FLAIR: Storing unbounded data streams on mobile devices to unlock user privacy at the edge

Olivier Ruas [1] Adrien Luxey [2] Rémy Raes [2] Romain Rouvoy [2]

[1]Pathway [2]Inria Spirals - CRIStAL

Introduction

- **Geolocation data is sensitive:** from those, one can infer **points of interest (POIs)** such as place of living, work, political or shopping habits;
- **Location privacy protection mechanisms (LPPMs)** exist to transform traces, so they can still be used by third-party service providers *without* leaking personal data;
- The best possible usage is to gather data needed to run LPPMs **directly on the device**, to perform the data protection steps *in-situ*;
  - Mobile devices are however very constrained from a resources point of view, having limited CPU, memory, storage *etc.* capacities.

FLAIR

FLAIR (Fast LineAr InteRpolation) is a new piecewise linear approximation technique used to model and store temporal data streams.

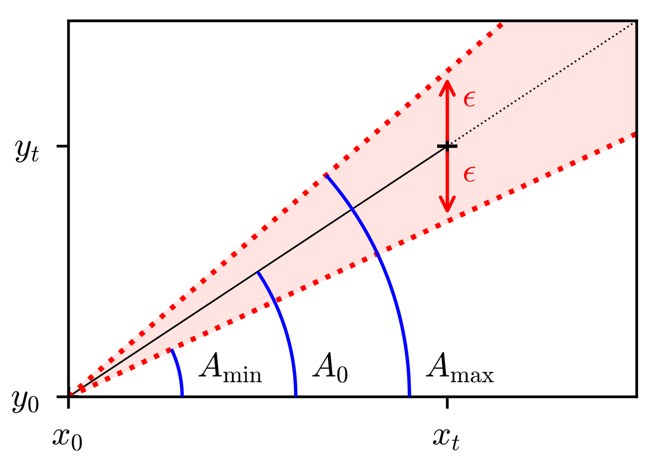

Figure 1. Tolerated error ϵ, persisted model (x_0, y_0, A_0) and in-memory $(A_{\min}, A_{\max})$ cone.

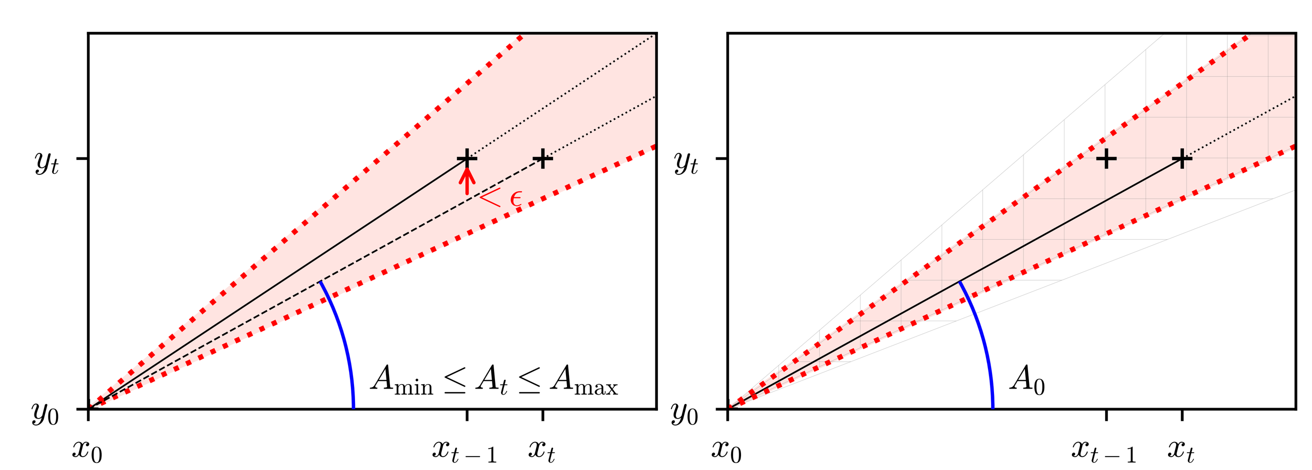

Figure 2. When new point fits the model, said model is updated.

