K-Area: An Efficient Approach to Approximate the Spatial Boundaries of Mobility Data with k-Anonymity

Maël Gassmann
Bern University of Applied Sciences
Biel/Bienne, Switzerland
email: mael.gassmann@bfh.ch

Annett Laube
Bern University of Applied Sciences
Biel/Bienne, Switzerland
email: annett.laube@bfh.ch

Dominic Baumann
Bern University of Applied Sciences
Biel/Bienne, Switzerland
email: dominic.baumann@bfh.ch

***Abstract*—Mobility datasets, being by nature potent in utility and complexity, are hard to work with when privacy has to be preserved. Existing solutions to balance utility and privacy are very specific to certain use case or dataset types, and usually strive to provide an absolute privacy while disregarding computational efficiency. K-area is an efficient method that uses geometric operations to calculate the boundaries of movement profiles that guarantee a certain degree of anonymity and exclude areas where privacy is at risk. It is applicable to most types of mobility datasets, as it only requires a set of Global Positioning System (GPS) points tagged to an identifier. K-area provides the largest areas of the dataset, which all validate a geometric k-anonymity condition. By already providing a level of indistinguishability, these areas are the perfect starting point for many applications.**

***Keywords*-Mobility Data; Privacy; Indistinguishability**

I. Introduction

Mobility data is very complex and can reveal sensitive information about its Data Collectors (DC). Anonymization is therefore a must, and many different methods were developed. Most of them are quite expensive and need to be customized to the intended application [1]. However, a fast and flexible mechanism is often needed to assess the anonymity of a dataset and exclude the parts where anonymity cannot be guaranteed.

Once the k-areas are calculated, they can be applied on the raw dataset to only consider data points inside their bounds.

Many use cases are envisioned, this is just a small list:

- Set a k-anonymity condition and run the algorithm periodically while collecting data.
- Generate heat-maps to visualize the readiness of a dataset and where data could be lacking.
- To be used as a pre-processing step before running computational expensive algorithms.

A mobility dataset is finite, hence, it always has clear spatio-temporal bounds; the first and last records define the temporal bounds, while the minimal spacial bounds is represented by a shape, which contains all the GPS points.

The k-area algorithm understands this and, by being aware to whom each point belongs, will strive to – roughly, but effectively – further reduce the size of the spacial bounds by cutting out distinguishable data points.

Depending on how the shape of the area is generated (e.g., a convex area), all distinguishable GPS points might not find themselves outside the bounding shape. But even if the k-area algorithm might not necessarily remove all distinguishable points, due to the nature of the shape, most outliers will be at the edge of its bounds. Thus, ensuring that the biggest portion of outlying points is cut out, and therefore greatly improve the computation efficiency of hypothetical further analyses.

The paper is structured as follows: Section II describes pertinent related works; Section III defines the concept of k-areas; finally, the work is concluded in Section IV.

II. State of the Art

There are already privacy enhancing methods that all have various effects, and all focus on anonymizing a specific aspect of data through different means [2]. Such privacy enhancing methods are described below.

a) Mitigation: Such methods are trying to mitigate the privacy risks with heuristics without theoretical or provable guarantees. Examples are swapping, obfuscation, spacial cloaking or segmentation.

b) Indistinguishability: Here, anonymity is measured in terms of how distinguishable is every DC inside the dataset. From the metric, one can reduce the risk of breaching the privacy of DCs by filtering out singularities [3]. The k-area is part of this set of methods.

c) Uninformativeness: Predominantly measured through differential privacy, uninformativeness is providing privacy warranties by assessing how much information each individual data buyer posses.

III. Concept

If a convex hull can represent the spacial bound of an entire dataset, it can also represent subsets of it. The heart of the concept of k-area is to calculate the spacial bounds of the GPS points of each data collector, and from their superposition, to extract the areas that at least k bounds intersect.

A. Mobility Dataset Definition

A mobility dataset can be structured in many different ways. One common point between all such structures is that they will contain GPS points, and each will be tagged to a data collector identifier. Only finite datasets are considered.

a) Data Collector: A data collector u has a subset of GPS points of the dataset that are all tagged with the same DC identifier.

b) GPS Point: A GPS point can have many attributes. The only ones pertinent to the algorithm are its latitude, longitude and DC identifier.

Courtesy of IARIA Board and IARIA Press. Original source: ThinkMind Digital Library https://www.thinkmind.org

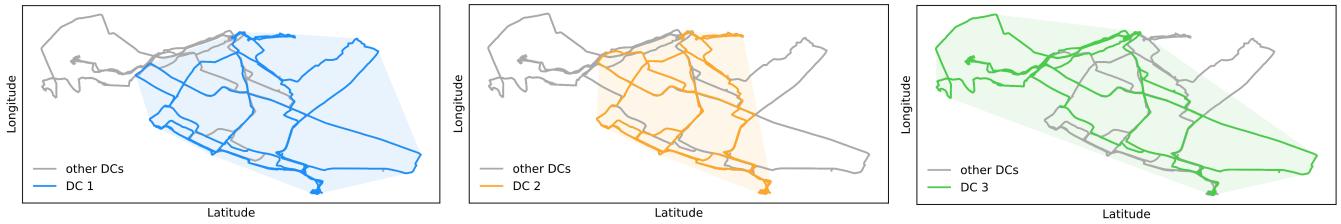

Figure 1. GPS traces of multiple DCs with their convex bounding polygons.

