Appendix A
Summary of notations

Table 3
Summary of notations.

Notation	Description
AP	Access Point
LP	Location Proof
DP/EP	Distance/Elevation Proof
R	Communication range of APs
ΔT	Duration of silence periods
$E_{\text{OPE}}(\cdot)$	Order-preserving encryption scheme
GK_{pub}	The public group key. It is known by everyone.
GK_{priv}	The private group key. It is known only by (all) the APs.
GK_{OPE}	The group key used for order-preserving encryption. It is known only by (all) the APs.
$\text{LP}_{i,j}$	The j-th LP collected at sampling time t_i. It has format $\{P_i, t_{i,j}, (x_{i,j}, y_{i,j})\}$.
P_i	The pseudonym which is used at sampling time t_i.
$t_{i,j}$	The time at which the j-th LP is collected at sampling time t_i.
$(x_{i,j}, y_{i,j})$	The coordinate of the AP from which the j-th LP is collected at sampling time t_i.
C_i	The set of APs from which the LPs are collected at sampling time t_i.
A_i	The time-aligned intersection of the communication discs of the APs from which the LPs are collected at sampling time t_i.

Appendix B
Formalization of the sampling problem

In this appendix, we detail the graph construction for the maximum-weight path formulation of the sampling problem in different scenarios (with one or multiple operators, with or without silence periods). Unlike in the base case, with silence periods, we must distinguish between the samples where a user collects a location proof to *start* a distance proof and those where she collects a location proof to *end* a distance proof. The case of a single operator with silence periods is illustrated in Figure 11. We build a graph with two vertices per time sample that correspond to both situations ("start" and "end"). The "start" vertex corresponding to τ_i is connected to all the "end" vertices corresponding to τ_j ($j > i$), which means that a distance proof started at time τ_i can be ended at any time τ_j ($j > i$). The weight of such edges (depicted with solid lines in the figure) is the value of the distance proof obtained, i.e., $d(A_i, A_j)$. To start a new distance proof, the users must observe a silence period. Implementing silence periods means that a user cannot start a new distance proof before ΔT time units after she ended her last distance proof. In other words, assuming that the time between two time samples is 10 seconds and that $\Delta T = 15$, if a user ends a distance proof at τ_i, she can start a new distance proof at τ_{i+2} at the earliest. Therefore, we connect the "end" vertex corresponding to τ_i to all the "end" vertices corresponding to τ_j ($j > i+1$). The weight of such edges (depicted with dashed lines) is zero.

In the case of multiple operators, we build a graph with two vertices per sample ("start"/"end"), *for each operator*. A distance proof started with a location proof of a given operator can be ended only with a location proof from the same operator, therefore the "start" vertices are connected only to the "end" vertices of the *same* operator. The "end" vertices, however, are connected to the start vertices of all the operators as a user can start a new distance proof with any of the operators. The construction of these edges follows the same rationale as above, except that a user does not need a silence period when starting a distance proof with an operator different than that of the previous distance proof.

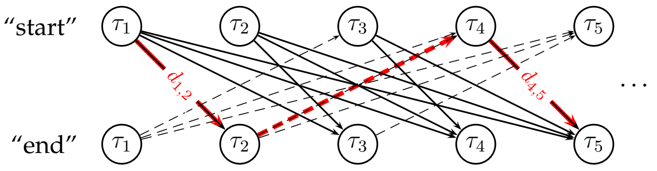


Figure 11. Graph construction for the maximum-weight path formulation of the sampling problem (one operator, with silence periods). Solid edges have weights equal to the corresponding distance proofs ($d_{i,j}$ is the short for $d(A_i, A_j)$). Dashed edges have zero weights. The thick, red path shows an example of a set of sampling points: The user starts a distance proof at time τ_1 (by collecting location proofs), which she ends at time τ_2 thus collecting a proof of value $d_{1,2}$; she then starts a new distance proof at time τ_4 (thus observing a silence period) which she ends at time τ_5. The total weight of this path is $d_{1,2} + d_{4,5}$.

