Strava Privacy Analysis: Understanding User Location Vulnerabilities

Rahul R. Jois
Department of EECS
University of California, Irvine
rjois@uci.edu

Maganth Seetharaman
Department of EECS
University of California, Irvine
seetharm@uci.edu

***Abstract*—Fitness applications have surged in popularity as more people turn to technology for assistance in maintaining a healthy lifestyle. These apps offer an array of features, including activity tracking, that encourages users to monitor their health metrics and progress. However, the continuous collection and storage of such detailed personal information raise serious privacy concerns. This study investigates the potential privacy vulnerabilities posed by one such fitness application, Strava. We analyze public data available on Strava, examining both spatial and temporal aspects of user activity data. We employ DBSCAN clustering algorithm for spatial clustering and derive temporal patterns based on the time of the day, aiming to detect key areas and times of user activities. Through the visualization of user data in the form of interactive maps and heatmaps, we identify segments of frequent user activities. Each segment provides details including the probable time range of user presence and the user's most active days of the week. We conclude by discussing potential countermeasures to mitigate identified privacy risks, such as setting profiles to private, concealing start and end points of activities, and masking frequent addresses. Future work involves more sophisticated clustering techniques, real-time alert systems for users, and expanding our study to other popular fitness apps. This research brings to light the significant privacy risks in fitness apps and emphasizes the importance of informed usage and robust privacy settings.**

I. Introduction

Fitness applications have become an integral part of our lives, assisting us in maintaining a healthy lifestyle. These applications employ GPS and other sensors in our smartphones and wearables to track activities, measure performance, and provide useful health insights. However, these benefits come with privacy risks. Many fitness applications, such as Strava, collect detailed personal information, including the geographical location of users and their activity patterns. This data, if misused or mishandled, could potentially lead to privacy breaches, with serious implications for the users.

Strava, for instance, publicly shares detailed activity data of its users by default, making it a gold mine for potential attackers aiming to track or understand the user's lifestyle, activity patterns, and even infer their probable home or work addresses. While Strava's rich datasets provide an excellent opportunity for analysis and pattern recognition, it also underscores the urgent need for robust privacy protection mechanisms.

In this study, we embark on a journey to identify potential privacy risks in the usage of fitness applications, focusing primarily on Strava. We leverage data clustering and visualization techniques to explore user activity data, discern patterns, and identify potential privacy vulnerabilities. Our approach involves spatial clustering, temporal analysis, and generating interactive maps and heatmaps to represent user activity.

By doing so, we aim to highlight the extent of personal information that could be inadvertently disclosed through fitness applications and propose countermeasures to mitigate these privacy risks. Our study underlines the importance of treating personal data with caution and being aware of the potential implications of sharing such data on public platforms. The complete source code, data preprocessing scripts, and analysis notebooks used for this study are available at our GitHub repository: https://github.com/RahulJois/strava-privacy-analysis. We invite interested readers and researchers to explore and use this resource.

Through this report, we seek to create awareness about the privacy risks in using fitness applications and encourage informed and privacy-conscious use of such technology. The report concludes by suggesting potential countermeasures to reduce these risks and identifying directions for future research.

II. Problem Statement

As fitness tracker applications continue to surge in popularity, so does the concern for privacy and security. These applications gather and store vast amounts of sensitive information like geographical location, physical activities, and patterns of user behavior. While this data plays a critical role in delivering personalized insights and motivating healthier lifestyles, it also presents substantial privacy and security risks. These risks stem from potential misuse of the data by malicious actors and inadvertent data disclosure, which could lead to users being tracked or having their personal routines exposed.

The primary problem we aim to address in this study is two-fold. First, we seek to better understand and visualize the spatial and temporal patterns of users' activities based on the data obtained from a popular fitness tracking application, Strava. Second, we aim to identify potential privacy risks and propose effective countermeasures to safeguard user data. The challenge lies in striking a balance between extracting useful insights from the data while ensuring the privacy and security of users' sensitive information.

To tackle these challenges, we employ clustering algorithms to segment user activities based on spatial and temporal data. The goal is to understand typical user behavior and identify high-frequency locations and time frames. Subsequently, we evaluate potential privacy risks that could arise from the disclosure of such information. Based on our findings, we then recommend a set of countermeasures users can adopt to protect their privacy while continuing to benefit from fitness tracking applications.

