they were as concerned about privacy, meaning that they perhaps might be more likely to still have

public accounts. In contrast newer users could be more educates about the importance of privacy and might be more likely to have private accounts.

The third and final variable I choose to use was age of the athletes. I hypothesised that age, might be a factor in an athlete's decision for their account's privacy settings. Older athletes, that grew up with less technology in their lives, might be more likely to be unaware of risks that are associated with having an online presence. In contrast, younger people, people who have had technology embedded in their lives from early childhood might be more aware of what they are posting and the risks that come along with it.

4.3 Solving the issue.

With a full understanding of the issue and importance of privacy, the next phase was to create an alternative, privacy-preserving leader board. Developing an algorithm that would mask the data which would allow for a user's original data to stay on the user side while the masked data is sent to the server side. This could potentially solve the privacy issues that occur with private accounts appearing on leader boards for networks such as Strava.

This solution would allow a user to remain anonymous, while still being able to maintain the competitive spirit of fitness networks such as Strava. Thus, allowing users to still compete and know their ranking without compromising their safety and privacy. This idea could also be applied to other similar platforms with similar privacy concerns to the ones discussed.

Chapter 5
Implementation

In this portion I will be thoroughly discussing the steps I have taken to develop the necessary scripts and algorithms to address the problem at hand. I aim to provide a more detailed explanation of the approaches taken and the challenges I encountered throughout the process.

5.1 Technical environment.

Each programming language has its own pros and cons and it's important to choose the right one for you and your research. Python is a popular choice in the industry due to its simplicity, it has a relatively simple syntax which also allows for people without a background in programming to comprehend it. Additionally, Python is known for its wide range of libraries and frameworks that allow for developers to integrate prebuilt tools and functions into their projects which can significantly speedup the development. Python is also known for its excellent support in various fields of Computer Science such as data analysis, machine learning and task automation [1]. Overall Python is a well-rounded and reliable programming language and allows a developer to achieve their goals with ease and efficiency. For these reasons, I have utilized Python for the

development of my work. Pythons' simple syntax, and extensive set of libraries such as NumPy, Pandas, Selenium, Requests proved to be very valuable during development.

5.2 Accessing Leader Boards

5.3 Analysing Account Status

Now that we have extracted the necessary data, we can begin to grasp the extend of the issue at hand. Analysing the data will allow us to identify patterns and trends that can reveal the extent of the problem. By establishing the number of private accounts versus public accounts on the leader boards, we can determine the number of individuals that not only care about their data and privacy but are being affected by a possible privacy issue of Strava. Understanding the extent of the issue, and understanding who is concerned about their data, is crucial for finding and developing solutions. Overall, analysing the data is an essential step that will provide us with a better understanding of the problem and help develop better strategies for a better solution.

5.3.1 Privacy Preferences

With the larger number of people among these leader boards, I found it crucial to analyse the percentage of private accounts on these leader boards to further understand the issue at hand.To achieve this, it was necessary to check the account status of each athlete on the leader boards I had downloaded. Of course, manually carrying out this process would be too time-consuming and would require a great deal of effort. Thus, I found it necessary to develop way to automate the process. While downloading the leader boards, I not only stored the information that is displayed within them such as name, pace, and time, but I also kept track of each athlete's unique ID. This proved to be important in this part of my research as it allows me to visit each user's account with ease.

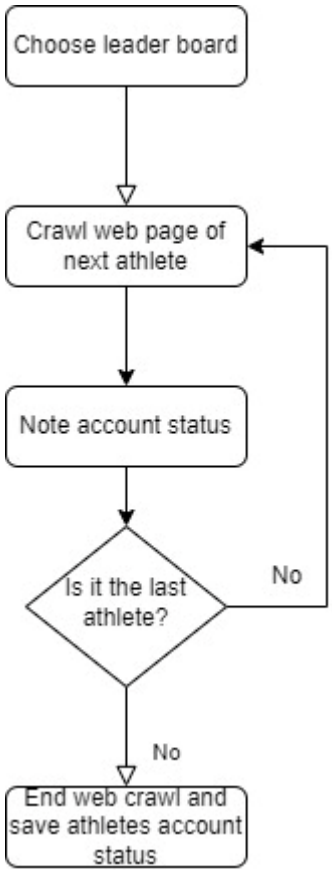

Figure 5. 6: Main idea to find account status.

As seen in figure 5.6, we can see the idea I followed to note the account status of the athletes. However, my goal was not only to match each account with their account status but to also keep track of the overall number of private and public accounts. To achieve this, I developed a script that would allow me to keep score of the number of private versus public accounts within the leader board, while also allowing me to match the account with account status.

I automated the process by allowing the script to read the unique ID of each athlete and creating two variables to keep track of the number of public and private accounts. The script opens a text file and updates the number of private and public accounts during the current running session. The script uses the Requests library to automate the process of logging into the Strava website and accessing each athlete's account. For each athlete the script finds their account status by locating a specific HTML element within their profile page. Finally, the code updates the private and public account counters to the text file, which will then be used in the next session to continue where it left off.

The process of finding the account status was time-consuming. One of the biggest obstacles I faced was Strava's limitation on the number of requests I could make. Each account access was one request, and with some of the leader boards having up to around 40 thousand requests, it became apparent that this would be a long process. Since the limitation was specific to the number of requests an account could make, this meant that I had to create multiple accounts so that I could access more profiles per session. However, decided to switch from using Selenium library to using the Request library, which allowed me to overcome the issue of request limitation. Moreover, since Request library did not have to render the HTML code, an increase in performance was observed. Although, I still had to be patient and persistent, as the process required a significant amount of time and effort. Despite the difficulties the insights gained were valuable and necessary for understanding Strava's users.

5.3.2 Trends and Patterns

The next step was to find any patters in the athletes' choices for their account status. Correlation between variables can reveal insights that might not be immediately obvious and can help understand the needs and preferences of Strava's user base. I decided to use three variables to test for any trends and patterns within the leader boards. As mentioned, I choose to look for any correlation between account status and an athlete ranking on a leader board.

I created a script that will calculate the percentage of private accounts in chunks of the leader board. The script goes through the CSV file containing the leader board and loops over the data in chunks tailored to the specific leader boards based on the number of efforts. For each chunk of rows, the code counts the number of 'private' and 'public' accounts, and the percentage of 'private' accounts is then calculated as a percentage of the total number of accounts in the chunk. The results for each chunk are then outputted into a new CSV file, where we can visualise the data by creating graphs to better understand the results.

Similarly, I decided to look for any correlation between account status and the efforts date. Privacy concerns have become more prevalent with the advancements of technology over the years. This led me to believe that users may not have been as

concerned about privacy as they are now. Thus, users that joined Strava earlier may have their accounts public, due to privacy not being such a hot topic back when they joined. In contrast, newer users, might be more likely to be aware of privacy concerns and have their accounts private.

To test this theory, I created a script that would allow me to determine if there is a correlation between the date on which the effort was made and the choice of account status. The script opens the CSV file in which the leader board is stored, and it finds the oldest and newest year in the data. After, it loops through each year in the range from the oldest to the newest year and counts the number of private and total entries for that year. If there are entries for that year, it calculates the percentage of private entries. If there are no entries for that year, it simply outputs a message indicating so. The results again are put into a new CSV file, which will be used to create graphs and gain a better understanding of the data.

Lastly, I looked for a correlation between an athlete's age and their account privacy setting. The older and younger generation have had different experiences with technology, the older generation have had technology slowly enter their lives while watching it advance rapidly. While the younger generation has always had some level of technology in their lives, it is something that they grew up with and are more familiar with. For this reason, I sought to find out if Strava's users are more likely to have specific privacy settings based on their age.

Testing this was a straightforward process as I already had the account status of every athlete on the leader board. When you update your account to premium, you can filter the leader board by age group. So, all I had to do, was simply use the same script to download the main leader board, to download the filtered leader board by age. Then, I easily matched each athlete's privacy setting by comparing it with the main leader board. This was easy as I just had to merge the filtered leader board with the main one by matching athlete unique ID. After being able to match the account status to each athlete, I then found the percentage of private accounts in each age-filtered leader board. Finally, to better understand and present my findings, I used Excel to create graphs and charts to showcase the relationship between account status and the athlete's age.

5.4 Privacy Preserving Leader Board

The main goal of this section is to thoroughly explain the process of making a privacy-preserving leader board. The goal is to ensure the protection of a user's data, while still maintaining the competitive spirit that the leader board feature provides for its athletes. I will be describing the technical steps involved in the creation of such leader boards.

The main issue with the current leader board that Strava has is the lack of anonymity. Most Strava users probably enjoy leader boards, it gives them a goal to work for, and a sense of competitivity. But not everyone likes their data being available for the whole world to see. People might not like the idea of others being able to see when and where they ran, for safety reasons, and others might just like to keep their fitness journey and efforts private. For this reason, a privacy preserving leader board would be a good middle ground.

To create this privacy-preserving leader board that protects a user's sensitive information I developed a script to mask their data, while still maintaining the same ranking. To ensure anonymity of a user, it is essential to keep the data secure, meaning there are no possibilities for any third parties to access the information. This would mean that the server of the social network would not know the information that is currently displayed on these leader boards, but rather it would have the information masked, while preserving the original data on user side.

The implementation of this script could be adopted by Strava and other fitness networks, to ensure user privacy. As mentioned, this script allows for a user's real data to never be sent to the server, but rather saving the real data on the user's side and sending the masked data to the server. This can be integrated into the upload process of a fitness network, as soon as an athlete completes their run, and just before the data is uploaded to the server, the data could be run through the script and then sent off. This will ensure that no one could find this information from the fitness network server and making the server agnostic to everyone's data.

