Strava Privacy Impact Assessment

Carl Elgin - r0644084

December 31, 2018

1 Application Description

1.1 Functionality

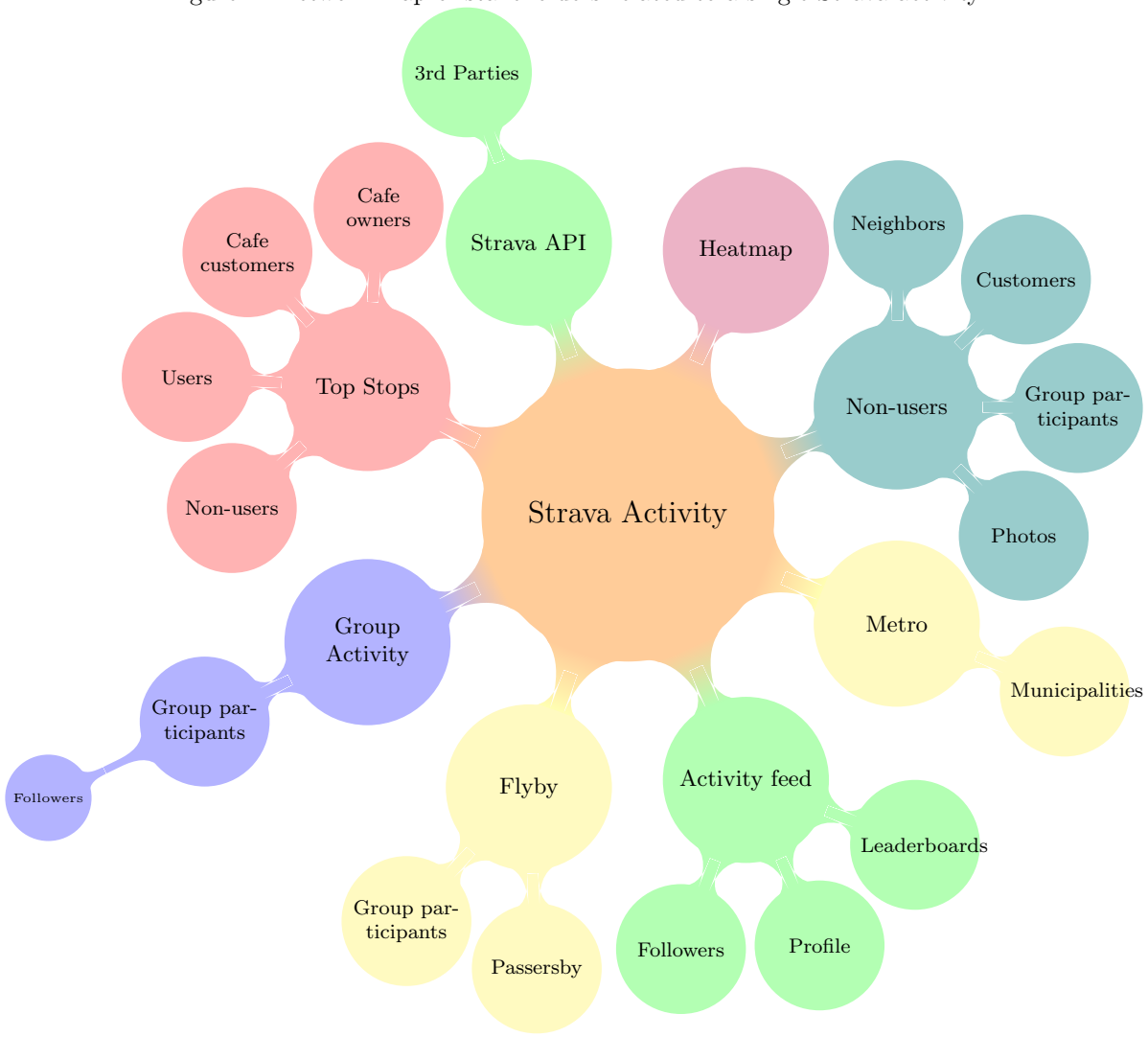
Strava is a mobile and web application that pioneers the intersection between training log and social network. Like other popular online fitness logs(TrainingPeaks, Today's Plan, Garmin Connect, etc) Strava is compatible with all major fitness trackers, in order to automatically sync data following activity completion. The automatic syncing of activities is compatible with devices that can export activities in .FIT, .TCX, or .GPX files, the most popular of which are cycling, running, and swimming[1]. However, in its current implementation, users are able to track activities in 25 different categories, in real-time, using the mobile application, or manually log activities in 33 different categories via the desktop version, after completing the activity.