Overall, this study contributes to the ongoing discourse on data privacy and security in the context of fitness tracking applications. It offers a comprehensive analysis of privacy risks and proposes user-centric solutions that can be implemented to protect sensitive user data, thereby promoting a safer and more secure use of these applications.

III. Related Work

This section explores the existing research in the field of fitness application privacy and security. The discussed works shed light on the vulnerabilities of fitness applications and their implications for user privacy and security.

Wei Zhou et.al. in their paper, "Security/Privacy of Wearable Fitness Tracking IoT Devices," explore the potential vulnerabilities in wearable fitness trackers with a specific focus on Fitbit [1]. They point out critical vulnerabilities in the communication between the Fitbit device, base-station, and Web server, and address the vulnerabilities of a proposed solution, FitLock. The authors conclude with a call for increased attention to security and privacy in wearable fitness trackers. Our work builds upon this by providing countermeasures that users can employ to safeguard their data.

The paper "You Can Run, But Can You Hide? A Deep Dive into Fitness Tracker Privacy and Security" by Jacob Leon Kröger et.al., takes a closer look at the privacy and security implications of fitness tracking services, particularly focusing on Strava's Endpoint Privacy Zones (EPZs) [2]. The authors expose the limitations of the EPZ mechanism, demonstrating its vulnerability to an adversary looking to determine the existence of an EPZ and recover the protected address. Our work complements this by suggesting user-specific countermeasures and a real-time alert system that could protect such vulnerabilities.

Finally, Hossein Fereidooni, Tommaso Frassetto, Markus Miettinen, Ahmad-Reza Sadeghi, and Mauro Conti, in their paper, "Fitness Trackers: Fit for Health but Unfit for Security and Privacy," evaluate the security and privacy features of various fitness trackers [3]. Their findings highlight the lack of robust security measures, leading them to suggest a shift towards more privacy-focused development processes. This work resonates with our research, as we also underscore the need for manufacturers to prioritize privacy and security to safeguard user data.

Our study distinguishes itself from the above works by not only identifying vulnerabilities but also by suggesting user-centric privacy countermeasures. Furthermore, our project also focuses on the importance of implementing more complex clustering techniques and integrating AI for predictive analysis.

IV. Data Collection and Preparation

The primary subject of this study was the first author's activities, as recorded on Strava. Strava offers features such as segments, which break down routes into various sections for monitoring and comparison, and Flyby, which enables users to see others they encountered during their activities. However, a loophole in Strava's system inadvertently exposes users' activity start times in weekly activity summaries, contrary to the settings in Strava's privacy controls. This vulnerability, though seemingly trivial, poses potential privacy risks. The perspective from the Activity Page is depicted in Figure (1). It's observable that only the date of the activity is shown on this page.

Figure 1: Start Time not visible in Activity Summary Page

Contrastingly, Figure (2) displays the same activity within the context of the weekly summary. Here, both the start time of the activity and the date are visibly presented.

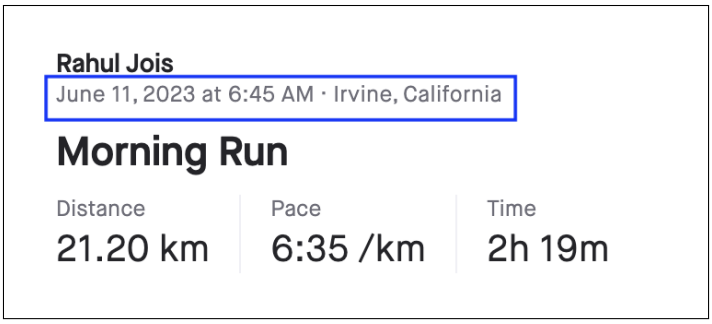

Figure 2: Start Time visible in the Weekly Activity Summary

In this study, we exploited this oversight, together with the comprehensive data Strava provides, to collect valuable information and construct an analytical framework. However, it's important to note that to mitigate privacy concerns, this project utilized only the first author's activities as the dataset. We outline our data collection and preparation process as follows:

1) **Data Gathering:** We extracted activity data (runs and rides) from the first author's Strava account using the Strava API. The data, comprising of three months' worth of previous activities, included geographical start and end points of activities, the route taken, and the activity's start time. We obtained the latter from the weekly activity summary due to a previously identified bug.
2) **Data Filtering:** The collected data was meticulously filtered. Activities with missing or inconsistent geographical coordinates were removed. We focused our

attention on activities that had complete and accurate geographical information.