5.4.1 How it works.

For the purposes of the implementation, I used data from already existing leader boards. The script reads the data that the leader board would need, such as name, date, time, and

pace and uses various methods to hide the original data. For instance, for the name, we want each user to have a unique id number that represents them, without anyone being able to track it back to the athlete and their account. For this reason, in the script, I hashed the original name to get a unique like username by using the SHA-256 algorithm. For the date, I simply just kept the year of the effort. This would allow users to compare efforts made, without exposing the day or month, which is a much safer alternative. For an athlete's pace, the code takes a pace time in the format of "MM: SS" and converts it into a 'datetime' object, I then extract only the time part of the datetime object. And a "score" is a calculated based on the pace. The score that can be used for analysis without revealing the actual pace. And finally for the time, I used the same logical approach as for the pace and converted it into a score that can be analysed without revealing the exact time.

Chapter 6
Evaluation

Throughout the research there have been two main questions. The first being about the demographic of fitness network users and specifically what influences their decisions surrounding their privacy settings, and who is more likely to prioritize their privacy. The second question focused on addressing what could be done about privacy issues surrounding the leader boards. In the previous sections I dove into detail about the measures I took to answer these questions. I explained the concepts, the issues, the questions, and the development process. And know that I have both understood the issue at a bigger scale, and worked to solve it, my next step is to analyse my findings and answer the initial questions. Finally, by doing so we can hope to better understand the issue and propose recommendations for fitness network providers to improve their privacy policies and practices.

6.1 The Statistics

To answer our first question regarding who prioritizes their privacy, I analysed many accounts based on three things, their age, their ranking in a leader board, and the time the effort was made and uploaded to the leader board. I stored each accounts privacy setting (public or private) and continues to find statistics based on this.

6.1.1 Account Preferences: Public or Private?

First and foremost, I investigated the number of private versus public accounts. As mentioned, I stored each user's account status, after downloading the leader board. I did this process for three different segments, Vondelpark in the Netherlands, Central Park in the United States, and a segment in Athalassa Park in Cyprus. Each of these segments' leader boards vary in size.

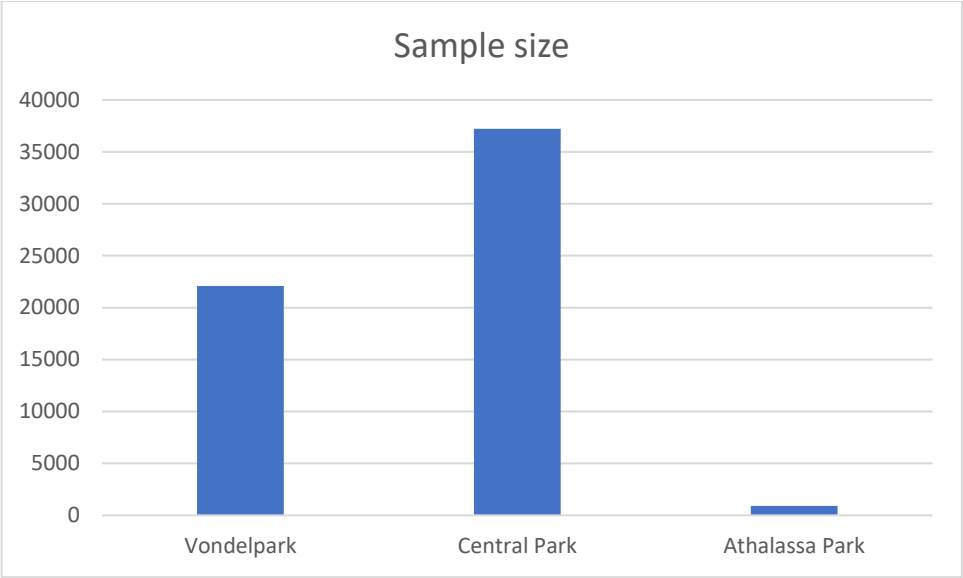

Figure 6. 1: Sample size of leader boards.

As seen in figure 6.1, I have three different sample sizes. The leader board in Central Park is the largest of the three with 37,208 athletes, the leader board from Vondelpark has 22,074 athletes, and finally the smallest leader board, from a segment in Athalassa park with only 917 efforts.

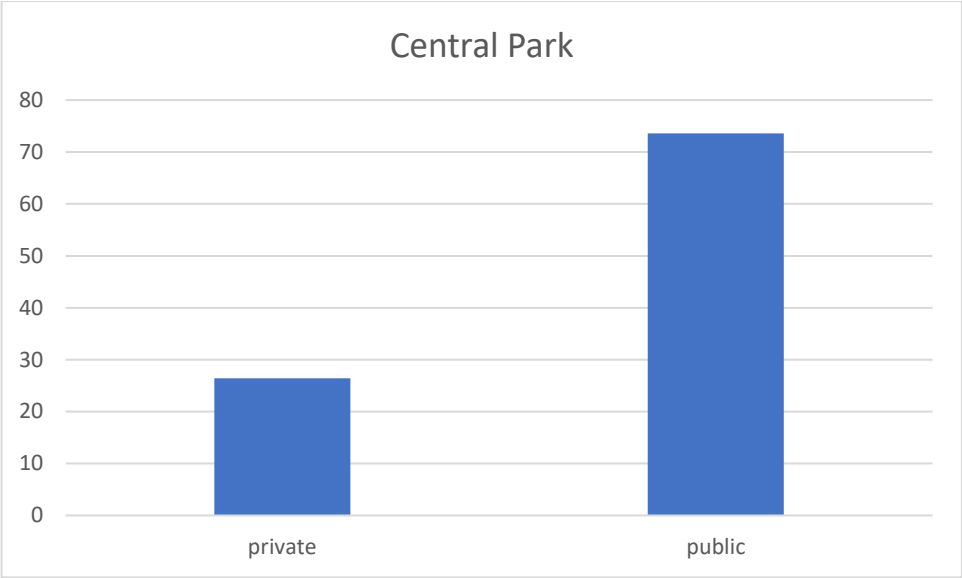

Figure 6. 2: Percentage of public vs private accounts for Central Park.

After exploring each user's privacy setting on Strava, I calculated the percentage of private and public accounts on the Central Park leader board. I found that surprisingly about 26% of users had private account while about 74% had public accounts.

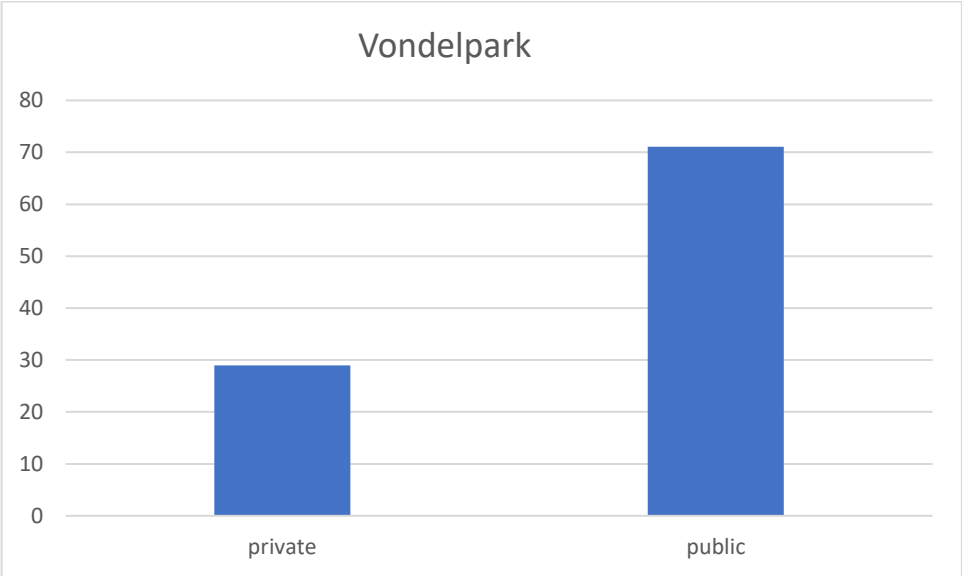

Figure 6. 3: Percentage of public vs private accounts for Vondelpark

Similarly, we see the percentage of public vs private account on the Vondelpark leader board. We have about around 29% private accounts and around 71% public accounts.

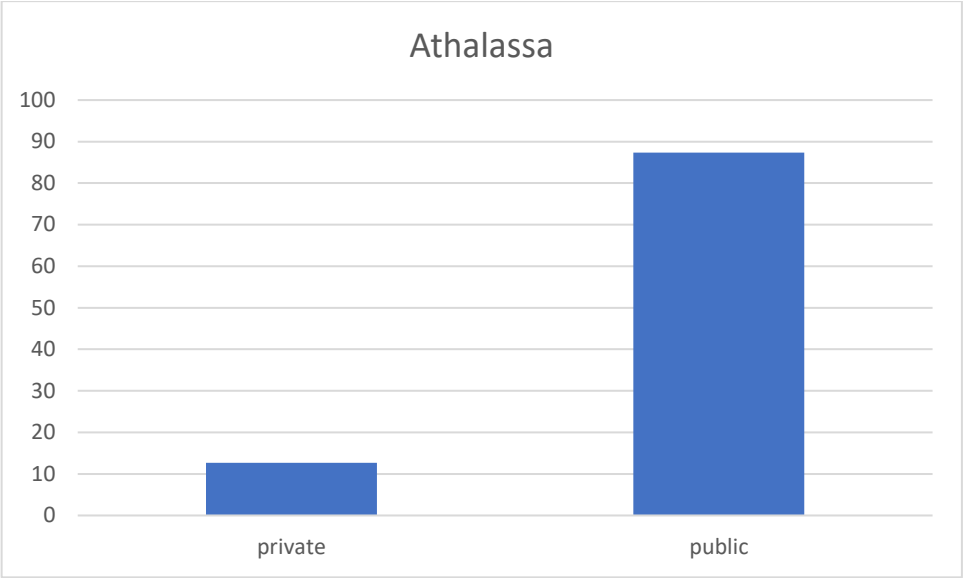

Figure 6. 4: Percentage of public vs private accounts for Athalassas Park.

Finally, here we can see the percentage of public versus private accounts for Athalassa park in Cyprus. Again, we see a relatively low number of private accounts at around 13%, while public accounts make up around 87% of the leader board.