Appendix C
Data-set details

In this appendix, we give more statistics about the datasets we used in the evaluation of SecureRun. Figure 12 and Figure 14 (top) depict, in the form of heatmaps, the densities of FON access points in Brussels, London and Paris and the densities of Free access points in Paris. Figure 14 (bottom) depicts the densities of Garmin activities.

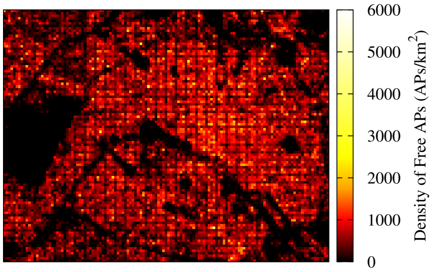


Figure 12. Heat-map of the density of Free access points in Paris.

Figure 13 shows the elevation map of Paris, where we evaluated the performance of SecureRun for elevation proofs.

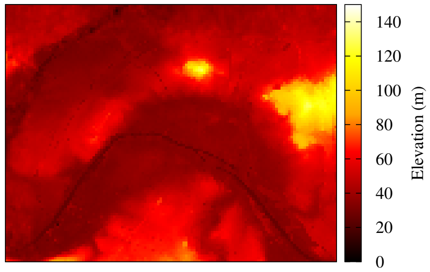
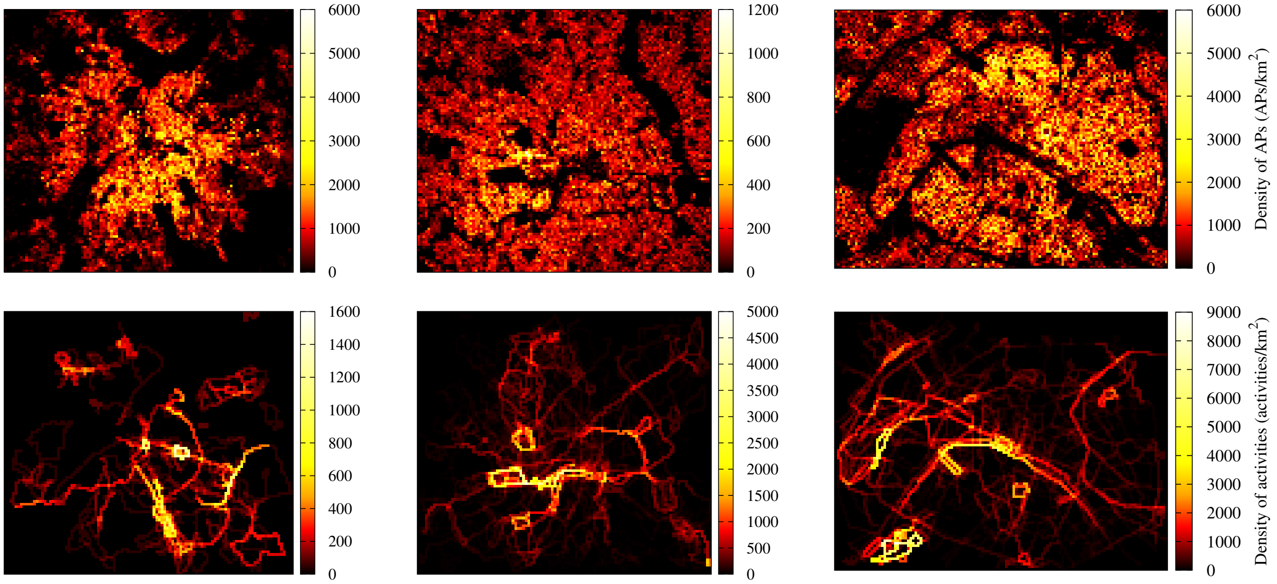


Figure 13. Heat-map of the elevation in Paris.

Figure 15 show the distributions (experimental CDFs) of the duration, length, elevation gain, density of APs along the



(a) Brussels (b) London (c) Paris

Figure 14. Heat-maps of the densities of FON access points (top) and of Garmin activities (bottom) in (a) Brussels, (b) London, and (c) Paris.

activity and the proportion of covered chunks (as defined in the data-set section) among the activities of the Garmin data-set (after filtering using the parameters from Table 1).