The ability to automatically sync activities is one of the foremost advantages in using such a product. For the performance-conscious user, such as a professional athlete, it allows the athlete to review key performance metrics, such a pace, heart rate, or power, immediately following an activity. On the other hand, for users who are less interested in an immediate analysis, it still provides peace of mind, knowing that your activities will be tracked should you wish to tune-in to view monthly or yearly exercise summaries. Most importantly, the automatic sync is what sets Strava apart from most other social network platforms, in that once a sync connection is made between a user account and a activity tracking device, Strava passively collects data, indefinitely, until an intervention from the user.

In addition to the ability to provide a historical or granular summary of exercise, Strava also provides a social element that has been absent from traditional logging platforms(TrainingPeaks, Garmin Connect, etc). In similar fashion of most popular social networking platforms, Strava allows users to "follow" and "be followed" by other users, whereby users can see one another's activities, posts, and training logs, and engage through comments or giving "kudos"(a virtual "thumbs up"). Furthermore, in the case of cycling or running activities, Strava is capably of leveraging the GPS data to insert additional social features into activities via the *segments* functionality. *Segments* are community-defined portions of road, trail, or path over which Strava aggregates activity logs in order to declare the users with the fastest times "King of the Mountain"(KOM) and "Queen of the Mountain"(QOM). The segments feature, unique to Strava at its inception, provides a competitive layer to the social functionalities of the platform, which is a testament to the quality of information being collected. Users can view leaderboards for each segment, as well as their own historical performances, in addition to logged and generated metrics for each completion such as average speed, power, heart rate, etc.. This privacy impact assessment will focus on the following main features of the Strava platform.

- Training Log: Provides a historical summary of activities.
- Segments: Community-defined portions of road where users are ranked by performance.
- Flyby: Feature that displays activities that occurred simultaneously within geographical proximity[2].
- Global Heatmap: Visualization of all activities that occurred across the globe[3].
- Strava Metro: A local, enhanced, heatmap sold to municipalities for urban planning purposes[4].
- Top Stops: Project using machine learning to identify the 150 most popular stops during activities[5].
- Strava API: Gateway through which third party developers interact with the Strava platform[6].

Figure 1: Network map of stakeholders related to a single Strava activity



1.2 Stakeholders

Due to the passive data collection enabled by the automatic sync of activities, the optional enhancement of activities through classification, photos, and tags, and the extensive capability to link users and activities through group activities, Flybys, and segments, the scope of stakeholders is extensive. The network map in Figure 1 exemplifies some of the means in which stakeholders are propagated from a single Strava activity. As the most basic social feature of Strava, a user's activities will be published on the *activity feed* of all followers, as well as published on their personal profile page. This functionality is enabled by default to allow engagement with followers, via comments and kudos, and also allows users to publish posts, or share other user activities. Because of these features, many stakeholders rapidly propagate from a single activity.

As Strava uses geographic proximity and correlation measures in the *Flyby* and *Group Activity* features, it is common for anonymous individuals to become virtually linked and identified by the Strava platform. As historical activity tracking yields insights into behavioral patterns, as demonstrated by the *Strava Top Stops*, it is possible for other people to become stakeholders in a users privacy. For example, neighbors, group participants, cafe customers, or owners could be partially identified through being present in photos during activities or stops, or even sharing similar activity patterns. Furthermore, by sharing user data through *Strava Metro* and the *Strava API*, users of these products are also stakeholders in privacy of individual users.

1.3 What data is collected?

In terms of the desired functionality of activity logging, Strava is relatively transparent with respect to the type of data collected from activity trackers and wearables, as it is used to present and overview of each

activity. Activities are typically recorded one smartphones, wearables, or portable GPS units, and then pushed to the Strava infrastructure upon collection. In this regard, the number of recorded activity metrics is largely determined by the capabilities of the recording device. Due to the extensive logging and analysis features, users are incentivized to record and push as much data as possible to the Strava platform. By default, the Strava mobile application records time and GPS data for each activity, which is used to determine speed/pace, distance, segment times, etc. at the minimum.