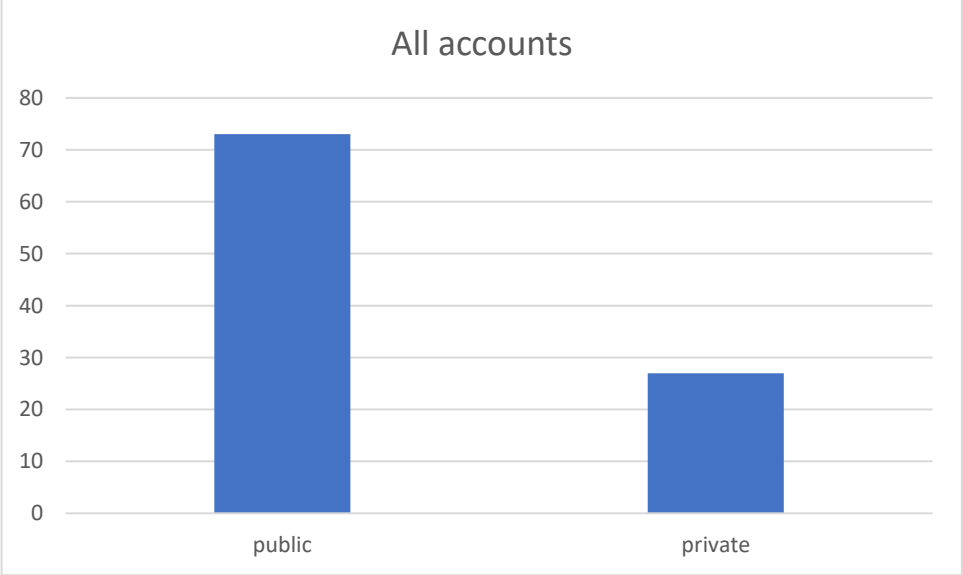

Figure 6. 5: Overall percentage of public vs private accounts.

As we can see on all three leader boards, most accounts are public with the overall percentage being around 73%. However, that still leaves around 27% of private accounts appearing on these leader boards. The percentage of private accounts

appearing on these leader boards raises concerns about privacy and data protection. As I have touched on before, many users may not want their data to be publicly available, yet their efforts are still being displayed. Moreover, the percentage of public accounts being so high could also be alarming, and an indication of a lack of awareness surrounding privacy issues and safety.

6.1.2 Distribution by Ranking.

After finally investigating the scale of the issue at hand and gaining a better understanding about the magnitude of athletes with private accounts on leader boards, I decided my next goal was to understand what influences people in their decisions regarding privacy settings. Was it possible that people with higher rankings on leader boards cared less or maybe more about their privacy? Perhaps, since they are better athletes and clearly take their efforts more seriously, it could be possible that they care less about their privacy, and more about sharing their efforts. Or perhaps athletes with higher rankings use Strava for example more often, therefore they are more likely to keep their accounts private to be safe.

To answer my question, I studied the leader boards that I previously mentioned. I created a script that would find the percentage of private accounts, within a certain split. This would allow me to observe any patterns between ranking and privacy settings if there were any.

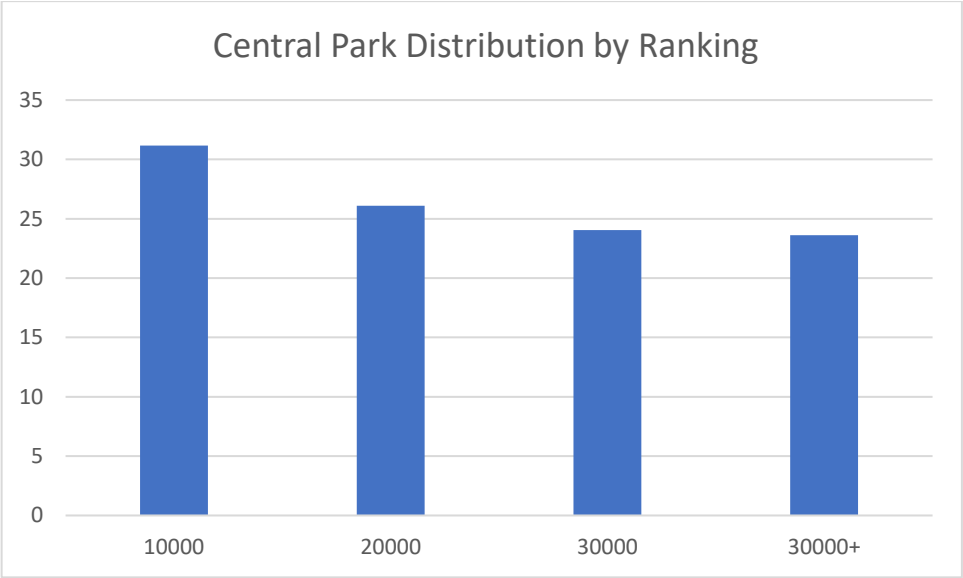

Figure 6. 6: Percentage of private account by ranking (split per 10000) for Central Park

For the Central Park leader board, the data shows that among the first 10,000 athletes, approximately 32% had private accounts. Between 10,000 and 20,000, there was a decrease in the percentage of private accounts to around 26%. In the range of 20,000 to 30,000, around 24% of athletes had private accounts. Finally, among the remaining accounts, a little under 24% had private accounts.

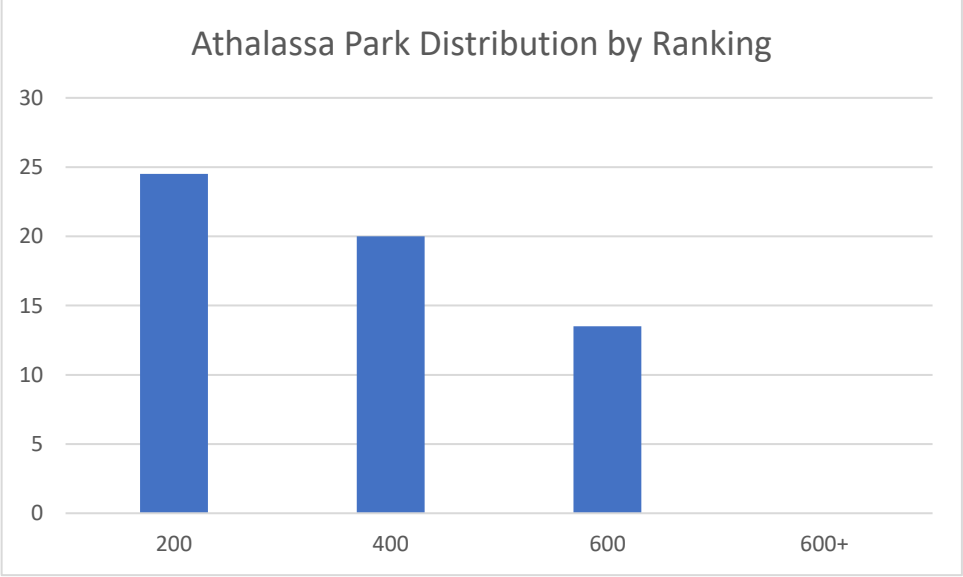

Figure 6. 7: Percentage of private accounts by ranking (split per 200) for Athalassa Park.

Also, for Athalassa Park, we can see that for the first 200 athletes the percentage of private accounts is about 26%. Between 200 and 400 it is around 20%, between 400 and 600 there was a significant decrease to around 14%. Finally, towards the end of the leader board there were no private accounts at all.

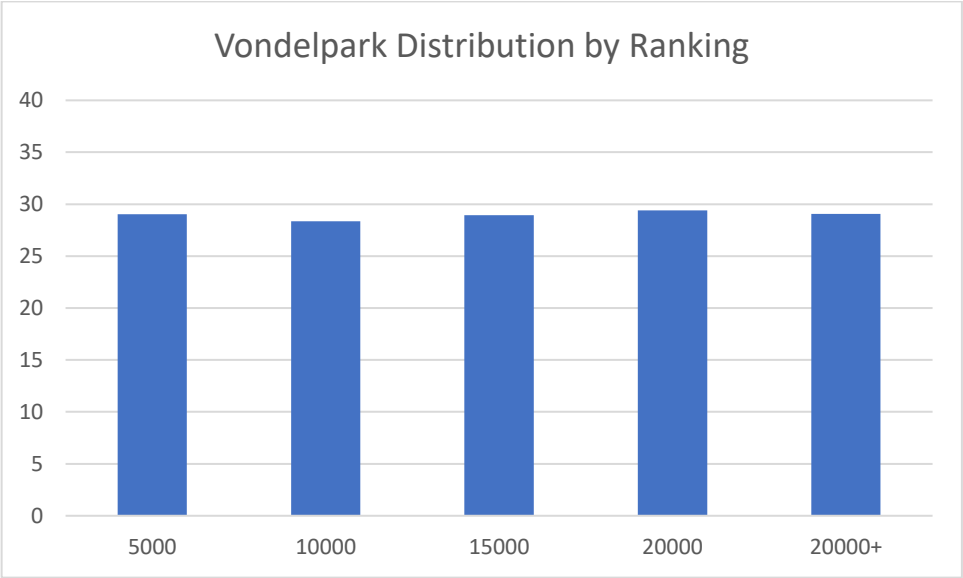

Figure 6. 8: Percentage of private accounts by ranking (split per 5000) for VondelPark.

For Vondelpark, however, we see a nearly consistent number of private accounts with each split having around 28 to 29 percent.

The data presented from graphs 13 and 14 could suggest a correlation between an athlete ranking on a leader board and their privacy preferences. It could imply that athletes who are more competitive and are ranked higher on the leader board may be more likely to protect their privacy. However, after also analysing the distribution for the leader board of VondelPark, which can be seen in graph 15, we see that the number of private accounts remained around the same throughout the leader board. This could suggest that there is no significant correlation between the rank and privacy settings for athletes.

The findings may not have given a clear answer about what influences an athlete's decision to keep their account private or public. However, it does raise several other questions about how factors such as the location of the segment, the culture surrounding fitness and privacy in that area, and the demographic of the athletes in the area, could

play a significant role in the percentage of private accounts. Furthermore, since out of all three areas Vondelpark had the highest percentage of private accounts throughout the entire leader boar, this could further show that location and culture could be a factor.

