Privacy aware analysis of spatial social media data

Jack Stephan

Primary Supervisor Marc Löchner and Alexander Dunkel

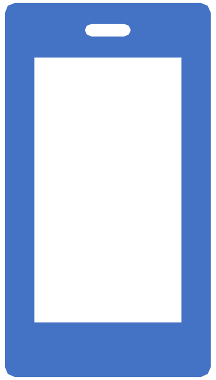
Promoter: D Burghardt

Reviwer: Wangshu Wang

Format of Presentation

1. Motivation
2. Problem Statement
3. Background
4. Method
5. Case Study
6. Discussion
7. Conclusions
8. Questions

Motivation



Big data and smart phone technology

Personal geographic information

Problem Statement

How can society continue benefit from the spatial analysis of social media data while ensuring the privacy of those individuals who are contributing the data in the first place?

Research

01

Analysis of social media data

02

Geoprivacy

03

HyperLogLog (HLL)

Research Questions

- What is meant by privacy regarding geographic information (GI)?
- Can treating GI with an HLL data structure increase the level of user privacy?
- Does treating this data with an HLL structure allow for the same quality of subsequent visualizations for social media research?

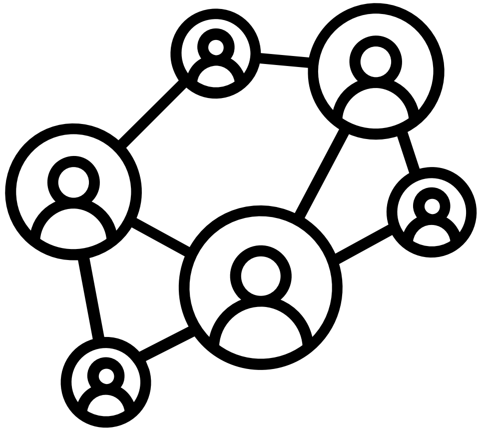
Research Questions cont.

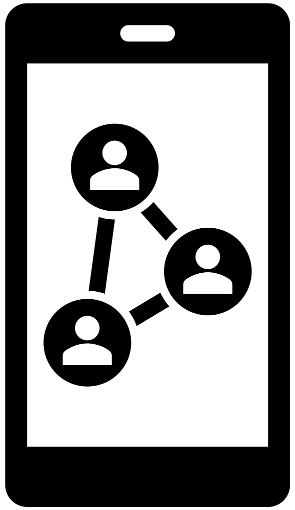
- Can the difference in privacy level be measured or qualified?
- Which set operations can be carried out on the HLL shards and what are the effects on the resulting visualization?
- What are the benefits and disadvantages to this database structure?
- What are the limitations of the HLL structure and its applications?

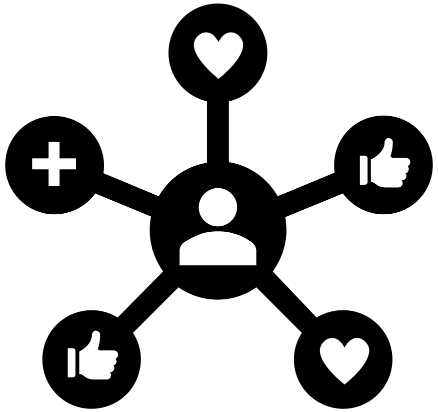
Background

Social Media Analytics

- LBS & LBSN
- Analysis of Social Media Data
- Analysis of Spatial Social Media







Privacy

What is Privacy?

- Nebulous Definition (free thought, autonomy, no surveillance, data ownership)
- Right to privacy vs concept of privacy
- Inherently relative

K-Anonymity

- Anonymity in a group i.e. city
- Property of data
- k-anonymity=$k-1$

Geoprivacy

- A subset of privacy
- Modern concept with modern tech
- Can build unique user profile
- Maintaining strict privacy is incompatible with LBSN

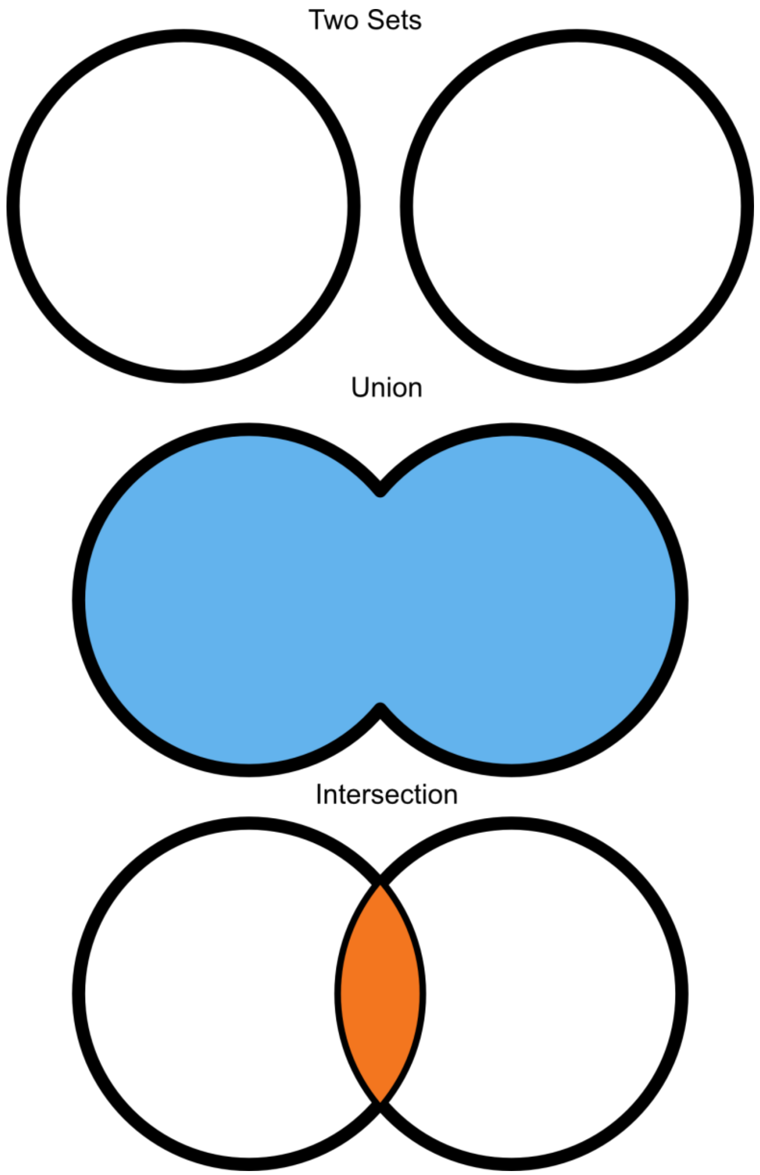
HyperLogLog

- Algorithm for estimating cardinalities
- Not for mapping or privacy, just a side benefit
- Cardinality Estimation
- Inherent error
- Lossless unions, continuous streaming