3) **GPX Reconstruction:** Strava provides the lap times for each kilometer of an activity. By coupling this with the activity's start time, we were able to reconstruct the Global Positioning System Exchange Format (GPX) data by adding a timestamp to each coordinate. This facilitated our ability to analyze not just where the activity took place, but also the timing of each segment of the activity.

Through this process, we were able to assemble a comprehensive dataset containing detailed spatial and temporal information about the first author's activities. This provided a solid foundation for our ensuing analysis.

V. Methodology

To begin, our choice of DBSCAN (Density-Based Spatial Clustering of Applications with Noise) over traditional clustering algorithms such as K-means was informed by a number of considerations unique to our project. First, unlike K-means, DBSCAN does not require the number of clusters to be specified in advance, making it more adaptable to the variability of real-world location data. Secondly, it can discover clusters of arbitrary shapes, thereby avoiding the limitation of K-means which can only form spherical clusters. Furthermore, DBSCAN's density-based nature makes it more robust against noise and outliers, which are common in geolocation data. Finally, DBSCAN is particularly well-suited for geo-spatial data clustering, efficiently capturing areas of user activity, thus making it an ideal choice for our purpose.

A. Spatial Clustering

The preliminary step in our methodology was implementing spatial clustering on the activity data with the aim of grouping activities occurring in close geographical proximity. This approach allowed us to identify and analyze recurring patterns in the user's behavior. For the spatial clustering process, we employed the DBSCAN algorithm. This density-based clustering approach was ideal for our study as it efficiently groups together points densely packed in a certain area. Given the nature of location data, which is often cluttered with noise and outliers, DBSCAN proved to be quite useful. We used geographic coordinates, latitude, and longitude, as input features.

In DBSCAN, a key parameter is 'eps', the maximum distance between two samples for them to be considered as in the same neighborhood. We set an 'eps' value of 0.1 kilometers in radians, corresponding to a real-world distance of approximately 100 meters. This threshold was selected based on our understanding of how close activities should be to potentially relate to each other. Another parameter was the minimum number of samples required to form a cluster, which we set to 5. This meant a group must contain at least 5 activities to be recognized as a meaningful cluster.

The outcome of the spatial clustering process was a set of clusters each signifying a distinct geographic area of user activity. These clusters served as the initial categorization of activities based on their geographic proximity and laid the foundation for our subsequent stages of analysis.

The Python code for the above steps is as follows:

```
# Run first DBSCAN
coords = df[['Latitude', 'Longitude']].values
coords_rad = np.radians(coords)

# Set Distance Threshold
kms_per_radian = 6371.0088
epsilon = 0.1 / kms_per_radian

# Run DBSCAN on lat, long
db = DBSCAN(eps=epsilon, min_samples=5, algorithm=
    'ball_tree', metric='haversine').fit(
    coords_rad)
df['SpatialCluster'] = db.labels_


```

Listing 1: Spatial Clustering using DBSCAN

The spatial clustering provides a preliminary view of the user's activity distribution and allows us to start distinguishing unique patterns of their routines. Figure 3 illustrates the spatial clusters formed from the initial DBSCAN run on the user's Strava data. Each point signifies an activity's start or end point, with the color indicating the different clusters. However, for a more refined analysis, we needed to go a step further.

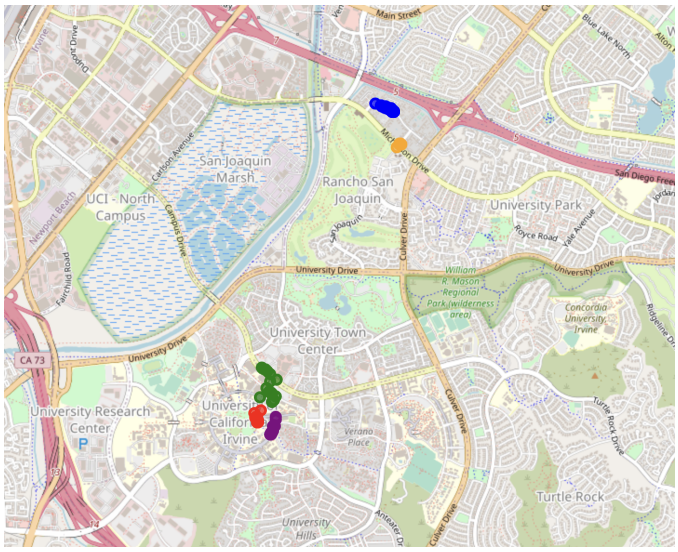

Figure 3: Spatial Clustering using DBSCAN

B. Temporal Clustering

Temporal clustering is an integral part of our methodology, aimed at creating a time-aware understanding of user activity. It's premised on the idea that activities conducted in the same location and time frame are likely related, thus aiding in unveiling patterns of user activity. We used the DBSCAN algorithm, as in the spatial clustering phase, but here the focus was on the 'HourOfDay' and 'PointType' features.