Table 4
Summary of the statistics of the filtered data-sets (FON and Garmin Connect) used in the evaluation (mean and standard deviation).

	Brussels	London	Paris
Number of AP	92,280	39,776	87,521
Number of activities	107	294	437
Density of AP (AP/km^2)	401±569	109±96.6	646±686
Density of AP along path (AP/km)	17.1±12.0	5.99±1.67	23.8±18.6
Proportion of covered chunks (%)	63.9±20.0	83.0±15.0	77.7±23.5

Table 4 gives some statistics on our (filtered) data sets of access points and activities (*i.e.*, FON and Garmin Connect). It can be observed that the density of access points is lower in London but they are more uniformly spread, especially along activities (as illustrated by the relatively small standard deviation compared to Brussels and Paris).

Appendix D
Investigation of the corner cases

In this appendix, we report on our manual inspection of the paths of the activities for which SecureRun provides low-accuracy summaries despite the high density of access points along the path and the high proportion of covered chunks. During our investigation, we found one typical case of such situations, which we show in Figure 16 and explain below. The general pattern is when a fraction of the path is very densely covered by access points but the rest is not; hence the average density over the whole path remains relatively high. More specifically, the user typically runs in a periodic fashion (e.g., around a stadium or back and forth on a street) on the part that is not well covered by access points, but this part is still covered by one or several APs (very close to each other) in a single location; hence the user cannot obtain any distance proofs (all the location proofs come from the same set of access points located close to each other) and almost all the chunks of the path are covered.

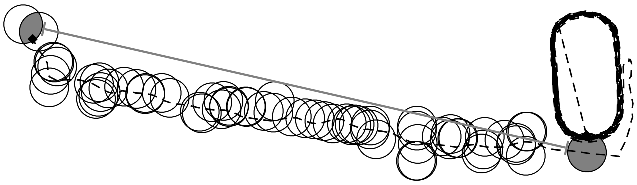


Figure 16. Example of an activity for which proportion of covered chunks is greater than 80% and the precision is smaller than 25% (planned sampling, ΔT = 60 s). The path is shown as a dashed line and the circles denote the communication ranges of the APs. The shaded areas represent the combined location proofs obtained at the sampling points.

It can be observed in the sample case that the user first runs to a stadium through a residential area and then runs a dozen of times inside the stadium on the 400-meter track. Because the stadium is covered by a single AP, all the chunks of the activity are covered, but it is not sufficient to increase the accuracy as all LPs are obtained from the same AP.

Appendix E
DP calculation with negative information

In Equation 1, we consider only the set C_i and do not take into account the fact that the user was *not* in the regions defined by the access points in the set $C \backslash C_i$ (hereafter, we call this *negative information*). Intuitively, if negative information is considered, the region that the user is inside at time t_i could be redefined as the intersection of A_i (as defined in Equation 1) and the complements of the regions defined by the APs in $C \backslash C_i$. Therefore, the negative information would provide a tighter estimate of the area the user was in at time t_i. This would provide better accuracy for the system, as the refined region is included in A_i, but it would also enable