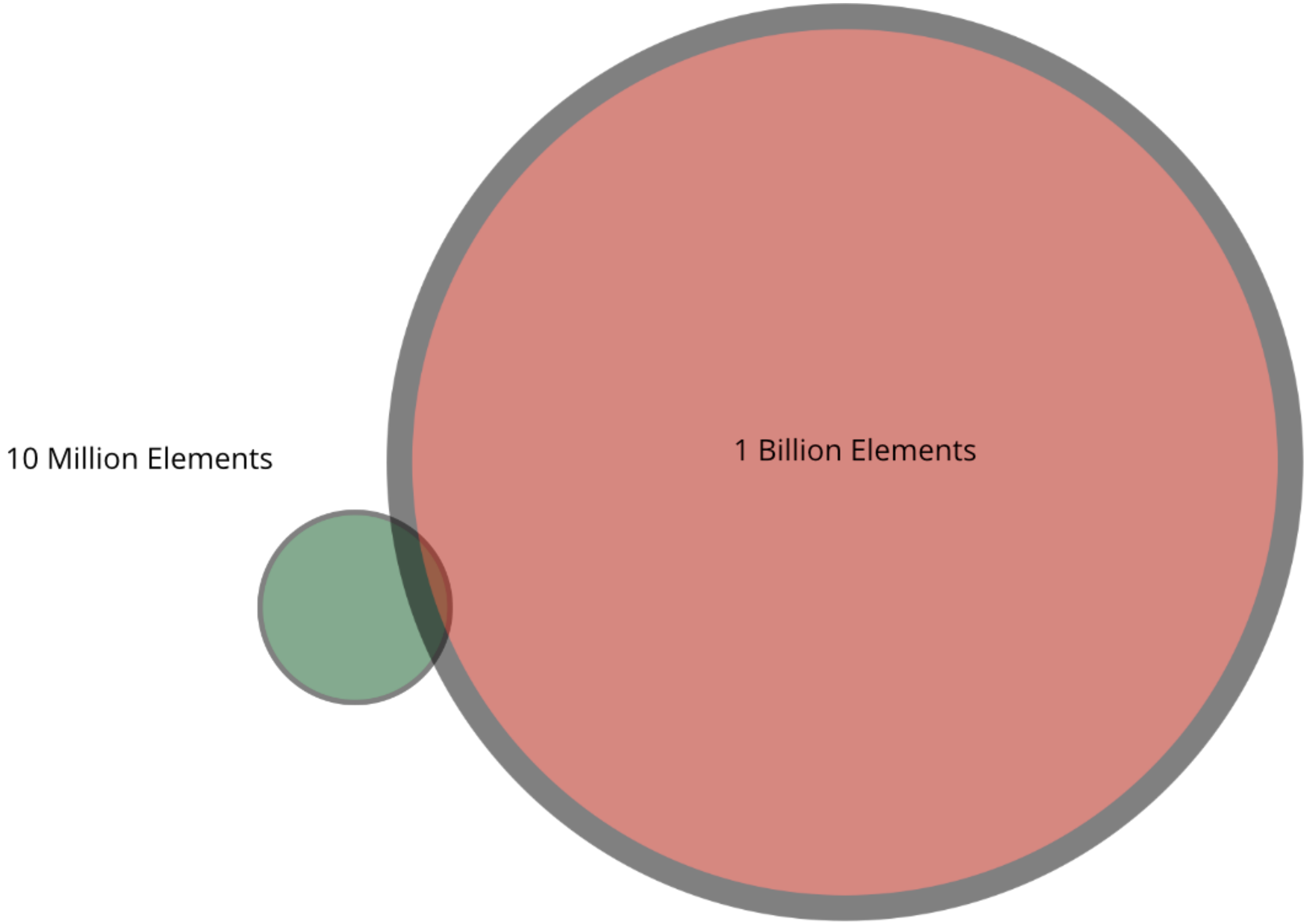
Intersections

Inclusion-Exclusion Principle

$$|A \cup B| = |A| + |B| - |A \cap B|$$



Intersection Error



Method

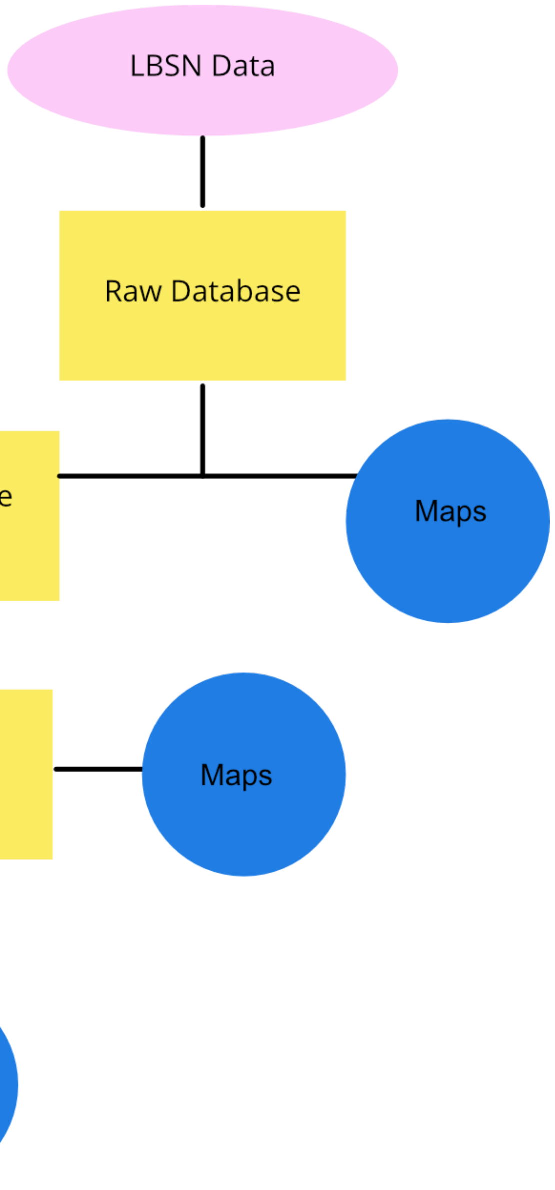
YFCC 100M Photo Database

- 100 m photo database
- Creative commons license
- Photos and videos

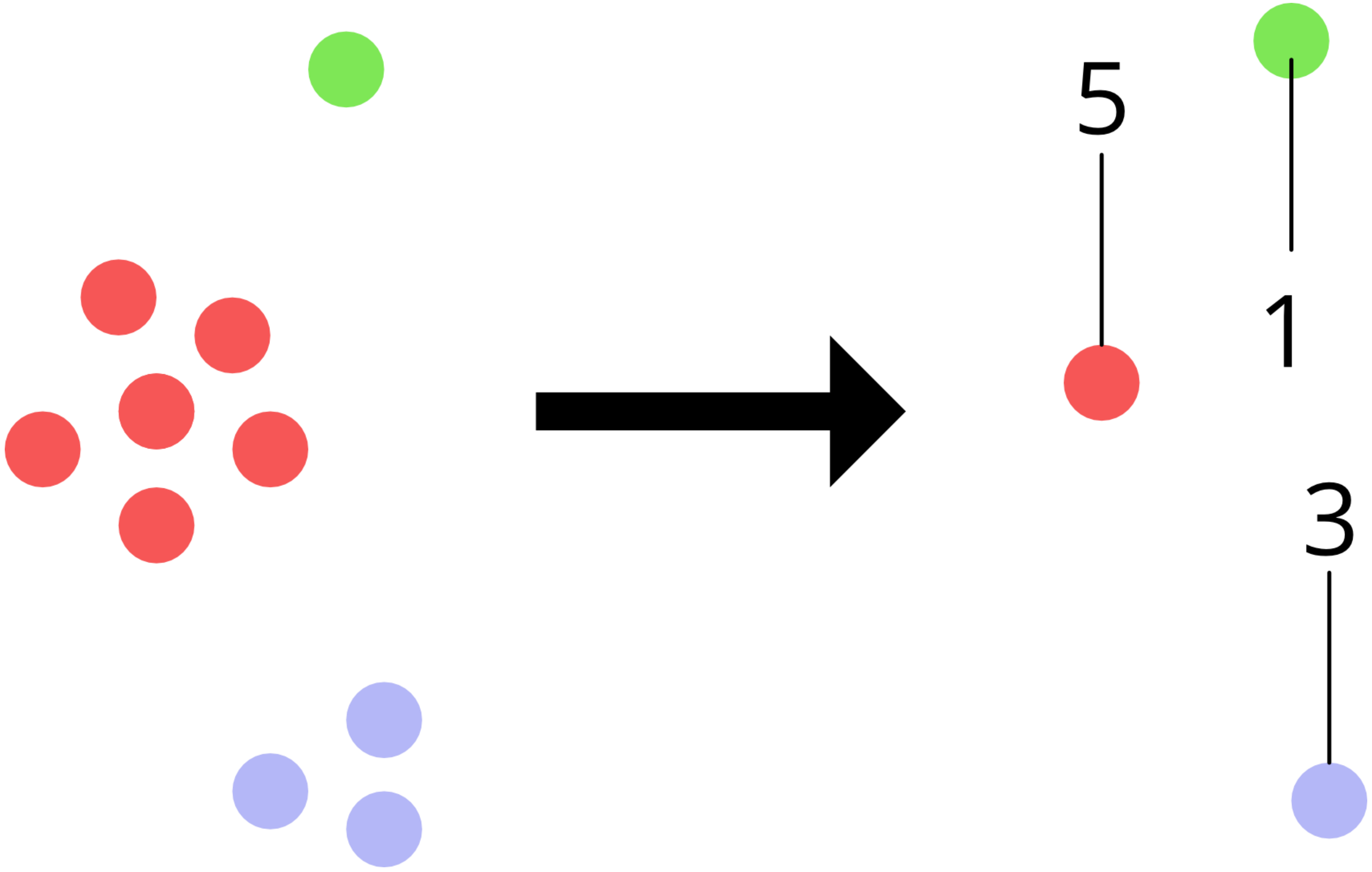
LBSN Data Structure

- Objects: entities from LBSN data e.g posts, users, places, and events
- Bases: characteristics of the object itself e.g. title, hashtags, post creation date
- Facets: topical, social, spatial, and temporal
- Overlays: also called metrics, are the bases used in the context of analysis e.g. post count, user count, user days