6.1.3 Distribution by Effort Date

After analysing the percentage of private accounts by athletes ranking, I wanted to further expand on the search for a correlation between athletes and privacy preferences. For this reason, I sought to find for a possible corelation between privacy setting and effort date. I hypothesised that individuals who signed up for the fitness network in earlier years might be more likely to have public accounts because privacy concerns were not as prevalent at the time. Thus, I investigated the percentage of private accounts for each year of efforts on the leader boards I have access to.

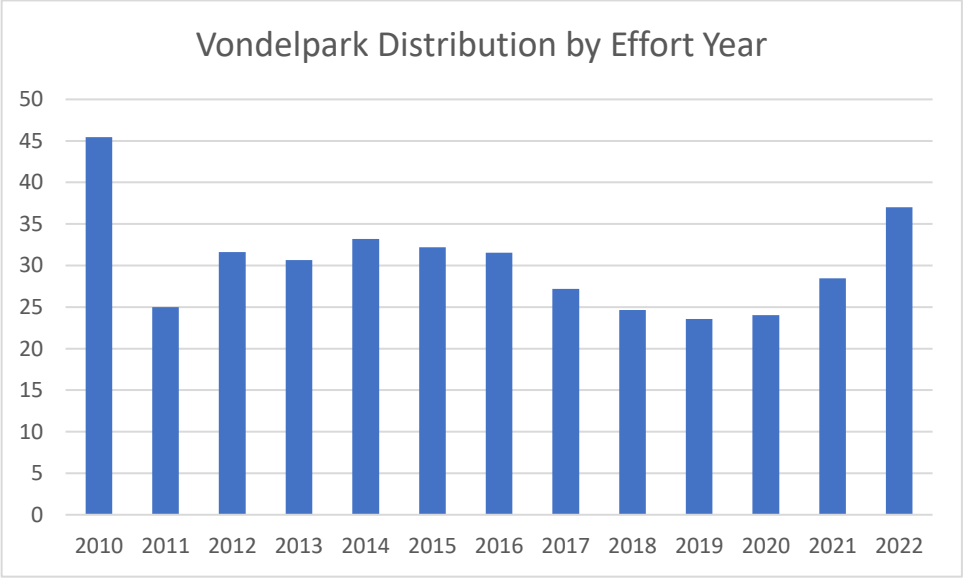

Figure 6. 9: Percentage of private accounts for Vondelpark by year of effort.

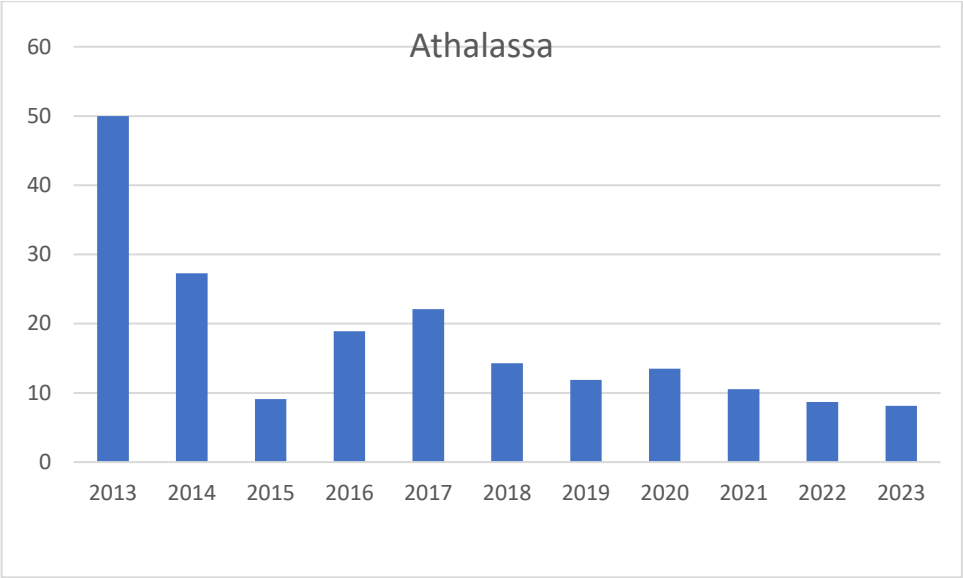

Figure 6. 10: Percentage of private accounts for Athalassa park by year of effort.

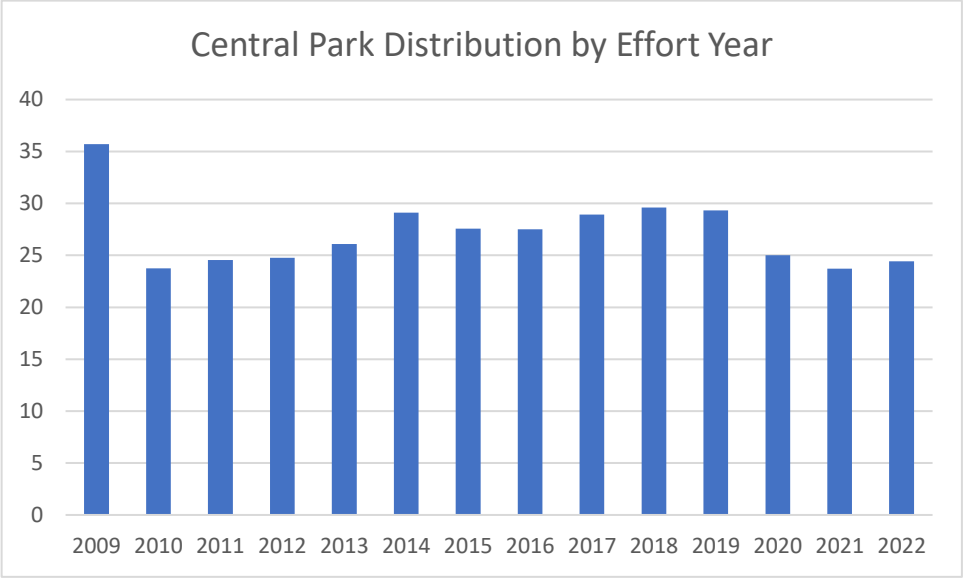

Figure 6. 11: Percentage of private accounts for Central Park by year of effort.

After analysing the percentage of private accounts within each effort year, I found no significant correlation between privacy preferences and effort date. In fact, the percentage of private accounts seemed to generally be evenly distributed throughout the years, with a few spikes that cannot be explained by any underlying trend or pattern. This would suggest that there is no pattern between the athlete's effort date and privacy preferences. It is important to note, that I only took into consideration efforts with a date

from 2009 and later. Although the leader boards contains earlier efforts, Strava was founded in 2009 and the very first effort was recorded on April 12, 2008 by one of the company's first employees, Davis Kitchel [3].Thus, I believe any efforts recorded before this date may not be legitimate or reliable enough to be taken into consideration.

6.1.4 Distribution by Age

Finally, the last hypothesis I wanted to investigate was the correlation between age, and privacy preferences. I hypothesised that generational differences could play a major factor in an athlete's decision for the privacy preferences. Perhaps younger generation were indifferent about their privacy due to the greater exposure to technology. On the other hand, older generation could be more cautious of technology and the safety of their data. It is possible to also observe the opposite, we might see that growing up in an era of rapid technological advancement, allowed for the younger generation to be more aware of the dangers, thus they might be more likely to prefer private accounts, and older generations might use the default public account.

To investigate this hypothetical corelation between generations and privacy preferences, I extracted the leader boards filtered by age group from Strava and found the percentage of private accounts in each age group. I decided to use the leader board from the Vondelpark segment as the sample size was not too big and not too small, and it also was the segment with the highest percentage of private accounts out of the three. Thus, I found it to be the most appropriate to use. Below we can find the analysis of this research, and by finding the percentage of private accounts, I hope to find some pattern, to help further understand what influences the decision regarding one's privacy settings.

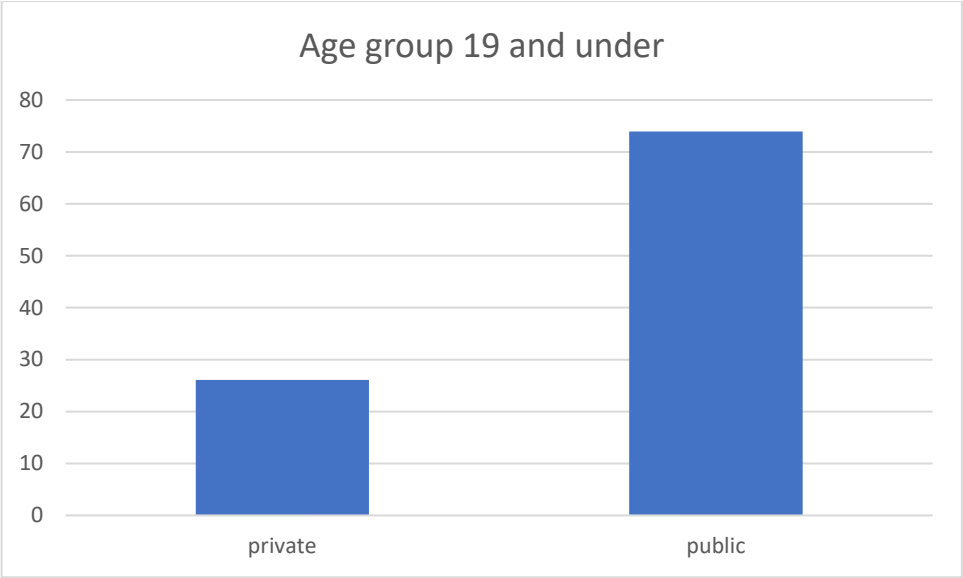

Figure 6. 12: Percentage of private accounts for athletes 19 and under.