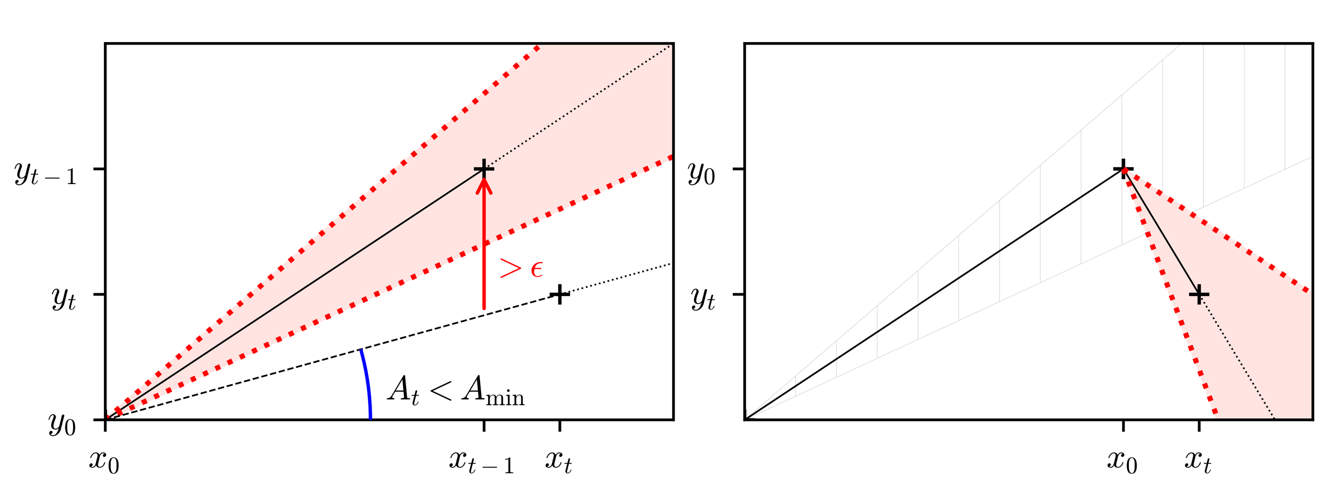

Figure 3. If new point does not fit, current model is saved, and a new one is created.

To check if a new inserted entry fits the model, *FLAIR* controls if it is in the cone modeled around latest entry.

- **if yes**, the model is updated: the in-memory cone becomes the intersection between previous cone and error cone around new entry (see figure 2);
- **if no**, the model is saved, and a new error cone is devised from latest entry and newly inserted point (see figure 3).