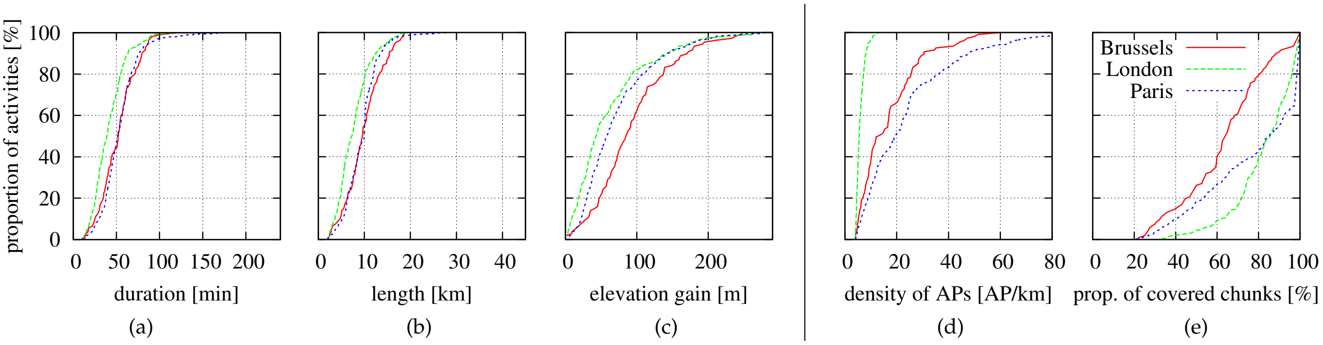


Figure 15. Experimental CDF of the (a) duration, (b) length, (c) elevation gain (d) density of FON AP (along the activity) and (e) proportion of chunks covered by FON APs, among the activities from the Garmin data-set.

the user to cheat by selectively reporting only a subset of the collected LPs (i.e., omitting some of the collected LPs to unduly increase the resulting distance proofs). For this reason, in our evaluation of SecureRun, we do not consider the negative information when calculating the lower-bound distance.

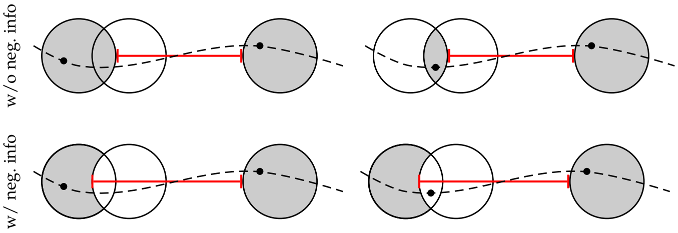


Figure 17. The lower-bound distance w/ and w/o considering negative information. The figures on the left show the case that using negative information improves the tightness of the lower-bound distance. The figures on the right show that using negative information can enable the users to unduly increase their lower-bound distance. The location samples of the user are shown with a dot.

These situations are illustrated in Figure 17: It can be observed on the left-most figures that, when the user is not in the intersection of the communication ranges of two or more APs, considering negative information (bottom) increases the accuracy of the distance proof compared to the base case (top). It can also be observed on the right-most figures that, when the user is in the intersection, by omitting to report one of the LP she collects (bottom), she can unduly obtain a larger distance proof (potentially higher than the actual distance) compared to the base case (top).

Appendix F
Survey details

In this appendix, we give more details about our online survey. First, we give the complete transcript of our survey questionnaire (Figures 19 and 20). Long lists of options have been truncated for the sake of conciseness. Our online questionnaire was designed with the LimeSurvey system and interfaced with the HealthGraph API in order to access the participants' RunKeeper account data. We used this data for screening purposes. Second, we give more detailed statistics about the survey participant's responses. In Figure 18, we show the repartition of the participants' concerns regarding the authenticity and the privacy implications of the activity data shared with activity-tracking applications.

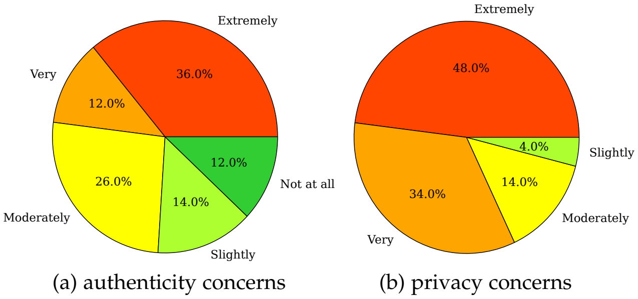


Figure 18. Survey participants' concerns regarding (a) the authenticity of the activity data shared by their friends and (b) the privacy implications of the activity data they share.

Demographics

1) What is your gender?

○ Male
○ Female

2) How old are you?

▭

3) What is your primary area of employment?

○ Retired
○ Unemployed
○ Student
○ Arts, entertainment, or recreation
○ Agriculture, forestry, fishery, or hunting
○ [. . .]
○ Transportation
○ Other: ▭

4) Besides RunKeeper, which of the following fitness applications are you a member of?