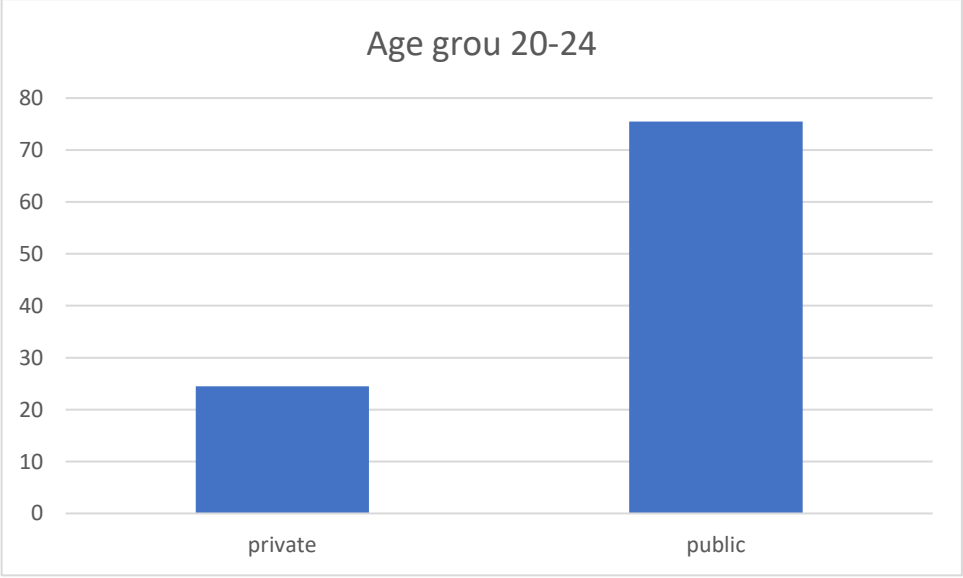

Figure 6. 13: Percentage of private accounts for athletes between 20 and 24.

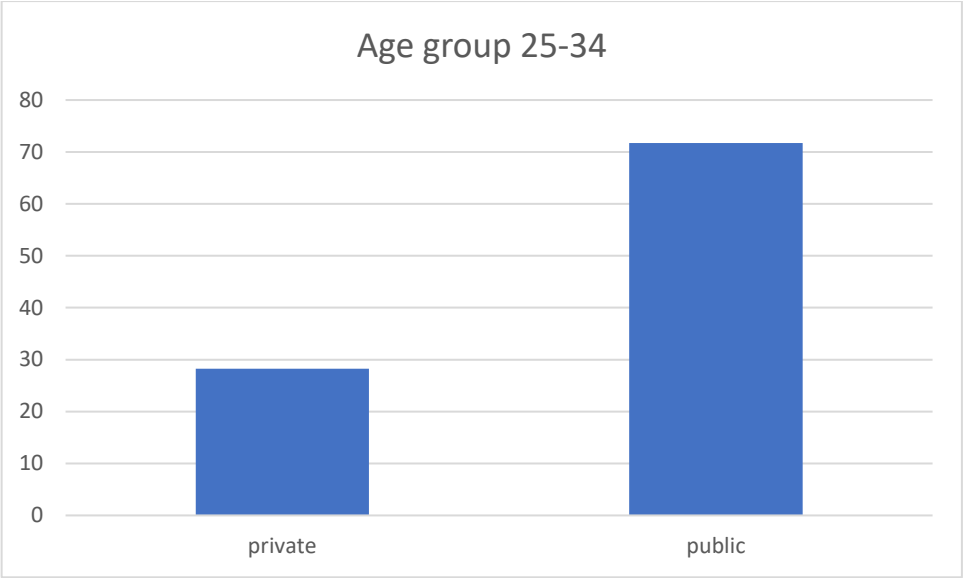

Figure 6. 14: Percentage of private accounts for athletes between 25 and 34.

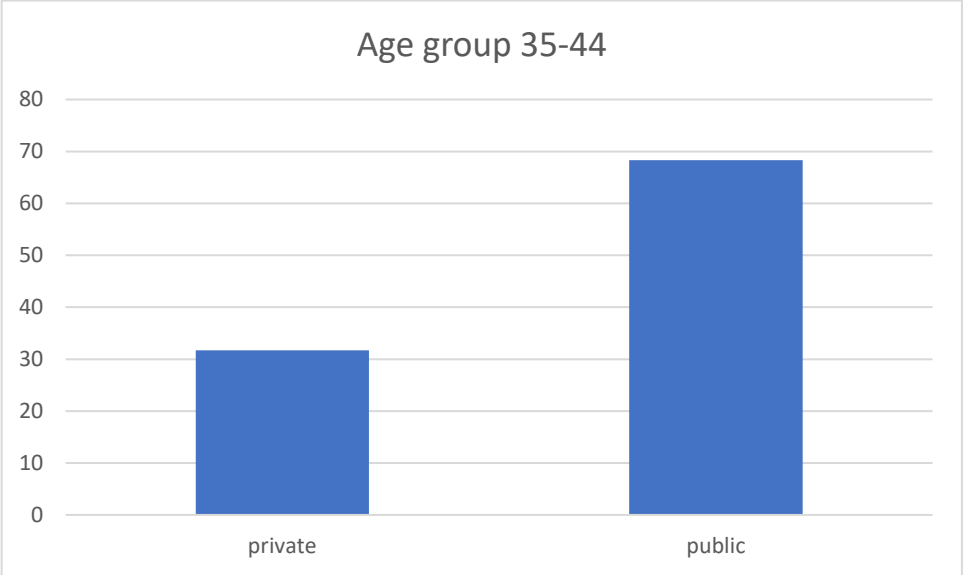

Figure 6. 15: Percentage of private accounts for athletes between 35 and 44.

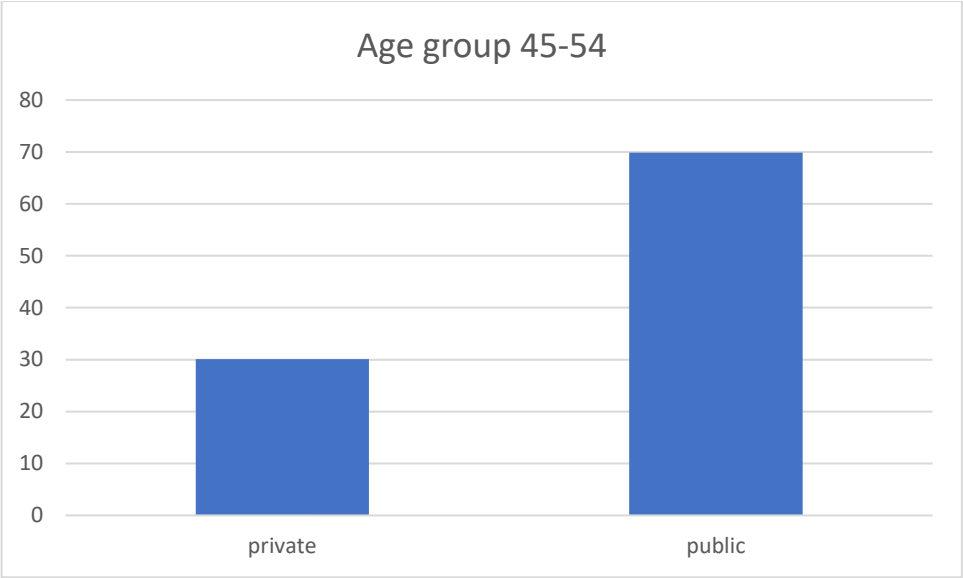

Figure 6. 16: Percentage of private accounts for athletes between 45 and 54.

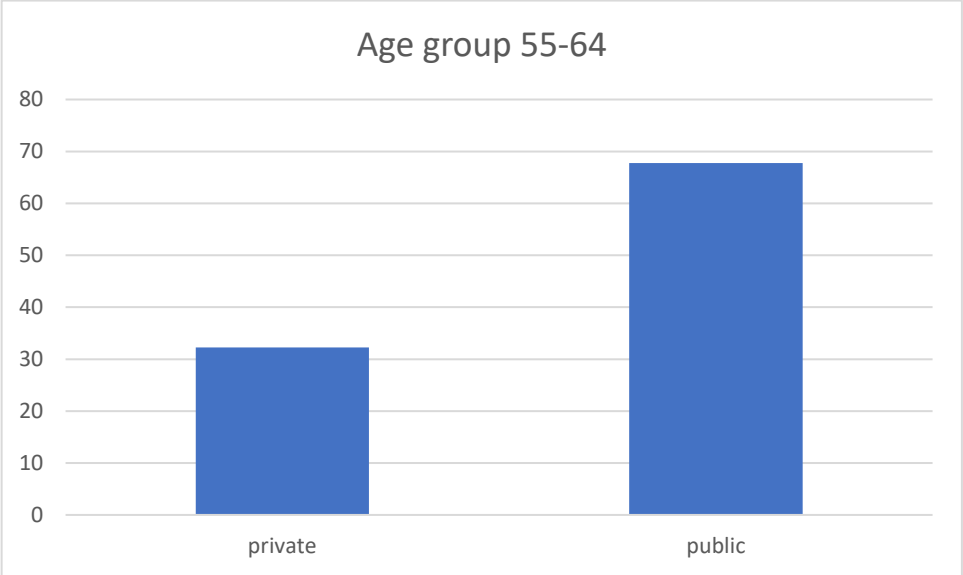

Figure 6. 17: Percentage of private accounts for athletes between 55 and 64.