In addition to what most users consciously push to the platform, Strava performs extensive data collection, as alluded to in the *Strava Privacy Policy*[7]. Although it may not be evident when using the platform, it is clear from the documentation that Strava will collect any information about activities or use of the platform in order to refine and improve its service offerings. As outlined in the *Strava Privacy Policy*, this also includes making inference about health information such as Suffer Score, Fitness and Freshness, and Training Load(among others), which Strava is able to refine through machine learning in large-scale[8].

As stated in the privacy policy, Strava will also collect data for the purpose of making inferences or products to be sold to other entities, as well as purely for marketing purposes. The primary example of the viability of the Strava data collection abilities, is the Heatmap and Metro features developed under the Strava Labs moniker. Additionally, Strava collects, shares, and aggregates data to and from any third-parties associated with a user, their followers, or any of their personal contacts, related to the platform or otherwise. With this information, Strava is able to partner with affiliate groups to directly target users through advertising campaigns, in-platform groups, activities, and challenges, as well as to engage potential users.

1.4 Design and Implementation

The implementation of Strava is not straightforward, and can be described as segmented into Strava and Strava Labs, as shown in Figure 2. As a whole, Strava represents the platform that users interact with most often, which includes the training log, activity feed, and social features. Furthermore, through the use of the Strava API, third parties allow users to interact with the Strava platform directly on activity trackers. One popular example of this is the *Strava Live Segment* feature, which allows GPS units to pull segment information during an activity and alert the user to upcoming segments, as well as provide real-time pacing information. Additionally, Strava withholds more detailed analysis of completed activities for premium membership called *Strava Summit*.

On the other hand, *Strava Labs* consists of the more computing-intensive applications that are not presented in the typical Strava user experience. *Strava Labs* projects represent the combination of advanced machine learning techniques and large stores of available data and include *Heatmaps*, *Flyby*, *Slide to GPS*, and *Top Stops*, among others. These projects are referred to as "proof of concept" as most are open source, and serve as the building blocks for their general service offerings. On the other hand, *Strava Metro* is a project in which Strava leverages their Heatmap functionality, in addition to extensive historical analysis, to provide a service catered towards municipalities for the purpose of urban planning.