Workflow


LBSN Data
Raw Database
Privacy Aware Database
Maps
Aggregated
Maps
Maps

Data Processing Graphic



Evaluation of Results

- Definition of geoprivacy

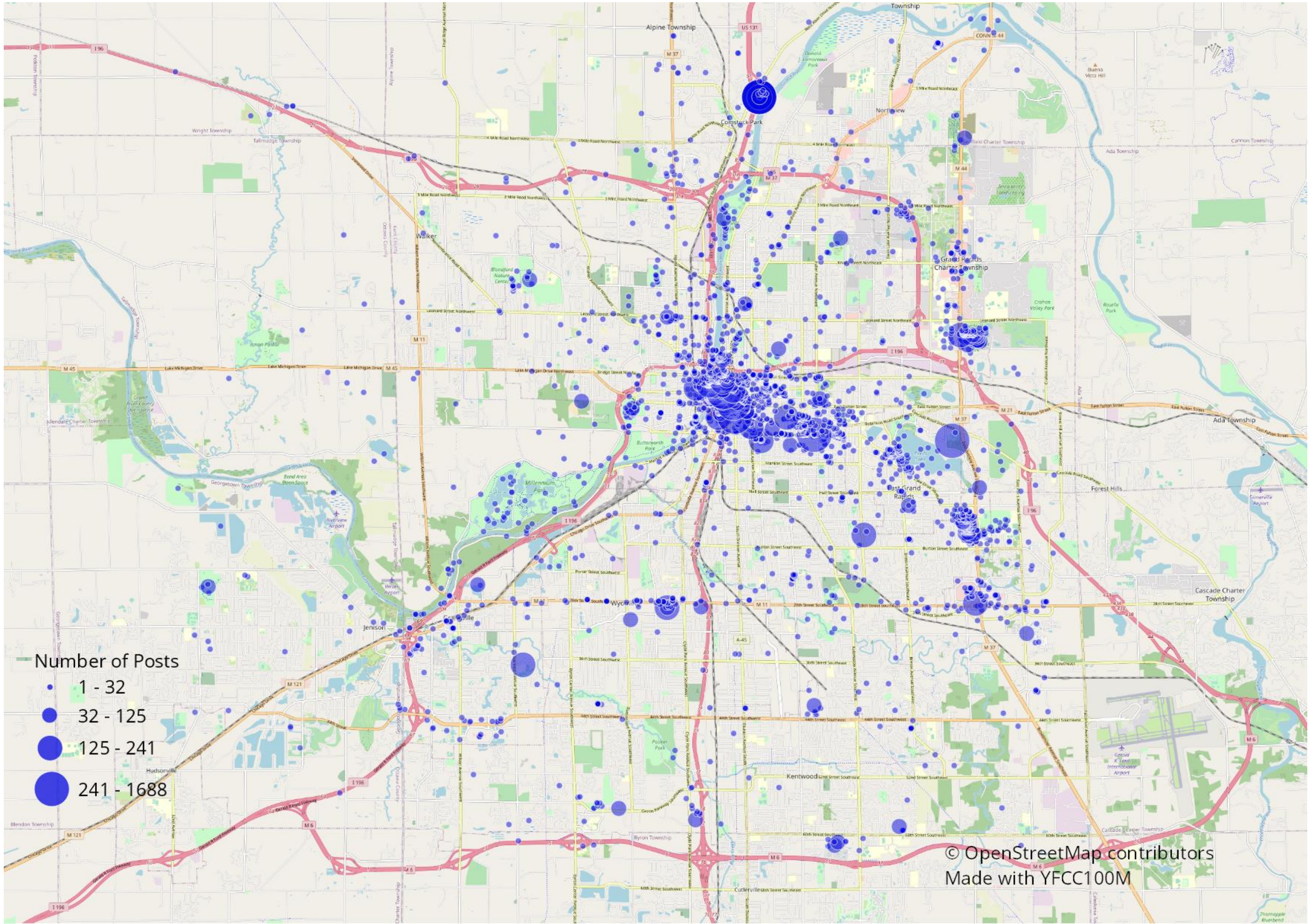
“whether or not a single user can be identified from the data set”

Case Study

Cardinality

- Number of unique elements in a set
- Cardinality is a useful measure when mapping

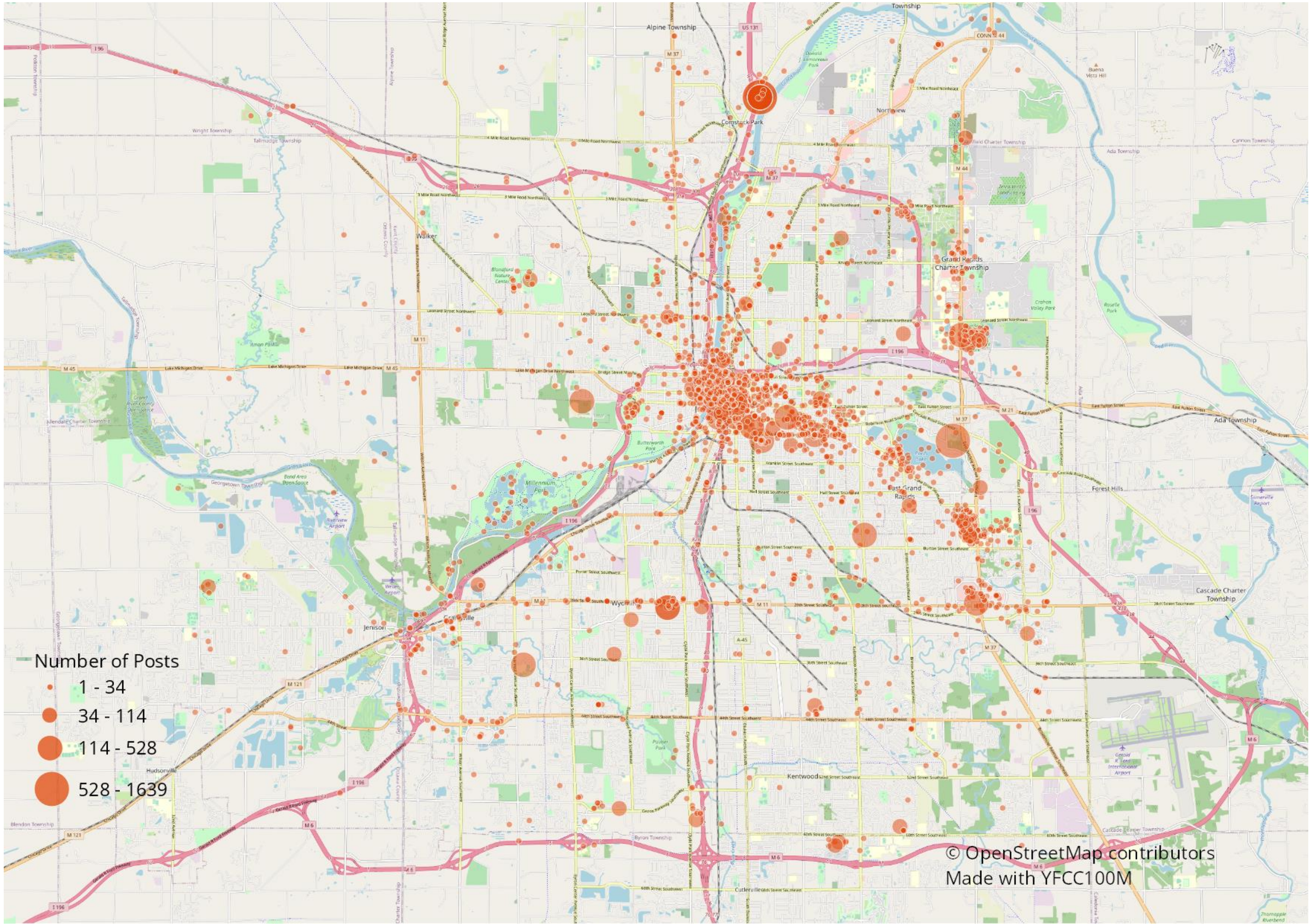
Number of Flickr Posts Grand Rapids, MI (HLL)