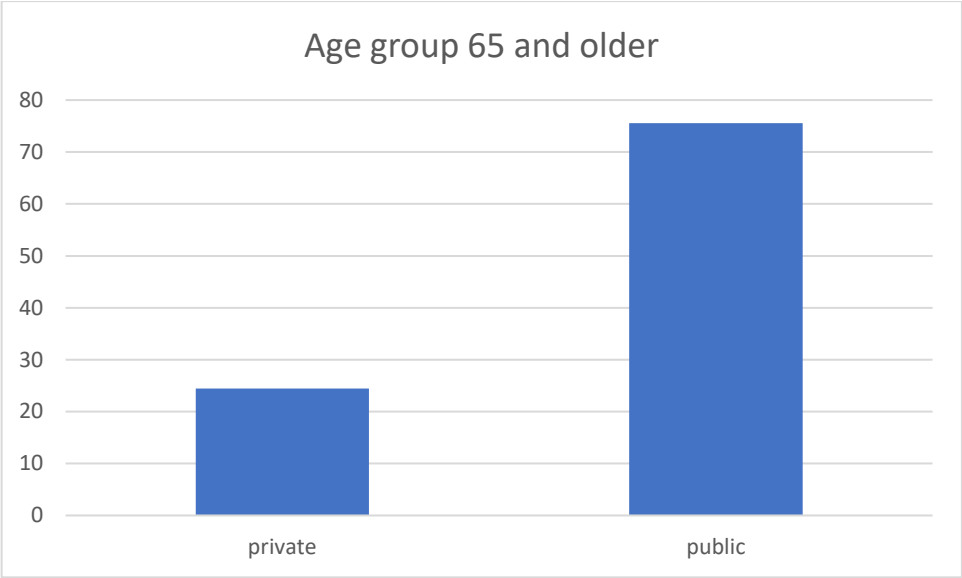

Figure 6. 18: Percentage of private accounts for athletes 65 and older.

In figure 6.12 we can see the percentage of private accounts for the age group 19 and under. The sample size for this age group was quite small, with it having only 395 athletes. After analysing the leader board, I found that 26% of the athletes had private accounts. For age group 20 to 24, we had a sample size of 2553 athletes. As seen in figure 6.13, the percentage of private accounts is around 25%, very similar to the age group of 19 and under. However, we can start to see a slight increase in the percentage after these two age groups. For the age group of 25 to 34 with a large sample size of 10849 athletes, we can see in figure 6.14 that around 29% of users have private accounts. Similarly in figure 6.15 for the age group of 35 to 44, we can observe the largest percentage of private account, at around 32% with a sample size of 4931 athletes. Moving on to figure 6.16 we can see that for the age group of 45 to 54 there is also a high percentage, with around 30% of accounts being private. Additionally, for the age group of 55 to 64 as seen in figure 6.17, we can also see a high percentage of private accounts at 32%. Lastly the percentage of private accounts drops for the age group of 64 and older, as seen in figure 6.18, the percentage of private accounts drops dramatically to 24%.

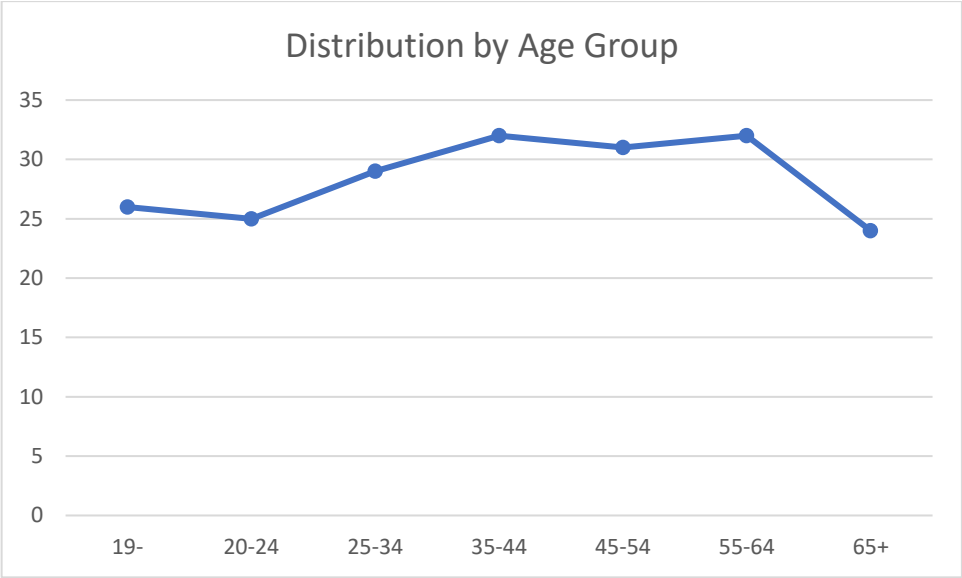

Figure 6. 19: Overall percentages of private accounts by age groups.

As we can see through figure 6.19, most age groups had a percentage close to the overall 29% of the leader board as seen in figure 6.3. However, we see a that for the age group of 24 and younger there is surprisingly a low percentage of private accounts. This could be because generation Z (which are currently between the ages of around 26 and under) have grown up with technology embedded in their lives. From childhood they have been exposed to technology and social media. Thus, they may be more willing to share personal information online and may not perceive privacy concerns in the same way that older generations do. Similarly, we also see that the older age group of 65 and older which could be defined as Boomer generation (which are currently around 60 to 77) are not as tech-savvy and are potentially unaware of the risks and dangers that come with an online presence. The in between generations, millennials (currently around 27 to 42) and generation X (around 43 to 58), are two generations that had technology slowly and gradually enter their lives [5]. Thus, it is possible that they are more likely to be cautious of their personal data being exposed or misused. Additionally, due to their age, they are more likely to have a mature and educated approach surrounding the issue of online privacy.

It is important to note that this research was only based on one leader board. For this reason, the finding may not be representative of the entire Strava user base. However, one thing is for sure, there is a large degree of people that care about their privacy, and it

is important provide assurance to these users that their data and privacy is safe with the platform they rely on.

6.2 Conversion Outcome.

After finally analysing these leader boards, I developed a better understanding of the issue and the extent of it. These insights meant I finally felt confident enough to come up with an idea and carrying it out. A balanced solution that satisfies both fitness enthusiasts and privacy-conscious individuals. A new type of leader board that would allow users to feel confident enough in the application to safeguard their data, while also being able to enjoy the perks of a leader board, such as the competitive spirit.

As stated earlier, I used both hashing methods and various masking methods to hide the original data from the server. This would allow for the real data to be stored locally and the masked data to be stored on the server. Thus, still allowing for a ranking system to be formed.

First and foremost, it was necessary to create a unique ID that would mask each athlete username. This was a crucial step towards ensuring the privacy of each athlete. By using a hashing function that would generate a unique identifier for each user and it would allow users to compete anonymously with other athletes. In figure 6.20 we can see an example of a hashed username.

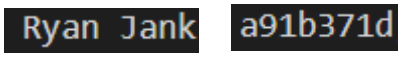

Figure 6. 20: The conversion of username to a masked version.

Next, I decided to conceal the effort date. Doing so we can ensure that any patterns and habits of users are not exposed, thereby preserving their privacy. This step is crucial to a user's safety and ensures no one will be able to track when a user runs or exercises. I found that problems arise when the entire date is visible for everyone, therefore I found that by only showing the year in which someone completed the effort, you can preserve their safety and still allow for a competitive atmosphere. For example, if the date was, 19-Sep-2022, each athlete would only know be able to see that another athlete completed their effort, in 2022.

Moving on, the next two crucial parts of the leader board is the pace and time of an athlete. As previously stated, to hide this information, and still maintain a leader board feature, I decided to mask the data, to still allow a ranking system. This would mean, that an athlete could still see where they are on a leader board in comparison to other athletes.

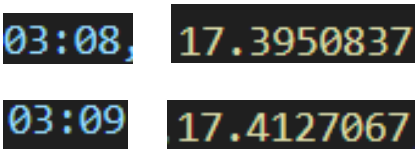

Figure 6. 21: The conversion of an athlete's pace to a masked version.

As seen in figure 6.21, we can see the format in which the new pace is masked. We can see that a ranking system, still applies, the fastest pace corresponds to the smallest number and the slowest pace to the largest number. In figure 6.21, we see the pace of two athletes, the first place we see 03:08 is the pace of the first athlete on a leader board, and the pace, 03:09 is the pace of the second athlete on a leader board. As we can see after the masking process, the first athlete's pace is masked to the number 17.3950837 and the second 17.4127067. This means that an athlete is still able to be ranked correctly.

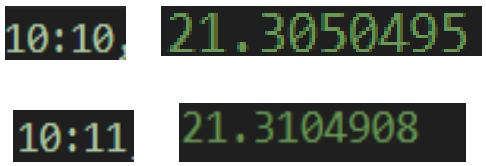

Figure 6. 22: The conversion of an athlete's time to a masked version.

In figure 6.22, we can see the format in which the new time is masked. Similarly, to the masking process of the pace, we can see that again a ranking system is still in place without revealing the time in which it took an athlete to complete its effort. As seen in figure 6.22, we can see that yet again, the fastest time, 10:10, was masked to the number 21.3050495, and the slower time of 10:11, was masked to the number 21.3104908. Thus, allowing for users to still be able to compare efforts and compete.

6.3 New Possibilities.

After working on a solution that could possibly satisfy both privacy and fitness enthusiasts by attempting to mask sensitive information, such as usernames, dates of efforts and paces. I aim to showcase possible implementation of a privacy-preserving leader board, that could be adopted by platforms such as Strava. These prototypes, incorporate the efforts to mask the data of a leader board I mentioned in previous sections, while still maintaining the feel and theme of the platform it is integrated into. I will be using Strava as an example platform.