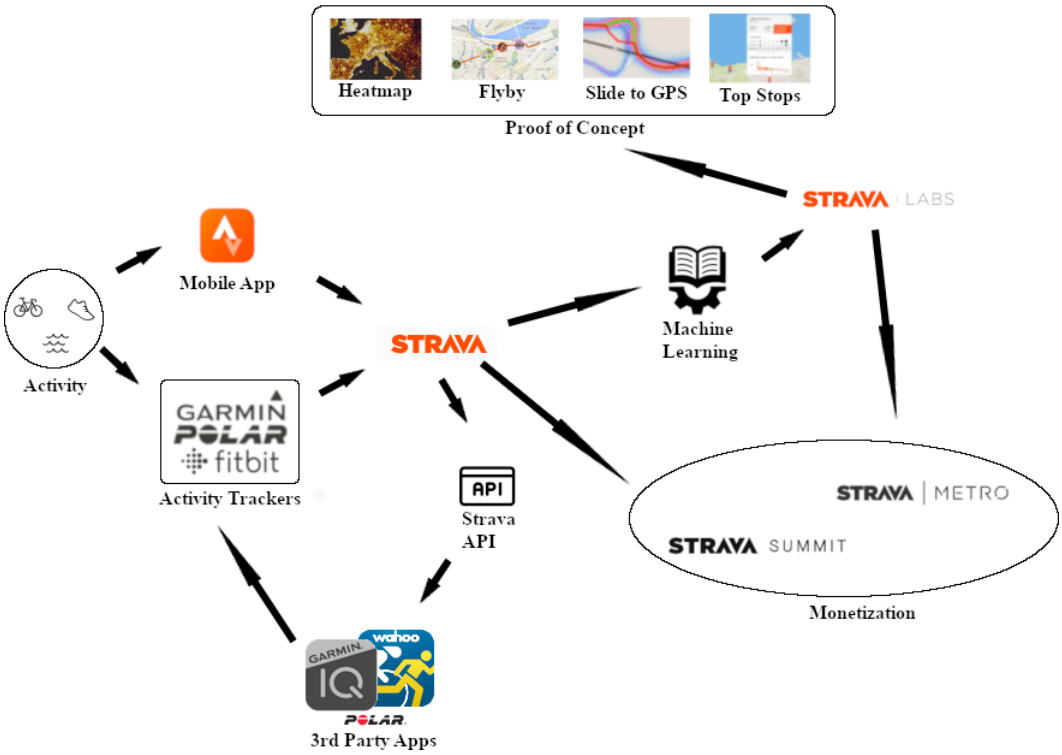


Figure 2: Overview of Strava implementation

2 Privacy Impact Assessment

2.1 Ethical

As the number of recorded activities and interaction with the Strava platform increases, so do the number of ethical concerns regarding the privacy impact of the stakeholders. However, both mobile and web platforms provide an encouraging atmosphere for users to increase the amount of activities they perform, markup their activities with photos, tags, and metadata, and subsequently increase the value of information being uploaded. For example, all user profiles and training logs display historical activity summaries, allowing users to view their activity patterns in daily, weekly, monthly, and yearly patterns. These types of summaries make Strava a highly valuable platform for users to invest in, as they allow for definitive fitness tracking, paired with social feedback elements.

Although users are presented with high-level activity analysis to incentize further activity sharing, the more extensive inferences that Strava can make from user data is not immediately visible. A primary example of this is the *Global Heatmap* project, developed under the *Strava Labs* moniker, which aggregates all of the GPS-tracked activities and overlays them onto a global map, to identify the most popular routes of its users, as shown in Figure 3. As a flagship project, this information directly benefits users through Strava's route building tool, which assists users in building GPS-based routes for future activities, using a local heatmap as a reference. This information also allowed *Strava Labs* to develop the *Slide to GPS* project, which uses heatmap data to make corrections to the Open Streetmap routing.

Figure 3: Strava's Global Heatmap identifies the most common activity routes in Leuven.

On the other hand, applications such as the *Global Heatmap*, also mean that, when overlayed on a map, patterns of activity can reveal much more information than users are aware of. First discovered by an Australian student Nathan Ruser, the *Global Heatmap* also makes it possible to identify secret locations, such as military bases and outposts[9]. In cases such as this, the methods that allow users to so clearly identify the best running routes also allow for clear identification of activities in "activity deserts". This type of exposure has been termed *fit-leaking*, and can be used to infer the *presence* of users, the rate of *activity*, the *profile* of the user, and even the *identity* of the user(PAPI)[10].

Furthermore, inferences about activities can be made by viewing the page for an activity itself, or viewing a replay in Strava's *Flyby* utility. Depending upon the specific device that tracked the the activity, each activity page will display a map of the activity and plotted elevation, distance, speed, time, etc.. Users can hover over the map or any plot to observe the data points for that specific point in time, including an locations the user stopped or visited during the tracked activity. Figure 4 shows a screenshot of the activity page for a bike ride in which I stopped for coffee at the Los Gatos Coffee Roasting Co. in Los Gatos, California. Despite the fact that Strava has removed the period of time in which my bicycle was parked stationary outside of the coffee shop, it is clear from the jump in the activity time-stamp that I spent approximately 29 minutes at the coffee shop during my activity. Additionally, the *Flyby* functionality provides a replay of an activity to be augmented with a replay of activities from users that Strava has identified in close geographic proximity to the activity. This means that viewing the *Flyby* of one cyclist's coffee stop could potentially identify other cyclists who have stopped at, or been geographically close to the coffee shop. Although the time and location of a single coffee stop may not be sensitive information, the ability to observe each data point in an activity can reveal a significant amount of information about the activity patterns of a given user.