Given the different scales of hours in a day compared to geographic coordinates, we first standardized the 'HourOfDay'

feature using the StandardScaler from sklearn.preprocessing. This process ensured the 'HourOfDay' was on a scale that did not disproportionately impact the clustering process.

We used an epsilon value of 0.4 in DBSCAN for temporal clustering. In this context, epsilon dictates the maximum difference in 'HourOfDay' for two activities to be considered in the same cluster. With our 'HourOfDay' standardized, an epsilon of 0.4 equates to roughly a 3-hour difference in activity start times. Therefore, we grouped activities starting within approximately 3 hours of each other, enabling us to capture user behavior patterns that varied throughout the day.

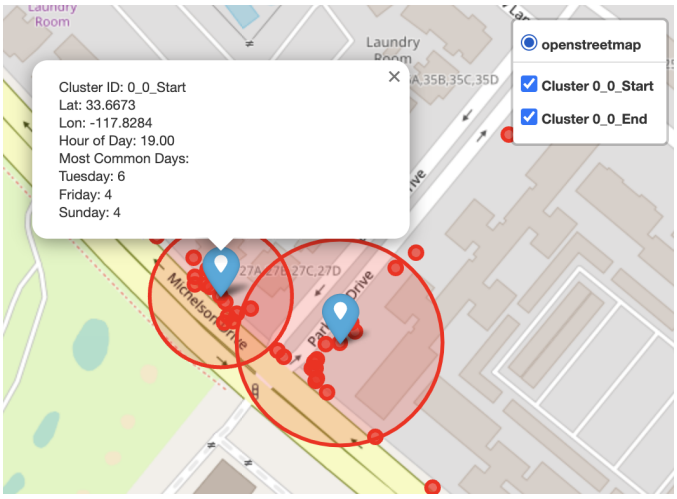

Figure 4: Temporal Clustering Output

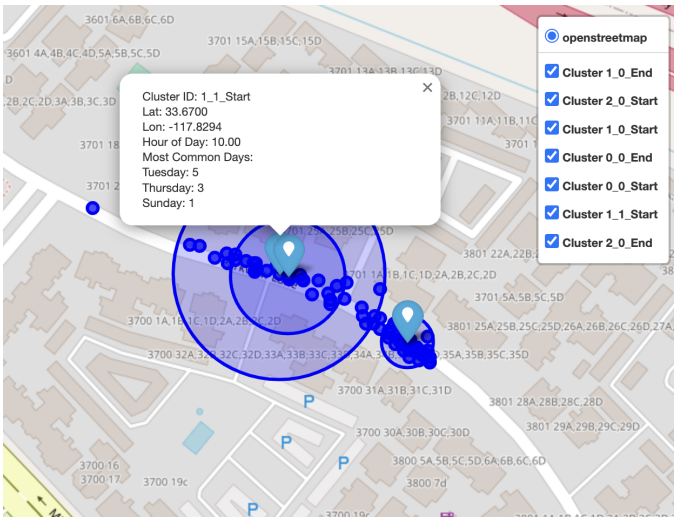

Figure 5: Temporal Clustering Output

The result of the temporal clustering process was a set of clusters that incorporated both spatial and temporal information. Each cluster represented a group of activities that were not only in the same geographic area, but also conducted around the same time of day, providing a nuanced picture of user habits. Figures 4 and 5 provide a visual representation of the temporal clusters. Each marker signifies a temporal cluster's center of mass, which represents the most frequent activity time and location. The popup attached to each marker offers more detailed insights, such as the cluster ID, the average latitude and longitude, the median activity start hour, and the most common days of activity in the cluster. This effectively combines spatial and temporal data to provide a richer understanding of user behavior.The temporal clusters thus set the stage for our heatmap-based segment generation, helping us identify detailed spatio-temporal patterns in the user's activity data.

