Privacy Aware Analysis of Spatial Social Media Data

by **Jack Stephan**

Spatial social media data is generated in massive quantities for the first time in history. Taking advantage of this vast new source of data is advantageous for researchers analyzing spatial social media, but the potential privacy conflicts of users losing control of their personal geographic information is concerning. In order to continue to benefit from spatial social media analysis while protecting the privacy of those individuals who are contributing the data in the first place, a framework for privacy aware analysis of spatial social media data was investigated. This technique leverages the HyperLogLog algorithm in order to create a privacy aware data structure for mapping.

HyperLogLog

HyperLogLog (HLL) is a type of probabilistic algorithm that estimates cardinality [2]. Cardinality is the measurement of unique items in a set. This is particular useful for mapping social media data where often times unique counts of individuals, reactions, or days are the main metrics being mapped.

Yahoo Flickr Creative Commons 100 Million Photos

The Yahoo Flickr Creative Commons 100 Million data set (YFCC100M) is a collection of 100 million Flickr posts presented in a uniform database structure [4]. The data set contains photos that were uploaded in a 10-year period between 2004 and 2014. This the data set used for analysis during this research.

Privacy Aware Data Structure

This YFCC100M data set was processed into a standard location based social network data schema using the schema proposed by Dunkel, Löchner, & Krumpe [3] The Data was processed using the HLL algorithm to estimate cardinalities i.e. unique users, unique posts, and userdays at particular coordinate.

HLL Operations

With the HLL data structure it is possible perform a number of operations to transform or analyze the data

- Cardinality
- Union & Aggregation
- Intersection
- Filtering by Topic
- Combination of Techniques

Mapping the Data

Once processed, this data can than be mapped. Figure 1 displays some of these techniques combined. The data has been filtered by reactions relating to the beach, the points have been aggregated to the state areas using unions, and the cardinality of user days is being displayed in the choropleth map. Figure 2 shows the same data with scaled symbols showing cardinality of unique posts.

This same map could be created using raw data but one would have to store a significantly greater amount of it and the aggregation to state areas would take more processing time. The raw data also allows personally identifying information to be stored in the data base, thus presenting a privacy risk to the data owners. The privacy aware data structure does not have this draw back.