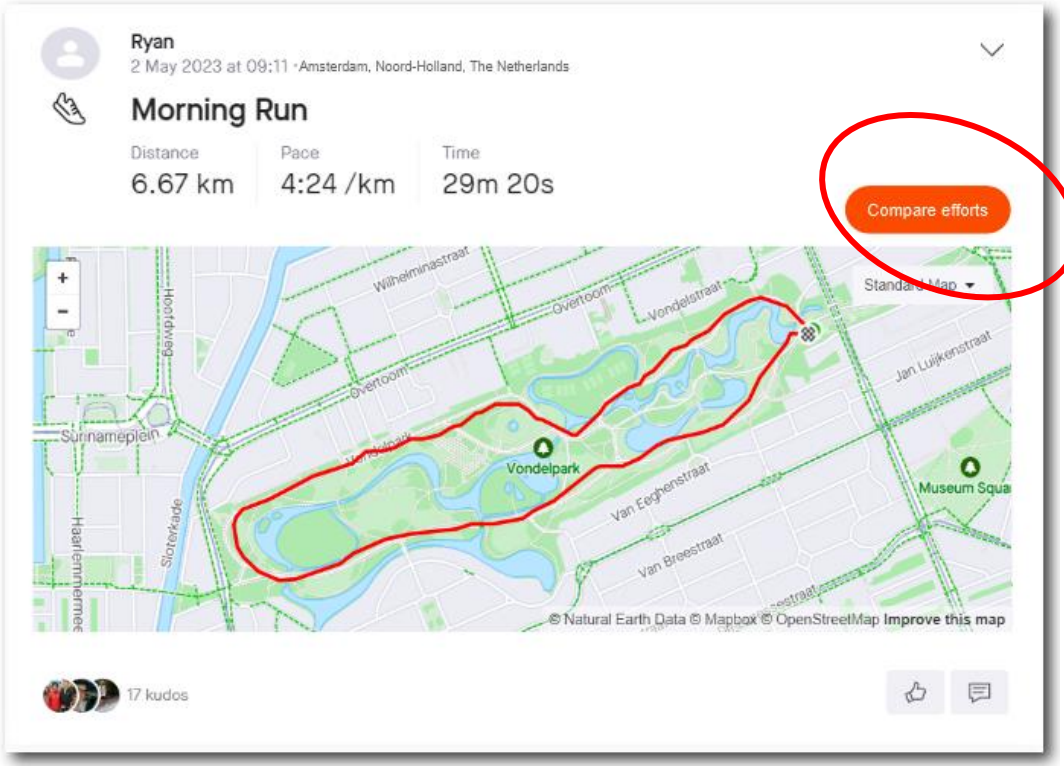

Figure 6. 23: An example of an effort on Strava after adapting to a privacy preserving leader board.

Figure 6.23 showcases a changes Strava would have to adapt to integrate the new privacy preserving leader board. In Figure 6.23 we can see, an effort made by an athlete, after an they compete their effort and the activity is uploaded to their account, they should have an option to compare their efforts with other users. After clicking on the option to compare efforts they will be redirected to another page.

6.2.1 First approach

The first approach for the new leader board, is percentage based ranking system. Since our main goal is for the server of the platform to not know the original data of an athlete, we have sent the masked versions to the server. Therefore, the server will have a ranking system, of the athletes based on the masked score. This would allow for the platform, in this case Strava, to inform an athlete, in what percentile they are in. In Figure 6.24 below, we can see a protype of what this adaptation could possibly look like.

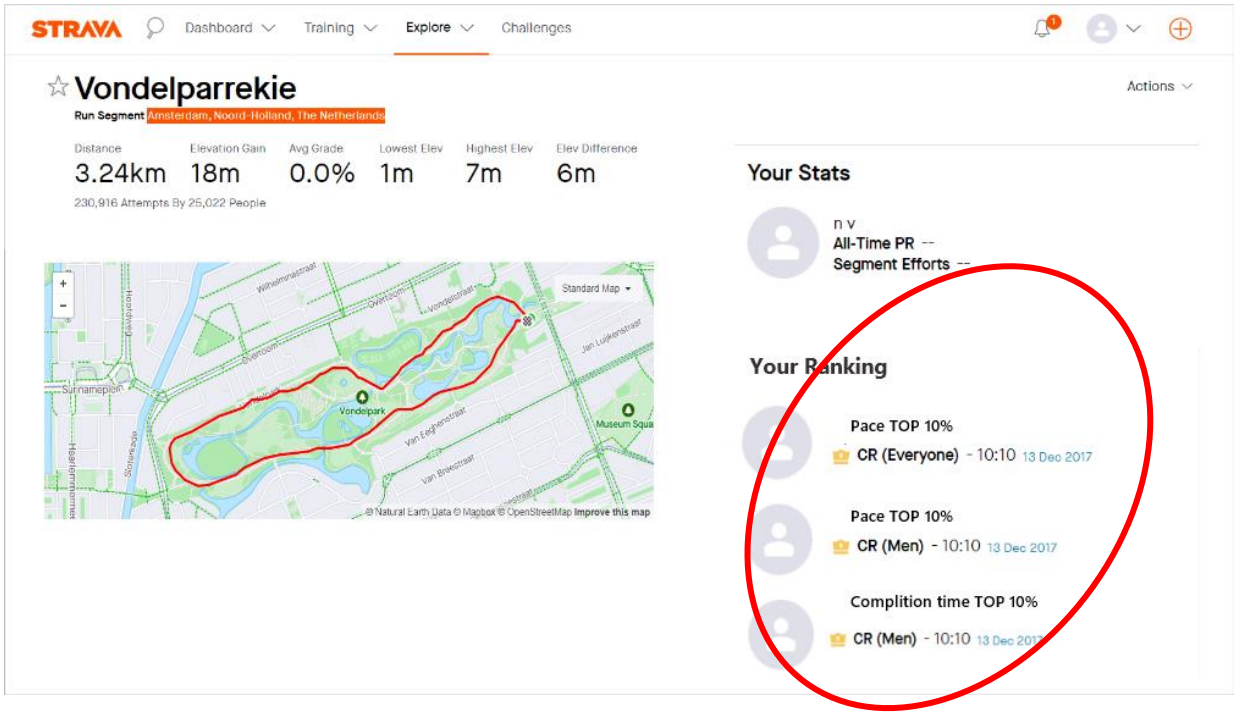

Figure 6. 24: Percentile System Example.

As we can see, after a user chooses to compare their efforts, they will be redirected to a page that could look like this. Similarly, to the original page of Strava, this page includes the name of the segment, the location, and all information about the segment such as distance and elevation and other information that is related to the number of attempts, this can help give perspective to the user without revealing to much information. To the left of the screen, we can see the athletes ranking. We can see that it gives them the percentile they belong to for their time and pace, among everyone and among their sex. This ranking system will allow users to preserve their data while still allowing them to compare themselves in relation to other efforts made, which preserves the original edge that comes with a leader board.

6.2.2 Second Approach.

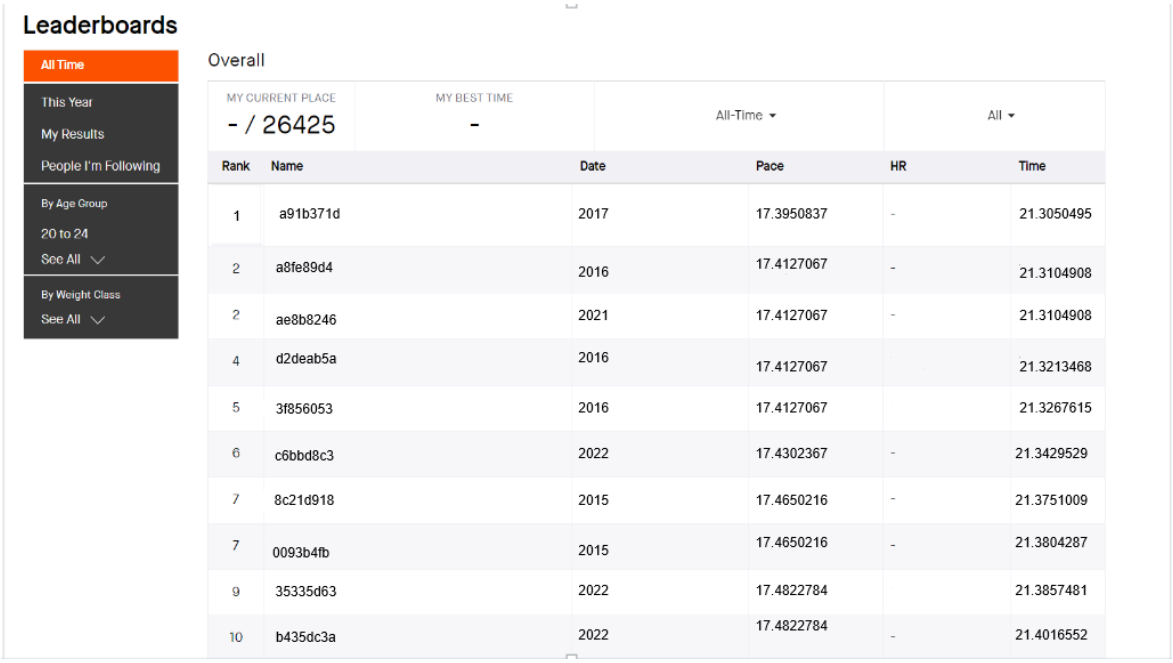

Figure 6. 25: Masked Leader board example.

The second approach would be to show the privacy-preserving leader board as it is. As seen below in figure 6.25 instead of the original data being showed, we will have the masked versions. This way each athlete will be able to see their ranking on the leader board based on their unique ID. With this leader board, an individual will still be able to

compare their efforts by viewing the masked data, it will allow them to have an estimated understanding of how close they are to the athletes above and below them. Finally, by still using unique IDs, a user can still share their IDs with people they trust, such as friends and family for some friendly competition.

Chapter 7
Related work

In this section we will examine previous work and research related to privacy issues of fitness based social networks. These publications aim to provide a comprehensive analysis of the privacy risks associated with fitness-based social networks, while others aim to explore a user's understanding to how these platforms work.

7.1 User Perceptions and Knowledge of Privacy in Social Fitness Networks.

In the paper Are Those Steps Worth Your Privacy? Fitness-Tracker Users' Perceptions of Privacy and Utility a study (N=227) was conducted to understand how users perceive the utility of their features and the privacy risks that are associated with them [16]. This paper showed that many individuals that use fitness trackers have a limited and incomplete understanding of how they work, what data they collect and what could be inferred from fitness tracker data.

Figure 7.1 depicts some of the research's findings. According to the results, many users were aware of some of the information that could be inferred from their fitness trackers. However, many failed to realised personality traits such as religion, political views, and sexual orientation could not be deduced from fitness tracker data. This however is not true, some research has indicated that these personality traits could in fact be inferred, for example someone could tell where someone is attending religious services [19]. It also shows that many users, are not worried about certain data being inferred.