HLL Data

latitude	longitude	post_hll	hll_cardinality
42.850770	-85.625230	\x138b40dba2	1
42.851861	-85.635927	\x138b4002c103e 20a420...	15
42.851900	-85.720267	\x138b4028a2	1
42.852057	-85.633707	\x138b40000108 2109a30...	44
42.852057	-85.569591	\x138b40ef62	1

Number of Flickr Posts Grand Rapids, MI (Raw)



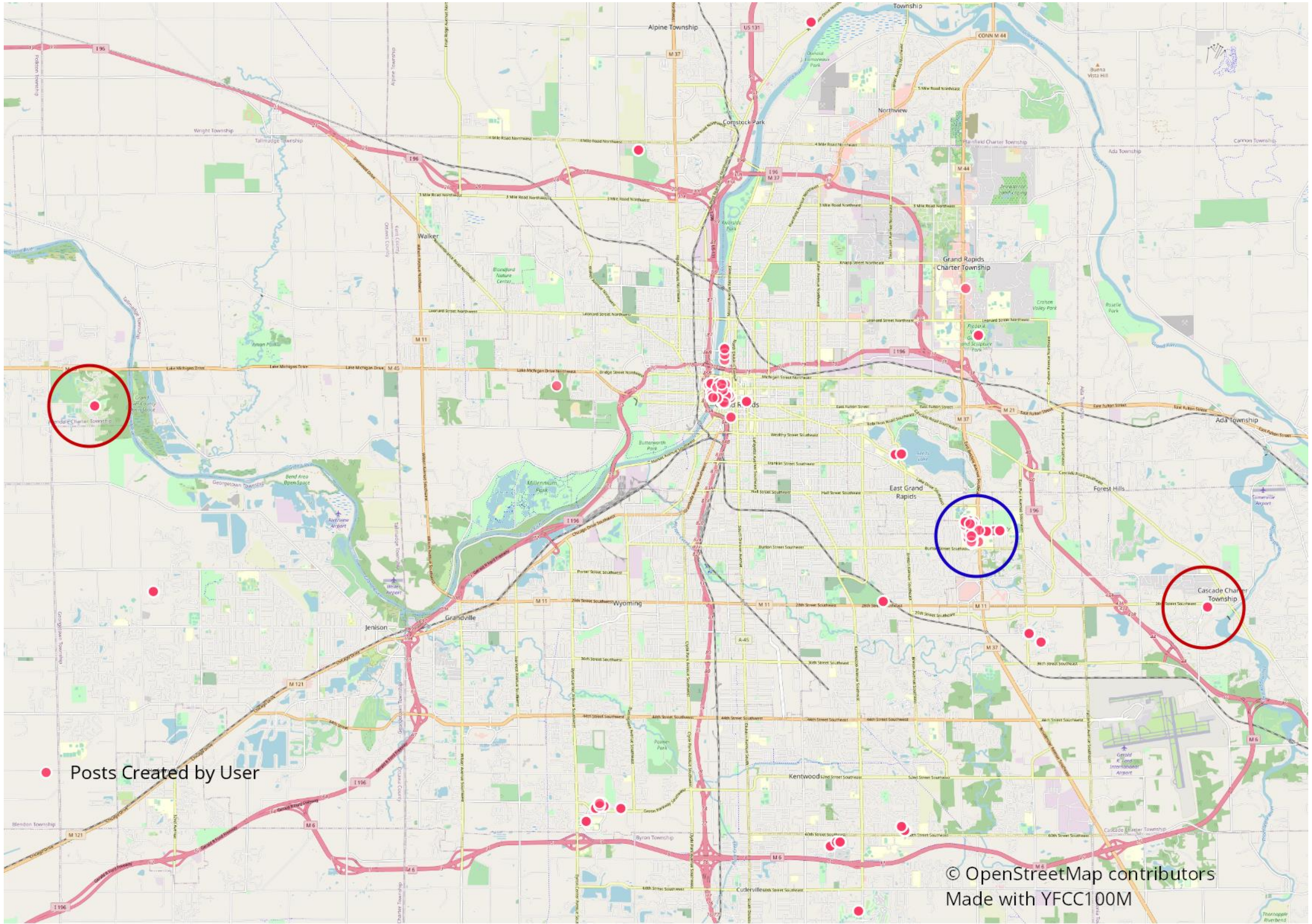
Comparison

	Raw	HLL	Percent Difference
Total Records	4518	4513	0.1%
Total Unique Posts	24110	23716	1.6%
Maximum Value	1639	1688	3.0%

Raw Data

post_guid	user_guid	post_publish_date	post_body	post_title	post_url	Longitude	Latitude
4514110401	19646481@N06	2010-04-12 15:39:55	"Grand Rapids" Michig...	Horse Abstract	http://www.flickr.com...	-85.610318	42.980979
6275585779	87815574@N00	2011-10-24 11:21:34	None	IMG_7024	http://www.flickr.com...	-85.593280	42.954035
796126334	87533529@N00	2007-07-13 07:54:33	Beloit @ West Michiga...	Michael Bertram	http://www.flickr.com...	-85.659531	43.040439
959276023	87533529@N00	2007-07-31 09:20:41	Peoria @ West Michiga...	Darwin Barney	http://www.flickr.com...	-85.659681	43.040557
10722950263	78629037@N03	2013-11-07 11:00:00	Micah.\n\nBreathe Owl...	Lamp Light: Breathe O...	http://www.flickr.com...	-85.637096	42.953250