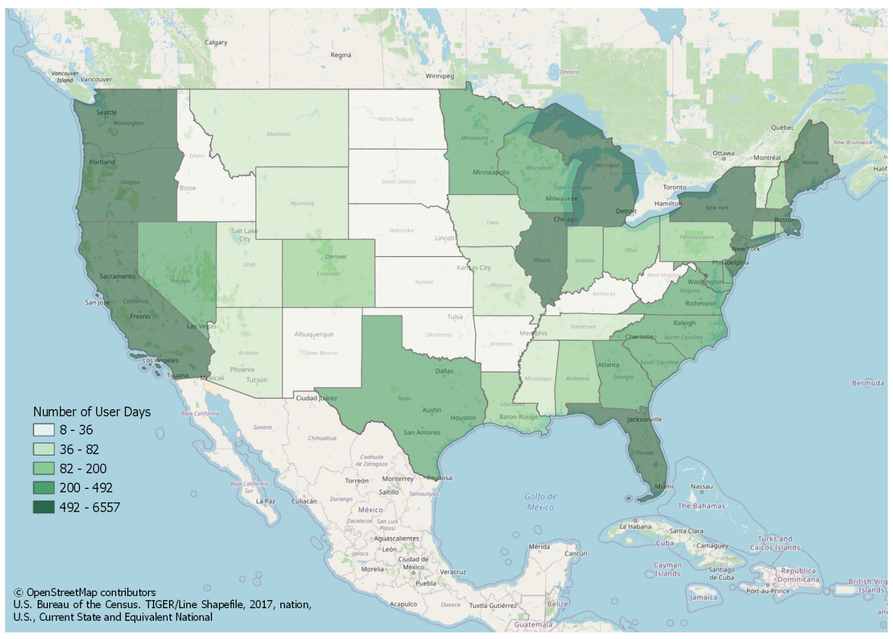

Figure 1 Number of User Days Reacting to Terms like "Beach" Aggregated by State

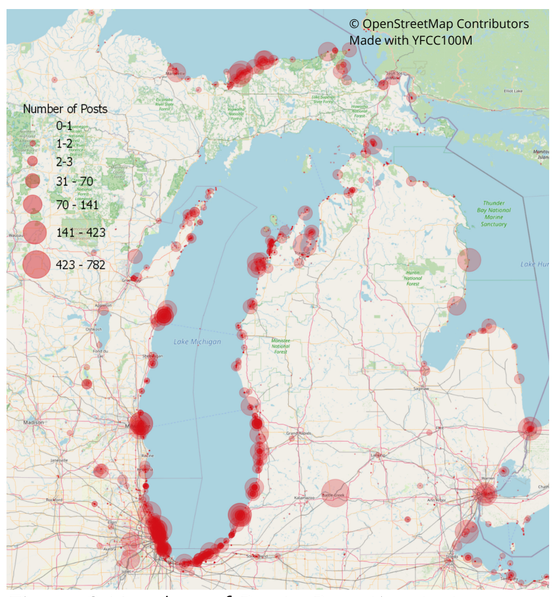

Figure 2 Number of Posts Reacting to terms like "Beach" around Michigan

Use Case

HyperLogLog is not privacy preserving unto itself; it must be combined with other methods to preserve privacy. In one study, it is demonstrated that cardinality estimators cannot preserve privacy [1]. In order to preserve privacy the database must be kept behind a restricted API and not made publicly available, as well as use a cryptographic hash to obscure the unique ids in the user data [1].

The use of this privacy aware data structure has the draw backs of limiting some types of visualizations that could be undertaken. For example, it would be impossible to show trajectories with this data structuresince trajectories require a unique identifier matching both a start and end point. Another disadvantage is that any researcher must know what they will be mapping before creating the privacy aware data structure.

Conclusions

The use of HLL for privacy aware analysis of spatial social media shows great potential. Its ability to be combined with other privacy aware techniques is one of its greatest advantages. Moreover, its potential to be incorporated into a privacy aware mapping library for a major programming language makes the continued research and development of HLL as a tool in the field of geoprivacy a worthwhile endeavor.

Thesis conducted at

Institute of Cartography
Department of Geosciences
Technische Universität Dresden

TECHNISCHE UNIVERSITÄT DRESDEN

Thesis Assessment Board

Chair Professor: Prof. Dipl.-Phys. Dr.-Ing. habil. Dirk Burghardt, Technische Universität Dresden

Supervisor: Dr.-Ing. Alexander Dunkel, Technische Universität Dresden

Supervisor: Marc Löchner M.Sc, Technische Universität Dresden

Reviewer: Wangshu Wang M.Sc, Technische Universität Wien

Year

2020

Keywords

Geoprivacy, HyperLogLog, Big Data, Social Media Data, YFC100M, Social Media Analysis, Spatial Analysis, LBSN Data, VGI, PII, Privacy

References

[1] Desfontaines, D., Lochbihler, A., & Basin, D. (2019). Cardinality Estimators do not Preserve Privacy. Proceedings on Privacy Enhancing Technologies, 2019, 26–46. https://doi.org/10.2478/popets-2019-0018.

[2] Flajolet, P., Fusy, É., & Gandouet, O. (n.d.). HyperLogLog: The analysis of a near-optimal cardinality estimation algorithm. 20.

[3] Summary—LBSN Structure. (n.d.). Retrieved August 28, 2020, from https://lbsn.vgisci-ence.org/.

[4] .Thomee, B., Shamma, D. A., Friedland, G., Elizalde, B., Ni, K., Poland, D., Borth, D., & Li, L.-J. (2016). YFCC100M: The New Data in Multimedia Research. Communications of the ACM, 59(2), 64–73. https://doi.org/10.1145/2812802

This master thesis was created within the Cartography M.Sc. programme – proudly co-funded by the Erasmus+ Programme of the European Union.

Technische Universität München · TECHNISCHE UNIVERSITÄT WIEN · TECHNISCHE UNIVERSITÄT DRESDEN · UNIVERSITY OF TWENTE. · Erasmus+ · cartographymaster.eu