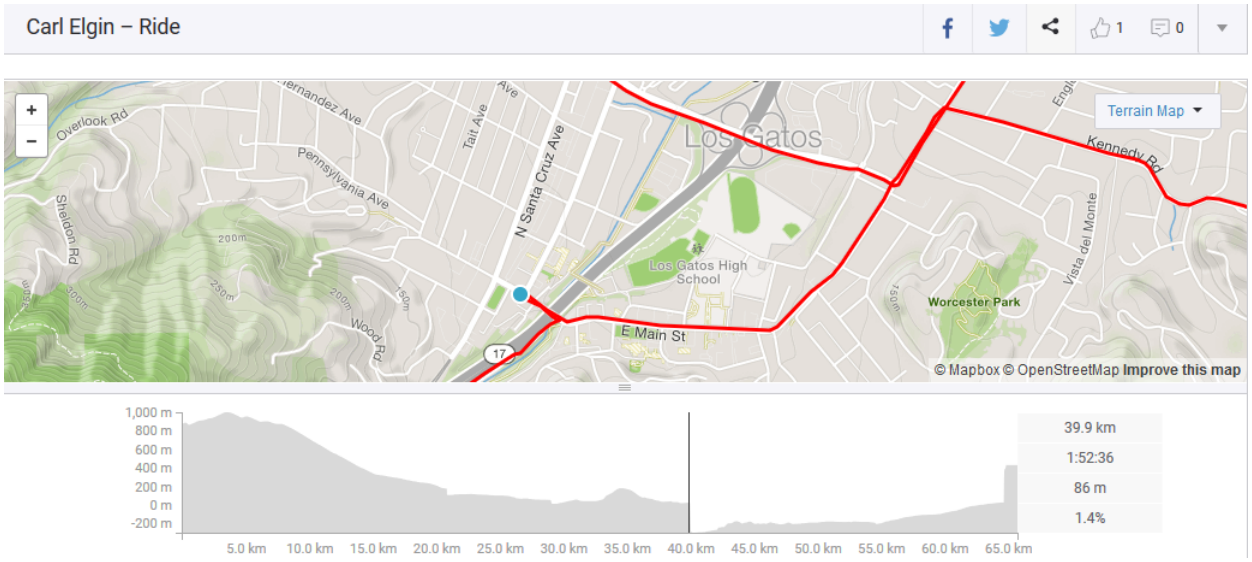


Figure 4: Location and duration of stops are clearly visible on the page for any activity. This ride includes an approximate 29 minute stop at the Los Gatos Coffee Roasting Co., as inferred from the jump in activity time.

Strava has recognized the capacity to identify user activity patterns, and has leveraged this type of inference into the *Strava Labs Top Stops* project[5]. Currently only available for the greater San Francisco Bay area, where Strava is headquartered, *Top Stops* uses machine learning to infer the 150 most common locations where its users stop, socialize, order coffee, or hang out[11]. Figure 5 shows a screenshot of the *Top Stops* profile for the Los Gatos Coffee Roasting Co., which includes the number of stops, average stop duration, and a distribution of the number of stops based on day and time. The *Top Stops* utility represents the massive amount of inference that can be made from having limitless access to global utility patterns simply through the use of time and GPS data.