The Python code for the above steps is as follows:

```
# Initialize a StandardScaler object
scaler = StandardScaler()
# Fit the scaler to the 'HourOfDay' column
df['HourOfDay_scaled'] = scaler.fit_transform(df[['HourOfDay']])
#Run DBSCAN within each spatial cluster using other features
df['Cluster'] = '-1'  # Initialize final cluster column
# Create separate clusters for 'Start' and 'End' points
for point_type in ['Start', 'End']:
    subset_df = df[df['PointType'] == point_type]
    for spatial_cluster in subset_df['SpatialCluster'].unique():
        # Skip spatial cluster -1 (noise)
        if spatial_cluster == -1:
            continue
        subset = subset_df[subset_df['SpatialCluster'] == spatial_cluster]
        if len(subset) >= 5:
            features = subset[['HourOfDay_scaled', 'PointFeature']].values
            dbscan = DBSCAN(eps=0.4, min_samples=5)
            clusters = dbscan.fit_predict(features)

```

Listing 2: Temporal Clustering

C. Heatmap and Segment Generation

After obtaining spatial and temporal clusters, our next step was to generate a visual heatmap representation for each of these clusters. The purpose of these heatmaps was to illustrate the density of the user's activities within each cluster, thereby highlighting the locations where the user frequented. To create these heatmaps, we used a pre-existing repository [4] specifically designed for this purpose. This repository employs Gaussian filters, a technique that assigns a weight to each point in the dataset based on its distance from the center. By applying this method, we transformed the raw GPS coordinates into a visually pleasing and smooth heatmap, where areas of greater activity were more intensely represented.

Once the heatmap was generated, we proceeded to further refine the output. The heatmap generation process results in two outputs - a visual image and a CSV file containing the coordinates and their corresponding intensity levels. We decided to filter this CSV data to retain only those points where the intensity was above 70% of the maximum observed within the cluster. This filtering step helped to emphasize the areas of highest activity within each cluster and discard less

frequently visited locations, allowing us to focus our analysis on the most active areas. Figure 6 illustrates the heatmap generated from the GPS coordinates within a specific temporal cluster. The varying color intensity indicates the frequency of the user's activities in different areas: darker areas represent a higher concentration of activities, thus indicating the user's most frequented routes or locations. This visualization aids in discerning patterns and trends in the user's spatial activity.

Figure 6: Heatmap

Following the heatmap creation, we shifted our attention to identifying 'segments' within each cluster. Segments represent distinct sub-regions of concentrated user activity within each cluster. Identifying these segments was an essential step towards understanding user's patterns and routines as they represent specific routes or locations frequently visited by the user. To do this, we combed through the GPS files in each 'start' cluster to find points present in the heatmap. Using these points, we created segments of approximately 500 meters each, a size that represents a meaningful distance in the context of outdoor activities. For each segment, we calculated the range of times it was visited and the frequency of visits, thereby gaining further insights into the user's activity patterns. Through this meticulous process of heatmap generation and segment identification, we could convert the raw GPS data into meaningful insights about the user's behavior and routines. We complemented our data processing and analysis steps with the generation of interactive maps. These maps displayed the spatial and temporal clusters, and segments identified through our process. The interactive nature of these maps allowed us to click on individual segments to reveal further details, enhancing our understanding of the user's activity patterns.

VI. Findings and Analysis

Following the application of spatial and temporal clustering, a distinct set of patterns emerged from the user's data. The spatial clustering successfully grouped together activities based on their geographic proximity. For example, Spatial clustering revealed five distinct clusters for the rides and surprisingly, only one cluster for the runs. This implies that the user consistently starts and ends his running activities at the same location.

In the subsequent temporal clustering phase, we detected more nuanced patterns in the user's activity schedule. The analysis showed that the user typically starts his activities on Monday, Wednesday, and Friday at around the 7th hour (approximately 7 AM) and on Tuesday and Thursday at around the 10th hour (approximately 10 AM). Interestingly, these activities typically take less than an hour to complete. The user seems to resume his activities at the 12th hour (around noon) on Monday, Wednesday, and Friday, and at the 13th hour (around 1 PM) on Tuesday and Thursday. Similar to the morning activities, these midday sessions also take less than an hour to complete.