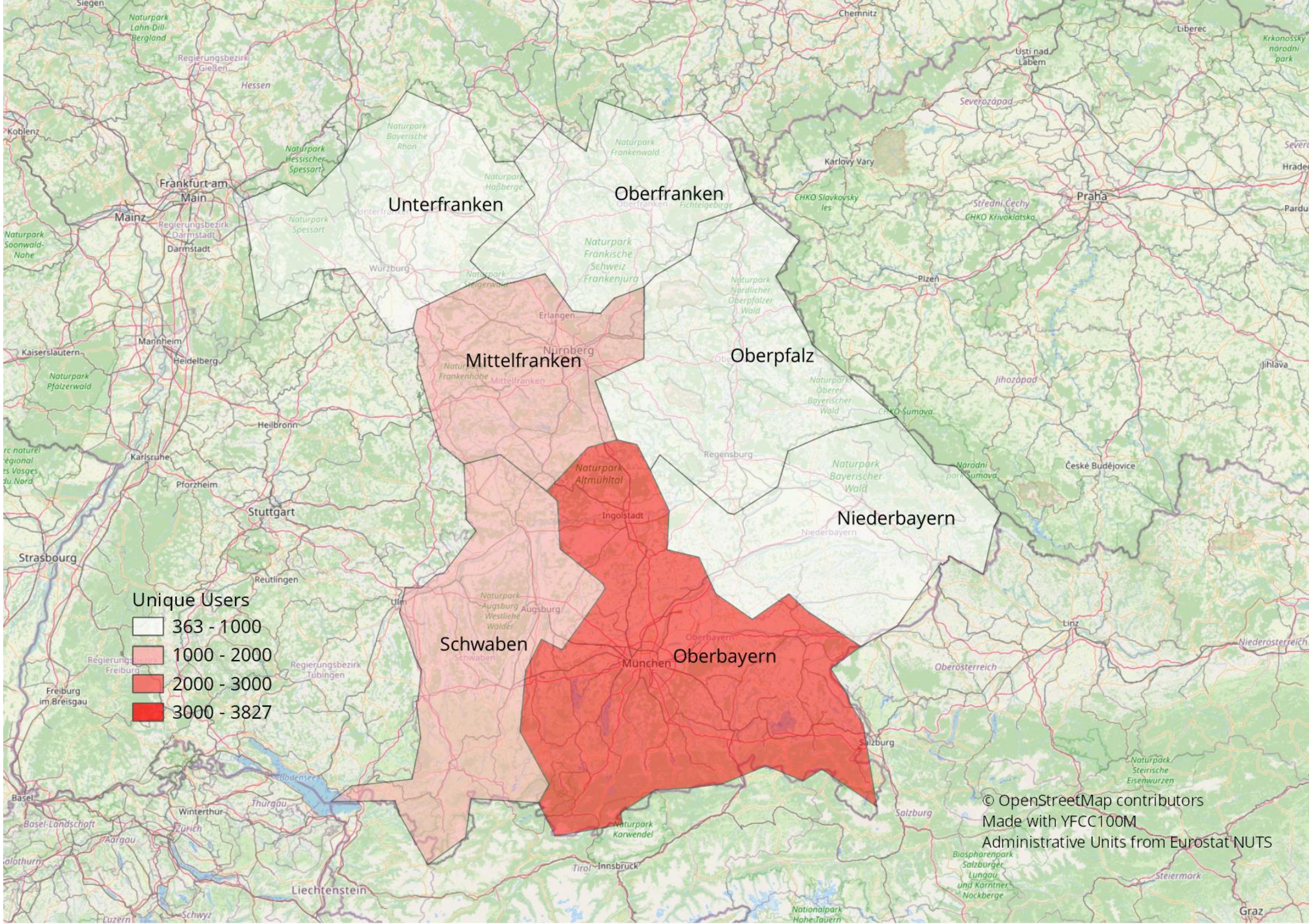
Reidentification of a User



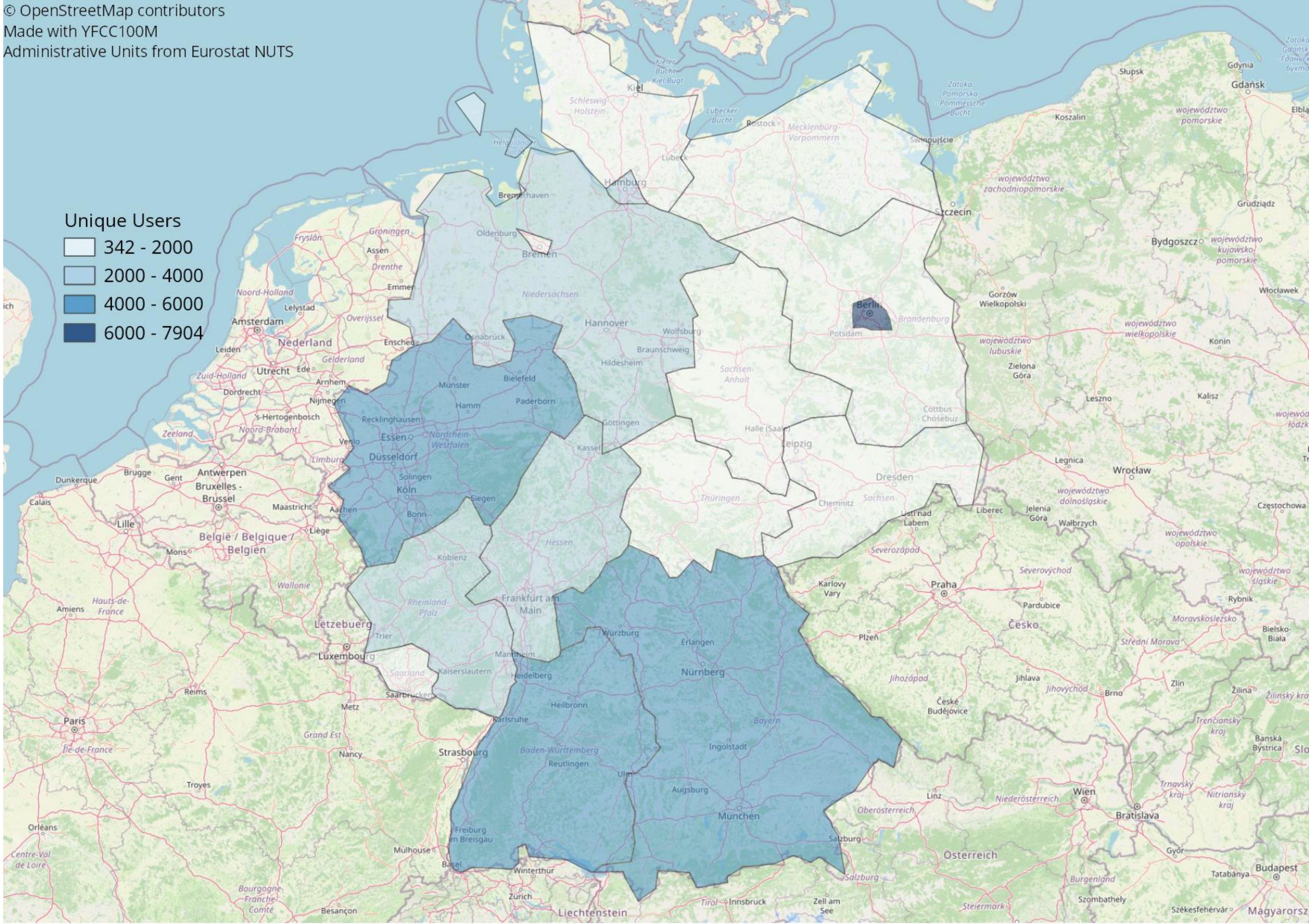
Union and Aggregation

- Lossless unions
- Can be aggregated spatially

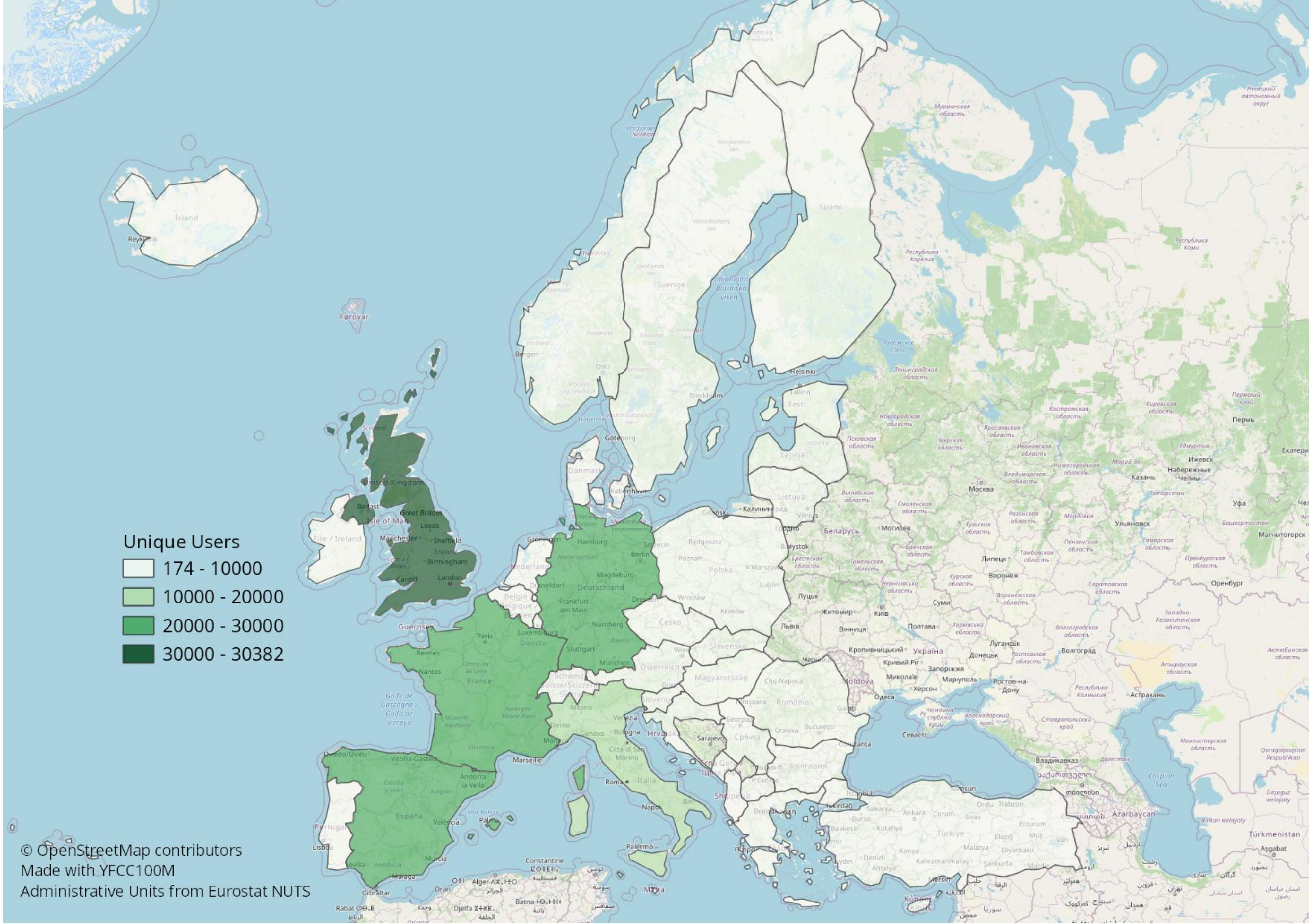