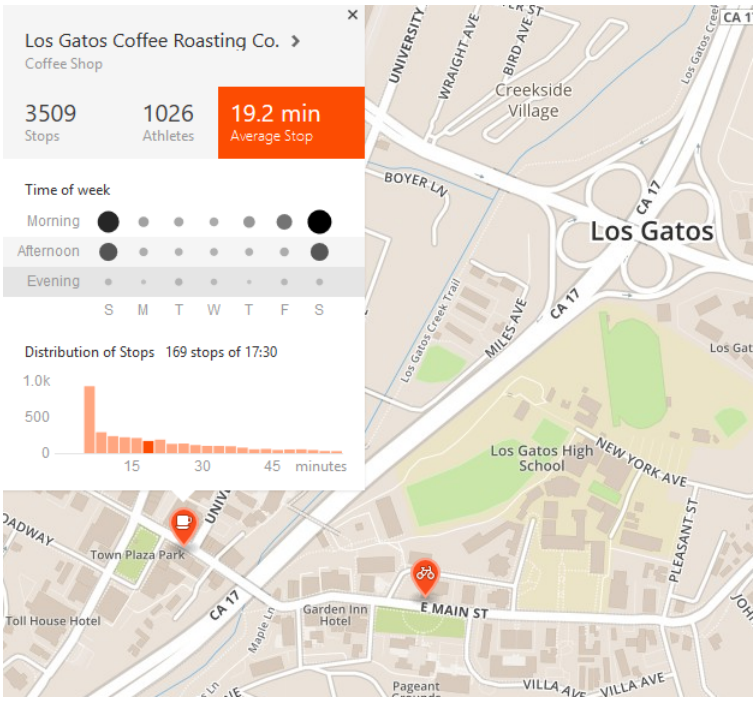


Figure 5: Top Stops are based on inference from historical activities

2.2 Technical

Due to their extensive service offerings, Strava places the burden of security on the user, emphasizing that "no system is 100% secure"[7]. The *privacy policy* maintains that apart from using a Secure Sockets Layer(SSL)

capable browser, the best method for ensuring that sensitive information is secured is through the privacy options in the user settings. Both mobile and web platforms offer privacy options that allow users to either contribute to or withhold information from various Strava services.

Within the account privacy settings, users can determine the specific audience with whom they would like to share their profile page, activities, group activities, and Flyby status. These options range from *Everyone*, *Followers*, *Only You*, and *No One* and are summarized in Table 1. The personal privacy controls also give users the choice of whether or not to contribute their activity data to the *Heatmap* and *Metro* projects. However, there is no opportunity to review the information that will be shared with these services, or grant permission on an activity-by-activity basis.

Table 1: Summary of the privacy control options within the Strava user account settings.

	Everyone	Followers	Only You	No One
Profile Page	✓	✓		
Activities	✓	✓	✓	
Group Activities	✓	✓		
Flyby	✓			✓

Additionally, users are allowed to set *Privacy Zones*, which are designed to hide a user's start and end of an activity within a specified perimeter of an address or coordinates. According to Strava, the zones are not centered at the location, ensuring that the perimeter itself does not identify the address[12]. However, due to the extreme inference capabilities regarding activity patterns of users, the privacy perimeter may not be enough to disguise a location in less populous areas, or when portions of the perimeter are inaccessible. Figure 6 shows a comparison of one of my own personal privacy zones with a public activity starting and ending in the center of the zone. In this example, as nearly half of the privacy zone consists of water, it is a less effective measure of protection as in other locations.

Figure 6: One of my personal privacy zones as per my user account settings(left) compared to a publicly viewed activity, starting and ending at the center of the zone(right).

2.3 Legal

As stated in the *Privacy Policy*, Strava maintains the right to use and distribute information that users provide to the platform, in accordance with their individual privacy settings. Additionally, Strava claims to abide by the General Data Protection Regulation(GDPR) stating that users have the *"right to access, rectify, download or erase your information, as well as the right to restrict and object to certain processing of your information"*[7]. Although the *Privacy Policy* clearly outlines how and when Strava can share user-submitted data with third parties, it is not clear whether Strava shares *inferences* about users, which are typically more informative and identifying.

In addition to the *Privacy Policy*, Strava outlines the legal guidelines for usage of the Strava API in the *Strava API agreement*[6]. The documents outlines the process through which developers can use the Strava API, which includes a thorough application and registration process. The agreement states that following registration, the API user assumes burden of legal compliance, including GDPR where applicable. This document also includes a limited liability statement directed towards API users and developers stating that Strava is not responsible for unlawful use of user data via the Strava API, as it relates to the *Privacy Policy* as well as GDPR regulations.

3 Recommendations

3.1 Ethical

Following the ethical concerns raised in the privacy impact assessment, there are several suggestions to be made in regards to user privacy. First and foremost, Strava has displayed their extreme inference capabilities through their *Strava Metro*, *Top Stops*, and *Flyby* projects, which reveal the extent to which troves of exercise activity data can reveal potentially identifiable patterns. As a result of these *Strava Labs* projects not taking a larger stage in the training log and social feed aspects of the platform, a large number of users may never be aware of their existence. Because of these inference capabilities, the Strava platform should make an effort to inform users of these capabilities, along with strategies for personal protection.