The generation of heatmaps and segments further illuminated the user's behavior patterns. The heatmaps effectively visualized areas of highest activity within each cluster. For instance, by examining the created heatmaps, it became evident that the user consistently follows similar routes for both runs and rides within each cluster. This pattern reinforced the user's regular routine and preference for specific paths. The figure 7 illustrates an interactive map that's generated from our data analysis, presenting distinct segments of the user's activity path. Each segment is distinguished by a unique color, allowing for a clear differentiation between different parts of the user's route. When clicking on any point within these colored lines, a popup window appears, providing comprehensive details about the segment. These details include the Segment ID, the start and end coordinates of the segment, the time range within which the user tends to pass through this segment, and the frequency of the user's activities per day of the week on this specific segment. This interactive map thus allows for a user-friendly and in-depth exploration of the user's activity patterns and routines.

Similarly, the identified segments within each cluster represented specific routes or locations frequently visited by the user. Our analysis showed that the user passes through each 500m segment at approximately predictable times during their activities. This level of detail allowed us to not only understand where but also when the user prefers to carry out their fitness activities, providing a comprehensive insight into their routines.

Overall, our methodology successfully transformed raw GPS data into meaningful insights about the user's behavior and routines. These insights could potentially serve as a basis for future studies or applications focused on [provide possible applications of your results or ways they could be used in the future]."

VII. Privacy Countermeasures

In response to the privacy vulnerabilities identified in our study, we propose the following countermeasures which can be implemented at both the platform level (by Strava) and the user level.

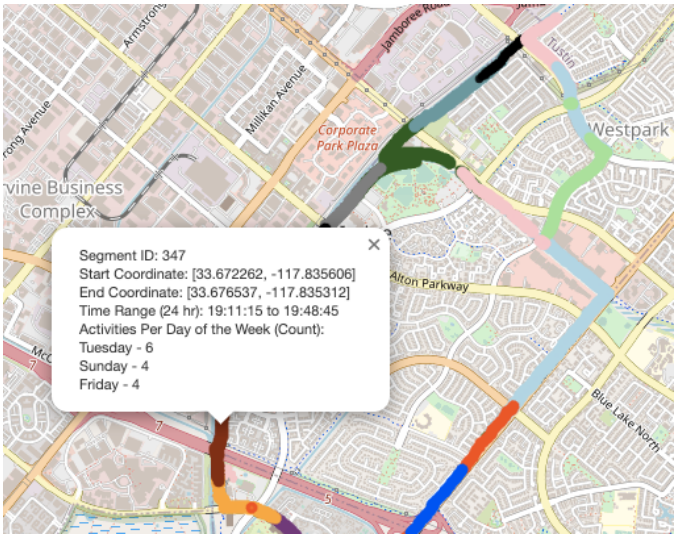

Figure 7: Interactive map showing segment information

A. Platform-Level Countermeasures

1) **Rectifying the Timestamp Bug:** The first step towards mitigating this vulnerability is by addressing the timestamp bug. Strava should fix the bug which displays the activity start time in the weekly summary, as it goes against their own privacy policy. This modification would prevent potential adversaries from being able to accurately infer the user's whereabouts based on activity start time.
2) **Adding Noise to GPS Data:** Another measure that can significantly improve privacy is the introduction of noise to the GPS data. This can be achieved by slightly altering the GPS coordinates of an activity. The alterations should be small enough not to noticeably affect the usefulness of the data (e.g., the total distance travelled), but large enough to prevent the identification of the precise location.
3) **User Awareness and Control:** Users should be made more aware of the potential privacy risks associated with sharing their activity data. Strava could implement a feature that visualizes the possible privacy implications, such as displaying a heat map to the user showing their most frequently visited locations.
4) **Differential Privacy:** As a more advanced solution, Strava could consider implementing a system of differential privacy. This involves adding statistical noise to data in a way that individual entries cannot be identified, but aggregate trends can still be analyzed. This would allow Strava to provide useful insights on user activity without compromising individual user's privacy.

B. User-Level Countermeasures

For users who want to proactively enhance their privacy while using fitness tracking applications like Strava, there are several actions that they can take:

1) **Set Profile to Private:** Restrict profile visibility to authorized users only. This measure prevents unauthorized access to the user's activities, thus adding a layer of privacy.
2) **Hide Start and End Points:** Users can adjust their privacy settings to conceal their activity's start and end points. This helps prevent potential adversaries from tracking the user's location.
3) **Mask Popular Addresses:** Frequent locations like home or office should be masked. Strava provides a feature to create privacy zones that do not display the precise start and end points within a certain radius of these locations.
4) **Disable Flyby Feature:** The Flyby feature on Strava allows others to see your activities if they happened to be nearby during the same time. Disabling this feature limits the visibility of your activities to others.
5) **Obscure Activity Maps:** Users can choose to hide the map of their activity, preventing others from seeing the routes they have taken.