Unique Users in BY



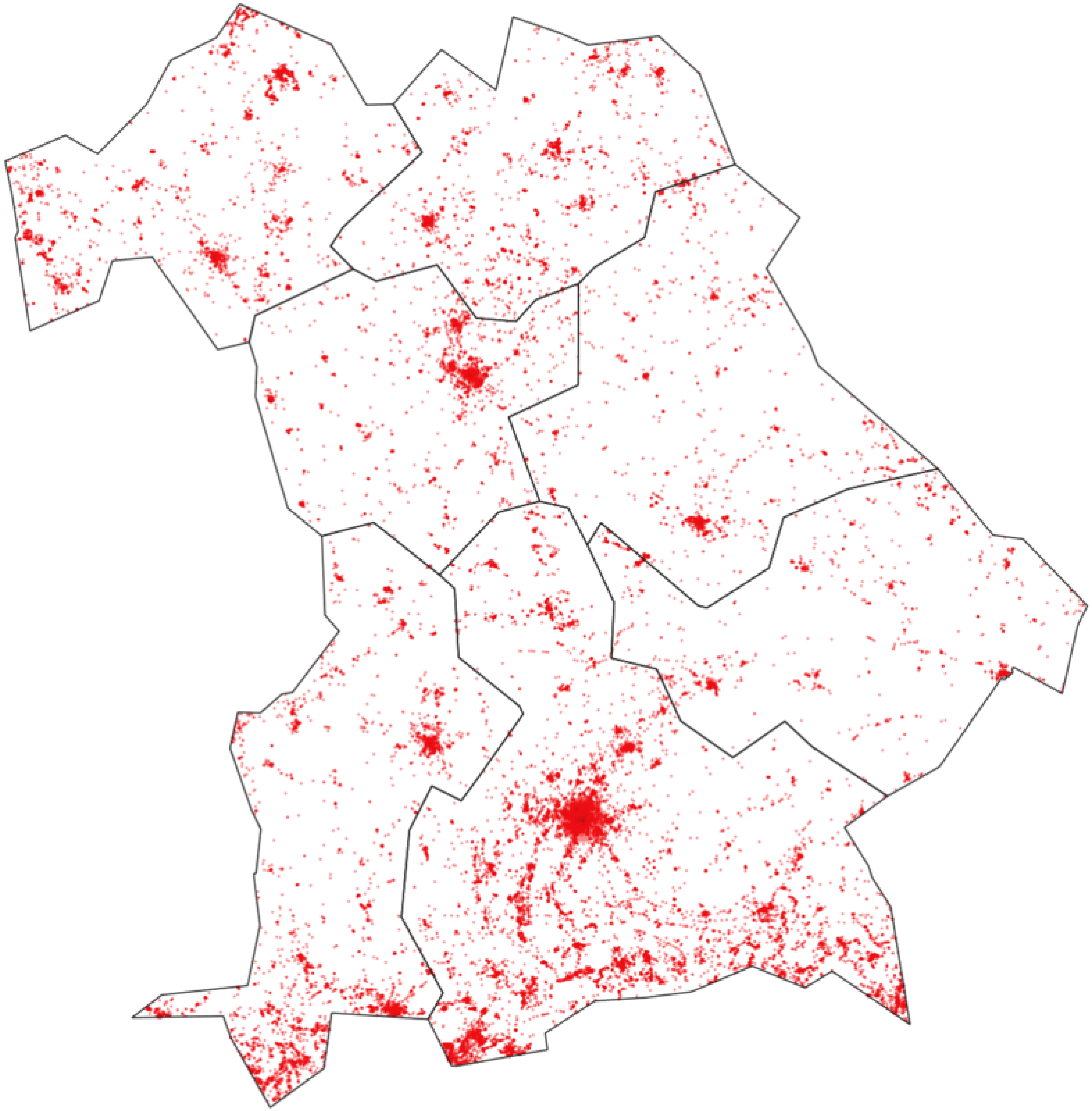
Unique Users in DE



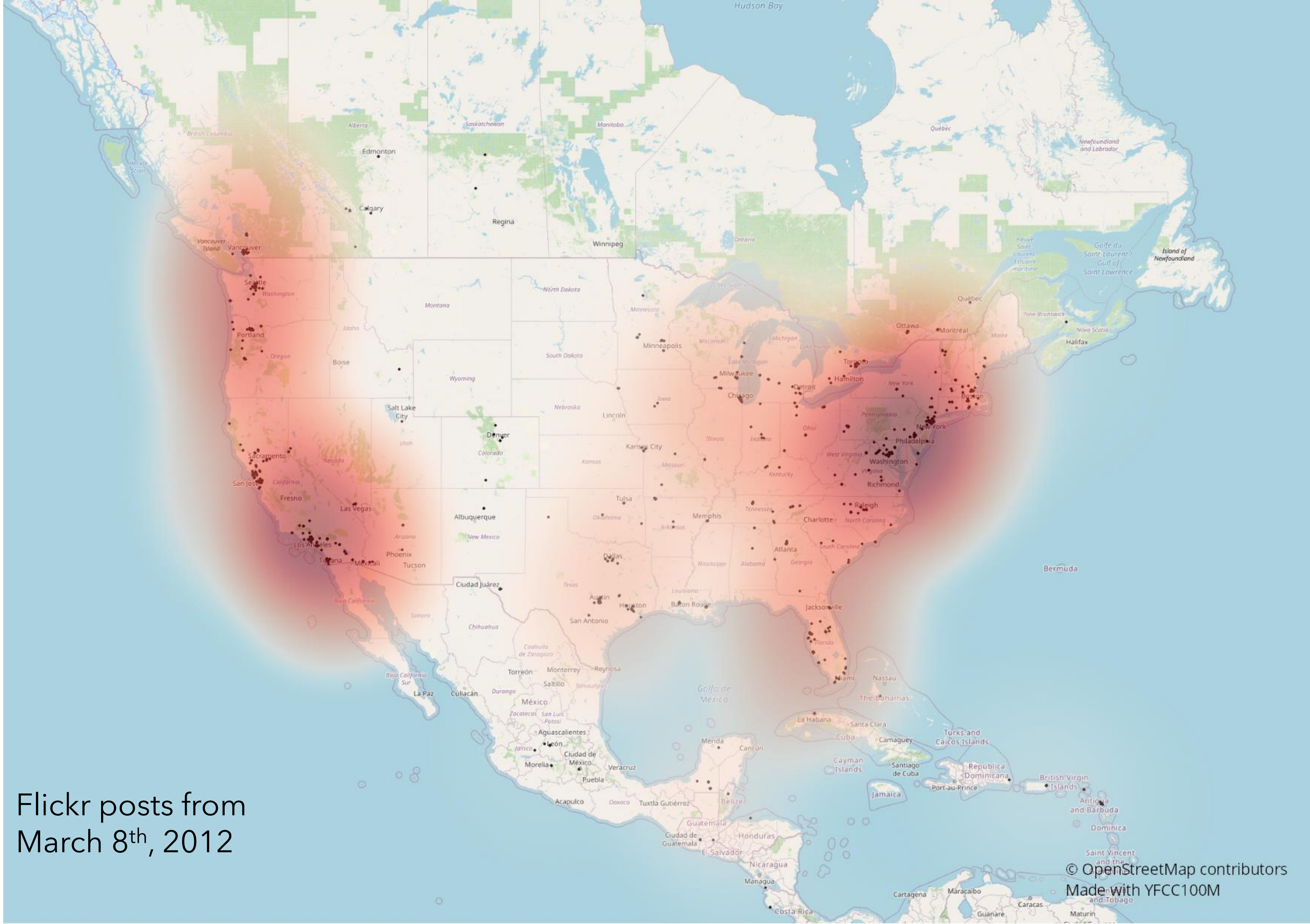
Unique Users in EU



Aggregation Animation

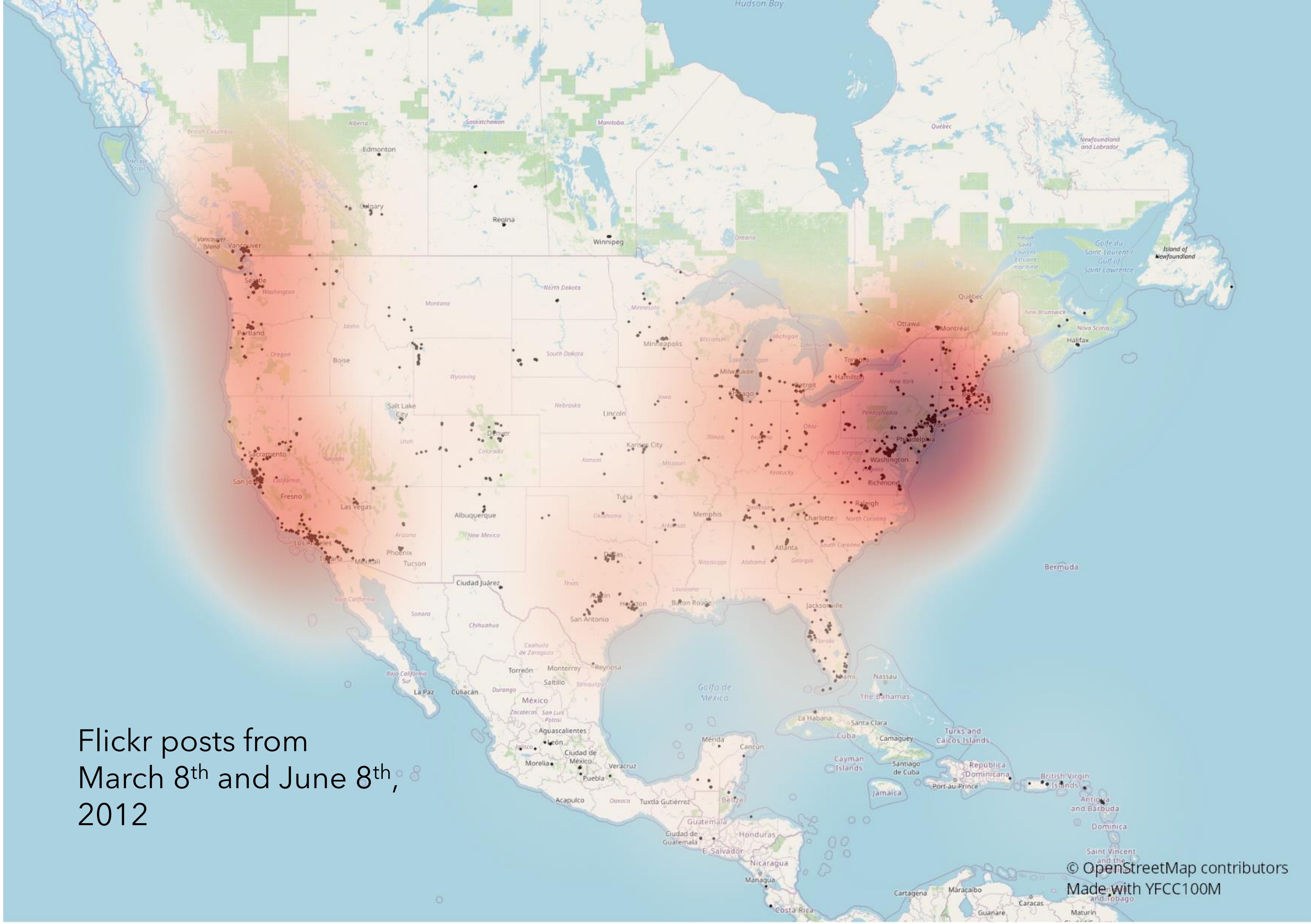