B. Polygon

The root of the concept is that the recorded GPS points of a DC u can be enclosed in a minimal polygon p_u. A minimal polygon can be, for instance, the smallest convex shape that contains all the DC GPS points (see examples for 3 DCs in Figure 1). It is not believed that there is an optimal minimal polygon type that fits perfectly each DC subset, a concave polygon or other types of shape might fit more complex cases, potentially at the cost of a less efficient algorithm. Further researches will focus on this subject. The inside of the intersection between two minimal polygons is following a k-anonymity of 2. To calculate the areas that follow a k-anonymity of 3, a third minimal polygon has to be intersected with the other two. Thus, if the algorithm was requested to yield the largest valid surfaces for these three DCs with a k-anonymity condition of 3, it will return their intersecting polygons.

C. K-Area

A k-area A_k is computed geometrically from all polygons p_u of P present in the data set (see illustrations in Figure 2). A k-area A_k is the union of the intersections between k polygons:

$$A_k = \bigcup (p_{i1} \cap p_{i2} \cap \cdots \cap p_{ik})$$
$$\text{for all } p_{i1}, p_{i2}, \ldots, p_{ik} \in P$$
$$\text{with } i_1 \neq i_2 \neq \ldots \neq i_k \text{ and } k \geq 2$$

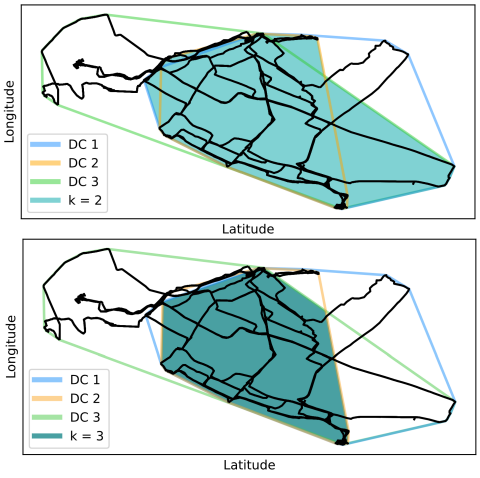

Figure 2. Three bounding polygons and the k2 and k3 areas.

D. Algorithm

To obtain an algorithm with polynomial runtime, polygon sets are used.

function K-AREA(P, k)
  $A \leftarrow$ array of size k with each entry as an empty set of polygons
  for all $p_u \in P$ **do**
    for $i \leftarrow k$ to 1 **do**
      if $i > 1$ **then**
        $A[i] \leftarrow A[i] \cup (p_u \cap A[i-1])$
      else
        $A[i] \leftarrow A[i] \cup p_u$ ▷ $i = 1$
      end if
    end for
  end for
  return A
end function

The function was successfully implemented in PostgreSQL with PostGIS extension using spacial data types.

IV. Conclusion

K-areas allow an approximation of the spatial boundaries of mobility data with a certain degree of indistinguishability. Based on geometric operations with polynomial order that can be implemented efficiently, the algorithm can significantly reduce the bounds of a dataset while ensuring relative k-anonymity. Further researches will focus on improving the results by using different types of shapes generated from GPS points while keeping the time complexity as low as possible.

V. Acknowledgment

K-area was developed in the innovation project “101.272 IP-SBM: Posmo Ethical Data Market” supported by Innosuisse.

References

[1] A. Kapp and H. Mihaljevic, “Reconsidering utility: unveiling the limitations of synthetic mobility data generation algorithms in real-life scenarios,” in *Proceedings of the 31st ACM International Conference on Advances in Geographic Information Systems*, ser. SIGSPATIAL ’23. New York, NY, USA: Association for Computing Machinery, 2023, pp. 1—12, [retrieved: 05, 2024]. [Online]. Available: https://doi.org/10.1145/3589132.3625661

[2] M. Fiore *et al.*, “Privacy of trajectory micro-data : a survey,” *CoRR*, vol. abs/1903.12211, 2019, [retrieved: 05, 2024]. [Online]. Available: http://arxiv.org/abs/1903.12211

[3] L. Sweeney, “k-anonymity: a model for protecting privacy,” *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.*, vol. 10, no. 5, pp. 557—-570, oct 2002, [retrieved: 05, 2024]. [Online]. Available: https://doi.org/10.1142/S0218488502001648

Courtesy of IARIA Board and IARIA Press. Original source: ThinkMind Digital Library https://www.thinkmind.org