It should be noted that while these countermeasures can significantly improve user privacy, they are not foolproof. Therefore, users should always be mindful of the information they are sharing and the potential implications thereof.

VIII. Future Work

While our study has demonstrated the potential privacy risks inherent in sharing GPS activity data on platforms like Strava, further research is needed to comprehensively evaluate and propose robust countermeasures. Here are a few potential directions for future work:

1) **More Extensive Data Analysis:** Our study was limited to a single user's data over a three-month period. Future work could consider larger datasets spanning multiple users and longer durations to identify common patterns and vulnerabilities.
2) **Other Fitness Tracking Applications:** We have focused our study on Strava due to its popularity and unique features. However, there are many other fitness tracking applications available that might have their own set of privacy implications. Future studies could consider a comparative analysis of privacy vulnerabilities across different fitness tracking platforms.
3) **Advanced Clustering Algorithms:** We utilized DBSCAN for spatial and temporal clustering in our study due to its suitability for our dataset. However, there are several other clustering algorithms that could potentially yield more accurate or insightful results. Future work could explore the use of these other clustering algorithms.
4) **Privacy-Preserving Data Sharing Mechanisms:** Future work could focus on the development and testing of privacy-preserving mechanisms for sharing activity data. This could involve methods to generalize or add noise to location data, or the application of advanced privacy techniques such as differential privacy.

The field of privacy-preserving data analysis is rapidly evolving, and we believe there are many fruitful avenues for future research. We hope that our study will inspire further work in this important area.

IX. Conclusion

In conclusion, our study showcased the ability to glean meaningful insights about a user's patterns and routines from raw GPS data obtained from Strava activities. By applying a two-stage DBSCAN clustering methodology—incorporating both spatial and temporal aspects—we succeeded in segmenting user activities based on their geographical proximity and the time of their occurrence. The patterns unveiled by these clusters provided a deep, multifaceted understanding of the user's activity trends.

Our heatmapping process, enriched by Gaussian filtering, provided a compelling visual representation of these clusters, illuminating areas of highest activity within each. This visual aid proved instrumental in our subsequent analysis, where we identified specific segments frequently visited by the user. This granular insight into the user's preferred routes and most active times could potentially be leveraged for tailored fitness recommendations, activity planning, or even risk assessment for personal safety.

Moreover, our exploration into privacy issues serves as a crucial reminder of the inherent risks associated with sharing location data, underlining the necessity of conscientious data-sharing practices. Strava users, and indeed users of any GPS-tracking fitness apps, must be aware of these privacy concerns and make effective use of available protective measures such as setting profiles to private, concealing activity start and end points, and obscuring activity maps.

Looking ahead, there's ample scope for enhancing this study. Future work could incorporate more variables such as elevation and weather conditions, apply different clustering techniques, or analyze a larger dataset comprising multiple users. Such extensions could provide an even richer understanding of user behavior, solidifying the foundation laid by this research.

This study stands as a testament to the valuable insights that can be extracted from location-based data, not just for the individual user, but also for wider applications in urban planning, health and fitness studies, and beyond. By continuing to develop and refine these data analysis techniques, we can uncover ever deeper layers of understanding from our increasingly data-driven world.

References

[1] W. Zhou and S. Piramuthu, "Security/privacy of wearable fitness tracking iot devices," in *2014 9th Iberian Conference on Information Systems and Technologies (CISTI)*, 2014, pp. 1–5.

[2] W. U. Hassan, S. Hussain, and A. Bates, "Analysis of privacy protections in fitness tracking social networks -or- you can run, but can you hide?" in *27th USENIX Security Symposium (USENIX Security 18)*. Baltimore, MD: USENIX Association, Aug. 2018, pp. 497–512. [Online]. Available: https://www.usenix.org/conference/usenixsecurity18/presentation/hassan

[3] H. Fereidooni, T. Frassetto, M. Miettinen, A.-R. Sadeghi, and M. Conti, "Fitness trackers: Fit for health but unfit for security and privacy," in *2017 IEEE/ACM International Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE)*, 2017, pp. 19–24.

[4] remisalmon, "Strava local heatmap," 2023, gitHub repository. [Online]. Available: https://github.com/remisalmon/Strava-local-heatmap