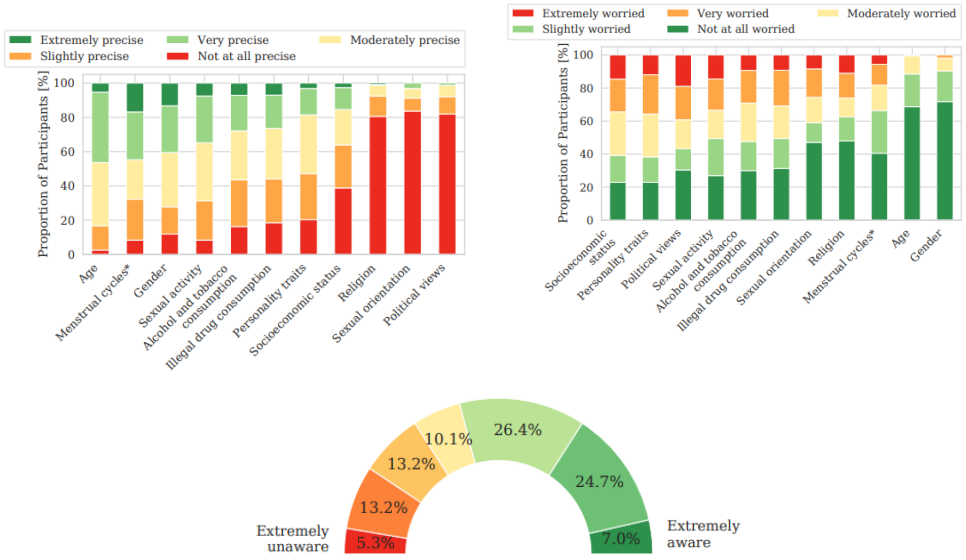

Figure 7. 1: Results of survey regarding a user's awareness of privacy risks, perceived precision of inferring sensitive information and level of concern.

7.2 Inferred location risks

In a survey (N=603), 45% of participants were somewhat or very comfortable sharing a map of their exercise [8]. While in another survey, (N=60) found that 90% participants indicated their start of activity is either home, school, or work, and an overwhelming 98% of the participant indicated those to be the end point of their activities [11]. Additionally, in another research it was found that 95.1% of regular Strava users are at risk of having sensitive locations such as their homes exposed [17].

Chapter 8
Conclusion

The aim of my thesis was to grasp the privacy issues related to fitness networks and I am confident this was achieved to a significant level by the end of the thesis I touched on the privacy issues related to location data being exposed, I discussed the potential risks regarding private accounts appearing on leader boards, and I came up with a possible solution to mitigate the issue at hand.

8.1 Future work

In terms of research and implementation, there is still much to be done. While my thesis made progress in uncovering the privacy issues associated with fitness networks and their leader boards, further work is needed. Due to the time limitation of my thesis, I could not fully analyse the privacy risks and sensitive infromation that could be inferred from the leader boards, nor could I fully develop a privacy preserving leader board.

To implement a solution any problem it is key to understand the problem and the needs of those using the system. The next steps in this research would be to thoroughly research these privacy issues, and further understand a user's view on them. Once this is achieved, further steps should be taken towards implementing solution that caters to both privacy concerns and the need for a ranking system for athletes on fitness networks.

8.3 General Conclusion

In conclusion, I believe that everyone with an online presence should take measures to protect themselves and their data. As seen in the previous sections of my thesis, there are many online safety and privacy concerns even where we least expect it. Just like any other platform, fitness networks have been subjected to various data breaches and instances of data misuse. Thus, understanding and finding privacy issues within these platforms is vital to safeguard users' data. It is important for networks such as Strava to recognize potential problems in their features such as leader boards. This would allow them to then work towards a solution that would allow them to maintain the feel of a leader board without compromising one's right to privacy.

Bibliography

[1] *Applications for python Python.org*. URL: https://www.python.org/about/apps/ (Accessed: 28 October 2022).

[2] "Art. 1 GDPR – Subject-matter and objectives - General Data Protection Regulation (GDPR)," *General Data Protection Regulation (GDPR)*, Aug. 30, 2016. https://gdpr-info.eu/art-1-gdpr/ (Accessed: 25 January 2023)

[3] BikeBiz, "One billion activities recorded on Strava," *BikeBiz*, Mar. 2019, [Online]. Available: https://bikebiz.mystagingwebsite.com/one-billion-activities-recorded-on-strava/ (Accessed: 3 March 2023)

[4] P. Bhatia, "10 key GDPR requirements: A short summary," *Advisera*, Dec. 2022, [Online]. Available: https://advisera.com/articles/a-summary-of-10-key-gdpr-requirements/ (Accessed: 2 January 2023)

[5] K. Brunjes, "Age Range by Generation - Beresford Research," *Beresford Research*, Jan. 19, 2023. https://www.beresfordresearch.com/age-range-by-generation/ (Accessed: 12 March 2023)

[6] A. Dini Kounoudes, G. M. Kapitsaki, and I. Katakis, "Enhancing user awareness on inferences obtained from fitness trackers data," *User Modeling and User-Adapted Interaction*, Jan. 2023, **Published**, doi: 10.1007/s11257-022-09353-8.

[7] EUR-Lex - 31995L0046 - EN - EUR-Lex," *EUR-Lex - 31995L0046 - EN - EUR-Lex*. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A31995L0046 (Accessed 18 December 2022)

[8] Jaron Mink Amanda Rose Yuile Uma Pal Adam J Aviv and Adam Bates. 2022. Users Can Deduce Sensitive Locations Protected by Privacy Zones on Fitness Tracking Apps. In Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (New Orleans LA USA) (CHI '22). Association for Computing Machinery New York NY USA Article 448 21 pages. https://doi.org/10.1145/ 3491102.3502136

[9] H. Jiang, J. Li, P. Zhao, F. Zeng, Z. Xiao, and A. Iyengar, "Location Privacy-preserving Mechanisms in Location-based Services," *ACM Computing Surveys*, vol. 54, no. 1, pp. 1–36, Jan. 2021, doi: 10.1145/3423165. [Online]. Available: http://dx.doi.org/10.1145/3423165

[10] H. Li, L. Yu, and W. He, "The Impact of GDPR on Global Technology Development," *Journal of Global Information Technology Management*, vol. 22, no. 1, pp. 1–6, Jan. 2019, doi: 10.1080/1097198x.2019.1569186. [Online]. Available: http://dx.doi.org/10.1080/1097198x.2019.1569186

[11] U. Meteriz-Yildiran, N. F. Yildiran, J. Kim, and D. Mohaisen, "Learning Location from Shared Elevation Profiles in Fitness Apps: A Privacy Perspective," *IEEE Transactions on Mobile Computing*, pp. 1–16, 2022, doi: 10.1109/tmc.2022.3218148. [Online]. Available: http://dx.doi.org/10.1109/tmc.2022.3218148

[12] Meteriz Ü. Yıldıran N.F. Mohaisen A.: You can run but you cannot hide using elevation profiles to breach location privacy through trajectory prediction (2019). arXiv preprint arXiv:1910.09041

[13] D. Ruby, "Smartwatch Statistics 2023: How Many People Use Smartwatches?," *Demand Sage*, Mar. 06, 2023. [Online]. Available: https://www.demandsage.com/smartwatch-statistics/ (Accessed 3 April 2023)

[14] Strava Developers," *Strava Developers*. [Online]. Available: https://developers.strava.com/docs/reference/#api-SegmentEfforts-getEffortsBySegmentId (Accessed 14 March 2023)

[15] Strava Revenue and Usage Statistics (2023)," *Business of Apps*. [Online]. Available: https://www.businessofapps.com/data/strava-statistics/ (Accessed 2 May 2023)

[16] L. Velykoivanenko, K. S. Niksirat, N. Zufferey, M. Humbert, K. Huguenin, and M. Cherubini, "Are Those Steps Worth Your Privacy?," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 5, no. 4, pp. 1–41, Dec. 2021, doi: 10.1145/3494960. [Online]. Available: http://dx.doi.org/10.1145/3494960

[17] Wajih Ul Hassan Saad Hussain and Adam Bates. Analysis of privacy protections in fitness tracking social networks-or-you can run but can you hide? In USENIX 2018.

[18] What is Privacy," *What is Privacy*. [Online]. Available: https://iapp.org/about/what-is-privacy/ (Accessed: February 2023)

[19] [1]"Inferring religious beliefs from fitness data," *John D. Cook | Applied Mathematics Consulting*, Apr. 01, 2019. [Online]. Available: https://www.johndcook.com/blog/2019/03/31/fitness-data-privacy/ (Accessed 28 April 2023)

Appendices

```
driver = webdriver.Edge(executable_path="msedgedriver.exe")


driver.get("https://www.strava.com/segments/1624225")

driver.get("https://www.strava.com/segments/1624225/leaderboard?age_group
=75_plus&amp;filter=age_group&page=1&per_page=200&partial=true")
time.sleep(60)
p=driver.page_source
file1.write(p)

for index in range(2,450):
    driver.get("https://www.strava.com/segments/1624225/leaderboard?age_g
roup=75_plus&amp;filter=age_group&page="+ str(index)
+"&per_page=200&partial=true")
    p=driver.page_source
    file1.write(p)
file1.close()
```

Web crawler for downloading leader board.

```
line = 0
for id in myList:
    line= line+1
    if line <= public_accounts + private_accounts:
        continue
    # driver.get("https://www.strava.com/athletes/"+str(id))
    response = reqs.get("https://www.strava.com/athletes/"+str(id))
    if(response.status_code!=200):
        break
    if(response.text.find("Follow")!=-1 or
response.text.find("follow")!=-1):
        p = "Public"
        public_accounts = public_accounts + 1
    else:
        p="Private"
        private_accounts = private_accounts + 1
    print(str(id) + " : " + p)
    file1.write(str(private_accounts) + "    " + str(public_accounts) +
'\n')
```

Web crawler to see which accounts are private.