□ Strava
□ Runtastic
□ GarminConnect
□ Moves
□ Endomondo
□ MapMyFitness
□ Other: ▭

Fitness data sharing on online social networks

5) How often do you share your fitness activities with your friends?

○ Always
○ It depends
○ Never

6) When do you share your location-based fitness activities with your friends?
[shown only if the answer to the previous question is "It depends"]

□ When I take a new path
□ When I break a record
□ When I want to compete with myself or with my friends
□ Other: ▭

7) Applications such as digitalEPO.com and Fake Track enable users to claim a performance that they did not actually achieve.
Were you aware of this fact?

○ Yes
○ No

8) Knowing this fact, how important to you is the authenticity of the fitness activities your friends share?

○ Extremely
○ Very
○ Moderately
○ Slightly
○ Not at all

9) Sensitive information can be inferred from the data you upload on RunKeeper (e.g., home/work locations, medical conditions). Moreover, it has been shown that some popular fitness applications pass personal details about their users to insurance companies, e.g., to set premiums (Click here for more details*)
Were you aware of this fact?

○ Yes
○ No

10) Knowing this fact, how important to you are the privacy implications of the data you upload on RunKeeper?

○ Extremely
○ Very
○ Moderately
○ Slightly
○ Not at all

*http://www.dailymail.co.uk/news/article-2409486/

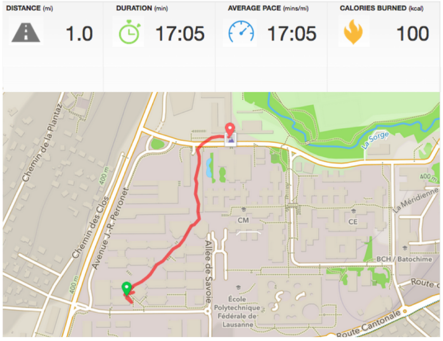
Figure 19. Transcript of our survey questionnaire (1/2).

Secure and private activity summaries

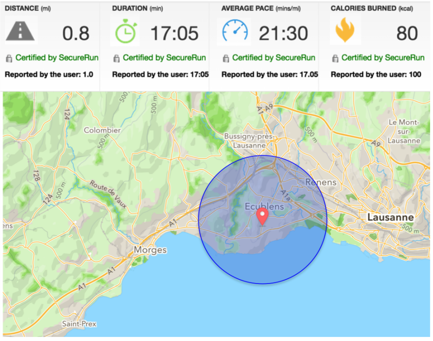
11) We designed a system, named SecureRun, that provides you with two main features:

- It protects your privacy. Specifically, the GPS traces of your activities is known only to you. You share only the summaries of your performance (e.g., the covered distance), and the coarse-grained information about the region where you perform your activities.
- It guarantees the authenticity of a fraction of the performance you report. For example, if you run 10 miles and report it, SecureRun can certify (based on cryptographic techniques) that you indeed ran at least 8 miles out of the 10 miles you ran.

We illustrate the differences between RunKeeper and SecureRun in the images below.



DISTANCE (mi)	DURATION (min)	AVERAGE PACE (mins/mi)	CALORIES BURNED (kcal)
1.0	17:05	17:05	100



DISTANCE (mi)	DURATION (min)	AVERAGE PACE (mins/mi)	CALORIES BURNED (kcal)
0.8	17:05	21:30	80
Certified by SecureRun	Certified by SecureRun	Certified by SecureRun	Certified by SecureRun
Reported by the user: 1.0	Reported by the user: 17:05	Reported by the user: 17.05	Reported by the user: 100

Assuming that you have run 10 miles, please choose your levels of satisfaction for the different values for which you would receive certification from SecureRun.

	Very low	Low	Medium	High	Very high
5 miles	○	○	○	○	○
6 miles	○	○	○	○	○
7 miles	○	○	○	○	○
8 miles	○	○	○	○	○
9 miles	○	○	○	○	○

Figure 20. Transcript of our survey questionnaire (2/2).