Union to Update



Flickr posts from March 8th, 2012

Union to Update



Flickr posts from March 8th and June 8th, 2012

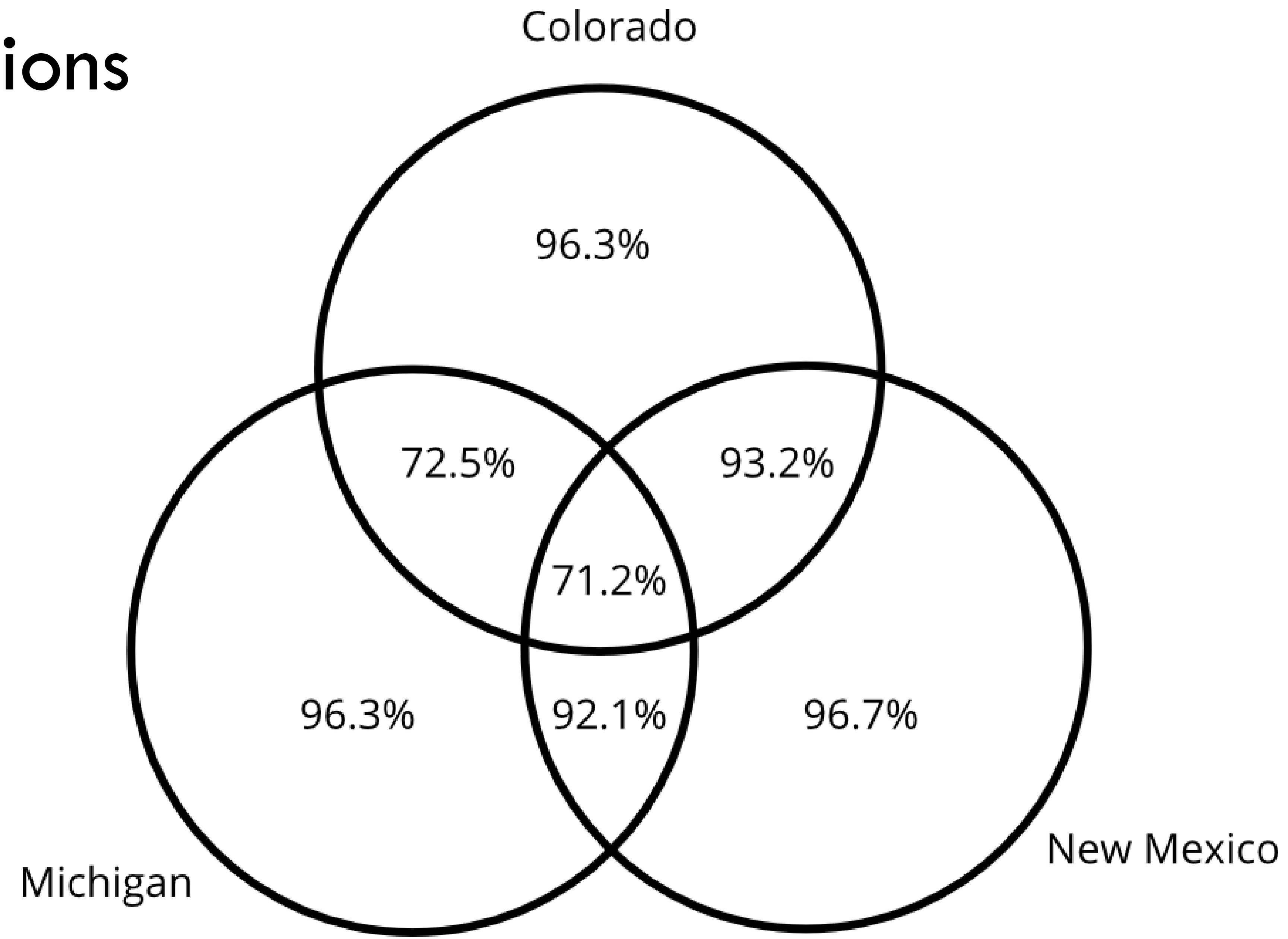
© OpenStreetMap contributors
Made with YFCC100M