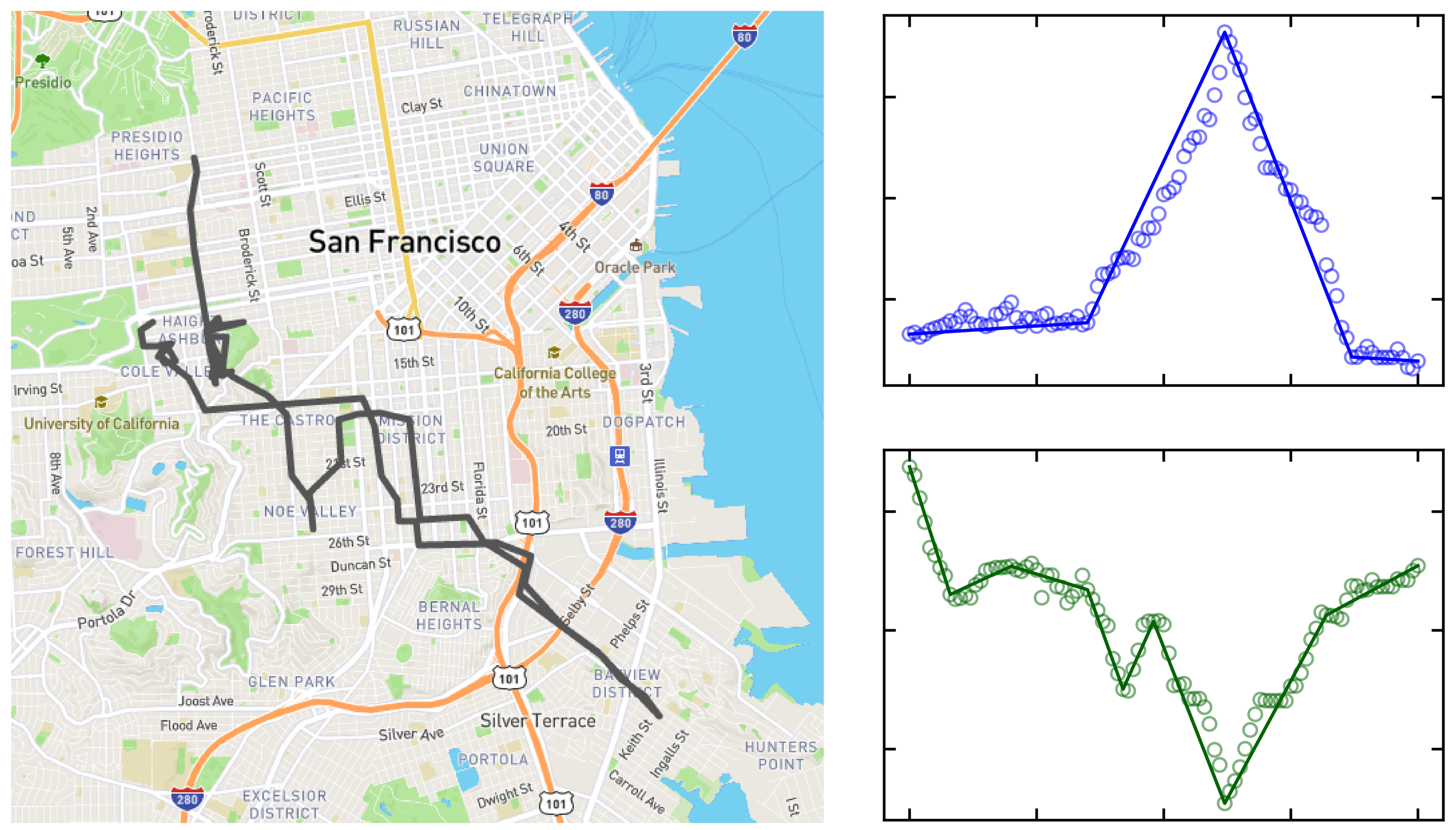

Figure 4. *FLAIR* modeling of latitude and longitude of a part of `user 0`'s trace from Cabspotting.

FLAIR performances

The main objective of *FLAIR* is to allow the storage of big temporal datasets on reduced memory. We check its storage capabilities by modeling two mobility datasets, PrivaMov [3] and Cabspotting [4]:

- PrivaMov (5GB) is modeled with 25MB (with ϵ = 0.001) → 99.95% gain
- Cabspotting size gain equals 21.02% (with ϵ = 0.001), due to traces lower density.

Moreover, checking the data throughput of *FLAIR* on random values, it is:

- **3505 times faster** on sequential writes
- **2343 times faster** on random reads

than SWAB[2] and GreyCat [1] competitors.

Protecting users' privacy

Promesse

Promesse [5] is a location privacy protection mechanism that removes POIs from traces by smoothing user speed along said traces. We implement it on mobile devices to provide protection to users; to assert Promesse's privacy protection, we simulate attacks on users' phones directly.

Divide & Stay

POI-attack is the state-of-the-art algorithm to infer POIs from location data; it is however slow (1h run time on desktop computer), and even worse on mobile devices.

To counter that, we propose a new algorithm, dubbed *Divide & Stay* (D&S), to compute POIs faster.

Platform	POI-attack	D&S	Speed-up
Desktop	59 min 20 s	32 s	×111
iOS	1 h 00 min 01 s	22 s	×164
Android	1 h 58 min 04 s	59 s	×120

Table 1. Computation times of raw POIs for PrivaMov's `user 1` on different platforms. *D&S* is at least 100 times faster than state-of-the-art approaches.

This shows that *Divide & Stay* enables infering points of interest from geolocation traces directly on mobile phones, which is not possible with the classical *POI-attack* approach, the latter taking too much time to execute.

D&S allows us to compute POIs fast, but are these as accurate as *POI-attack*-computed POIs? We compare in the table below the quality of all computed POIs: with/without *FLAIR* modeling, with/without *Promesse* protection, and using either *POI-attack* or *D&S*.

	without Promesse		with Promesse	
Algorithm	Raw POIs	FLAIR	Raw POIs	FLAIR
POI-attack	30	31	0	0
D&S	30	30	0	0
POI-attack ∩ D&S	21	20	-	-

Table 2. Impact of *FLAIR* and *D&S* on the number of inferred POIs from `user 0`'s trace in Cabspotting. Thanks to *FLAIR* and *D&S*, Promesse succeeds to protect user privacy at the edge.

More than 60% of *D&S*-computed POIs are the same as *POI-attack*-computed POIs, and 90% of the POIs are at a distance lower than 22 meters than a "real" one:

Using *FLAIR* to model data does not alter their utility.

References

[1] Thomas Hartmann, François Fouquet, Assaad Moawad, Romain Rouvoy, and Yves Le Traon. Greycat: Efficient what-if analytics for data in motion at scale. *Information Systems*, 83:101–117, 2019.

[2] Eamonn Keogh, Selina Chu, David Hart, and Michael Pazzani. An online algorithm for segmenting time series. In *Proceedings 2001 IEEE international conference on data mining*, pages 289–296. IEEE, 2001.

[3] Sonia Ben Mokhtar, Antoine Boutet, Louafi Bouzouina, Patrick Bonnel, Olivier Brette, Lionel Brunie, Mathieu Cunche, Stephane D'Alu, Vincent Primault, Patrice Raveneau, et al. Priva'mov: Analysing human mobility through multi-sensor datasets. In *NetMob 2017*, 2017.

[4] Michal Piorkowski, Natasa Sarafijanovic-Djukic, and Matthias Grossglauser. Crawdad data set epfl/mobility (v. 2009-02-24), 2009.

[5] Vincent Primault, Sonia Ben Mokhtar, Cédric Lauradoux, and Lionel Brunie. Time distortion anonymization for the publication of mobility data with high utility. In *2015 IEEE Trustcom/BigDataSE/ISPA*, volume 1, pages 539–546. IEEE, 2015.