Additionally, to improve the extent to which users are aware of the scale and scope of data that is currently in use in the platform, Strava should allow users to generate a personal data usage report. Such a report should clearly identify a breakdown of the type, category, and amount of individual user activity data, an explanation of the current privacy control settings, and a clear demonstration of how the specific settings effect the privacy of the given data. Due to the wide array of service offerings and the extensive API functionality, the current Strava implementation results in complete user ignorance with respect to their data fingerprint.

Furthermore, as of 2013, Strava has implemented a personal heatmap, essentially a non-interactive *Global Heatmap*, representing all of a user's activities within a given year. Unfortunately, this feature is restricted for Strava Summit users, who pay a monthly fee for additional functionality. Despite it's current implementation as a premium feature, Strava should provide the personal heatmap to all free users in order to be more transparent about how much activity data is being generated, what could be inferred by the data, and what Strava is currently selling in the *Strava Metro* product.

3.2 Technical

In regards to the technical issues of the privacy impact assessment, there are additional changes that Strava could implement to mitigate future problems. In the privacy controls page, each user can configure new privacy zones and also regenerate existing zones. The zones are circular regions with a user-specified diameter of between 200 and 1000 meters. According to the Strava documentation, the privacy zones are not centered on the specified address, so as to decrease the potential for identification. Despite these measures, the privacy zone may still leave users vulnerable in less densely populated areas, or when multiple activities can be compared to identify the circular zone itself.

In order to mitigate this vulnerability, the privacy zones should regenerate automatically, on a frequent schedule or an activity-by-activity basis. Additionally, the privacy zone implementation should be altered in order to allow multiple privacy zones to have an effect on a single activity. For example, this would allow a user to set one privacy zone for their home and another for their work, which would protect both locations while using Strava to track their commute. Furthermore, Strava should recognize privacy zones when grouping activities together, such that during a group activity, no private locations of either user appear on either activity.

3.3 Legal

Strava is well versed in terms of their legal obligations regarding user data, as evident in the *Privacy Policy* and *Strava API Agreement*. However, no process currently exists for Strava and API users to communicate their compliance with these regulations to users. The *Strava API Agreement* states that registered API users are obligated to report any instances of security breach or non-compliance back to the Strava organization, but developers should also have an obligation to report directly to the individual users as well.

4 Conclusion

In conclusion, the Strava platform has displayed the extreme capabilities of activity pattern recognition by having access to data from smartphones and activity trackers. Primarily marketed as a training log, privacy is not built into the platform by design, as users are encouraged to build a data repository that can be shared will friends and family. Instead, the burden of privacy is placed on the individual user, who must tune their individual account privacy settings, without ever understanding the extent to which their data may leave them vulnerable. Despite current compliance with privacy regulations, the recommendations presented in this privacy impact assessment would increase the available protections for all users of the platform.

References

[1] *How to get your Activities to Strava*. https://support.strava.com/hc/en-us/articles/223297187-How-to-get-your-Activities-to-Strava.

[2] *Strava Flyby*. https://labs.strava.com/flyby/.

[3] *Heatmap Updates*. https://blog.strava.com/press/heatmap-updates/.

[4] *Strava Metro*. https://metro.strava.com/.

[5] *Strava's Top Stops*. https://labs.strava.com/top-stops/.

[6] *Strava API Agreement*. https://www.strava.com/legal/api.

[7] *Strava Privacy Policy*. https://www.strava.com/legal/privacy.

[8] *Understanding Strava: how to use Suffer Score, Fitness and Freshness, Weighted Average Power, Intensity, Training Load and Power Curve*. https://roadcyclinguk.com/how-to/understanding-strava-training-metrics.html.

[9] *Why Strava's Fitness Tracking Should Really Worry You*. https://www.forbes.com/sites/thomasbrewster/2018/01/29/strava-fitness-data-location-privacy-scare/#220575ff55c3.

[10] *Fit Leaking: When a fitbit blows your cover*. https://www.johnscottrailton.com/fit-leaking/.

[11] *The Bay Area's Top Stops*. https://medium.com/strava-engineering/the-bay-areas-top-stops-d9516b7a1009.

[12] *Privacy Zones*. https://support.strava.com/hc/en-us/articles/115000173384.