Intersections

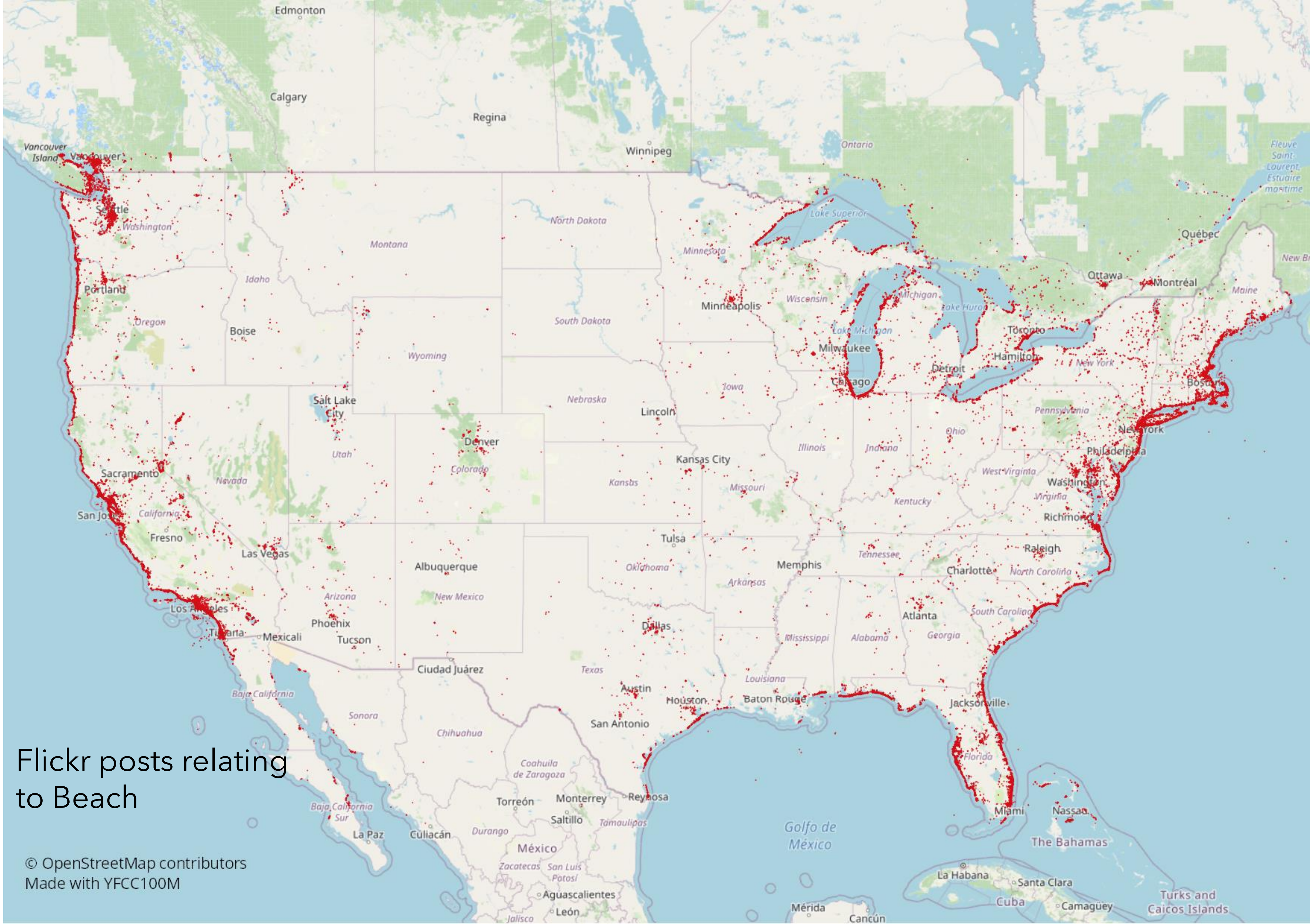
	Michigan	Colorado	New Mexico
Michigan	4127	890	279
Colorado	8275	5038	785
New Mexico	6118	6523	2270
Color Coding	**Union**	**Intersection**	**Cardinality**

$$|A \cap B \cap C| = 9724 - 4127 - 5038 - 2270 + 890 + 279 + 785 = 243$$

Intersections

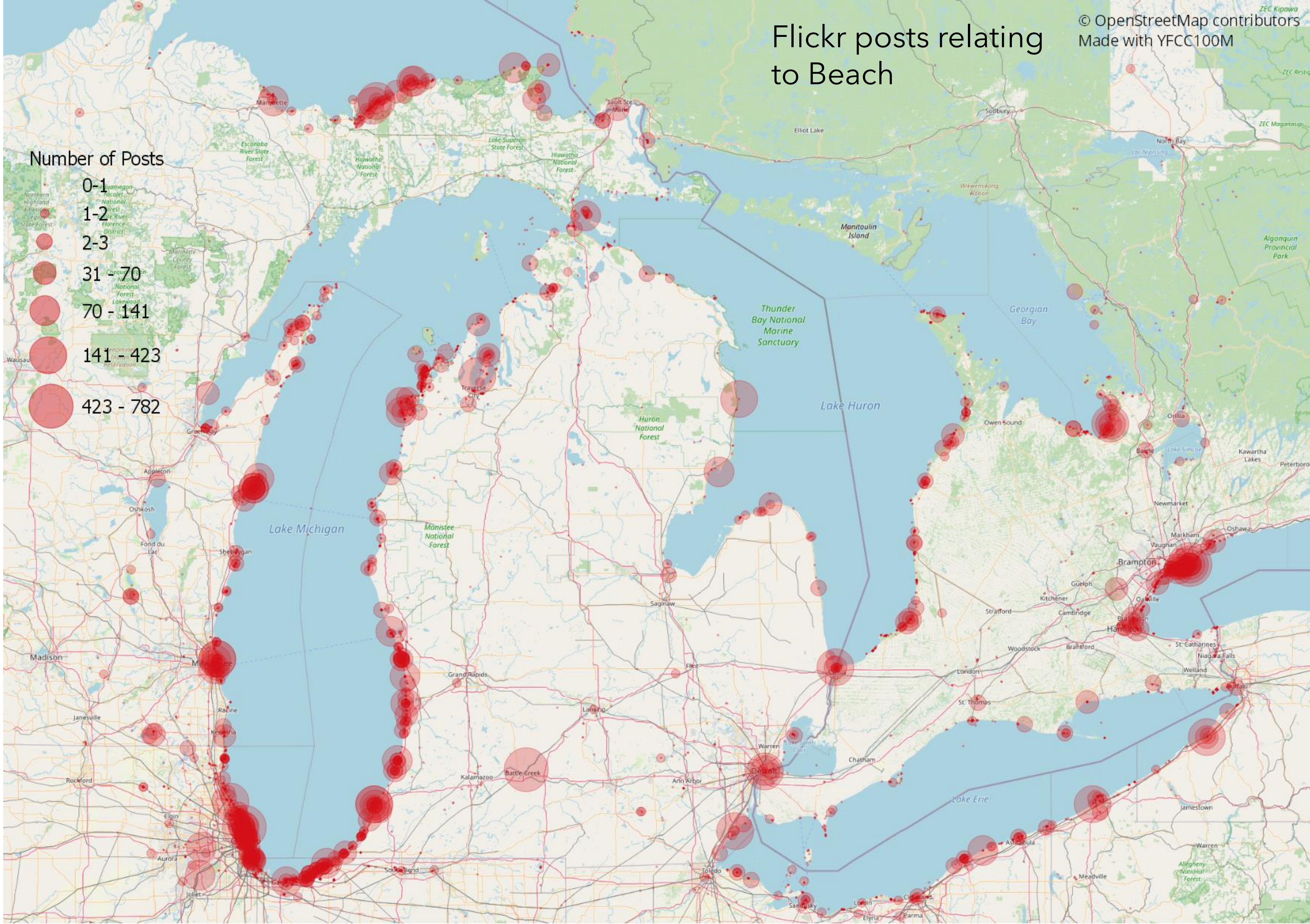


Filter by Topic

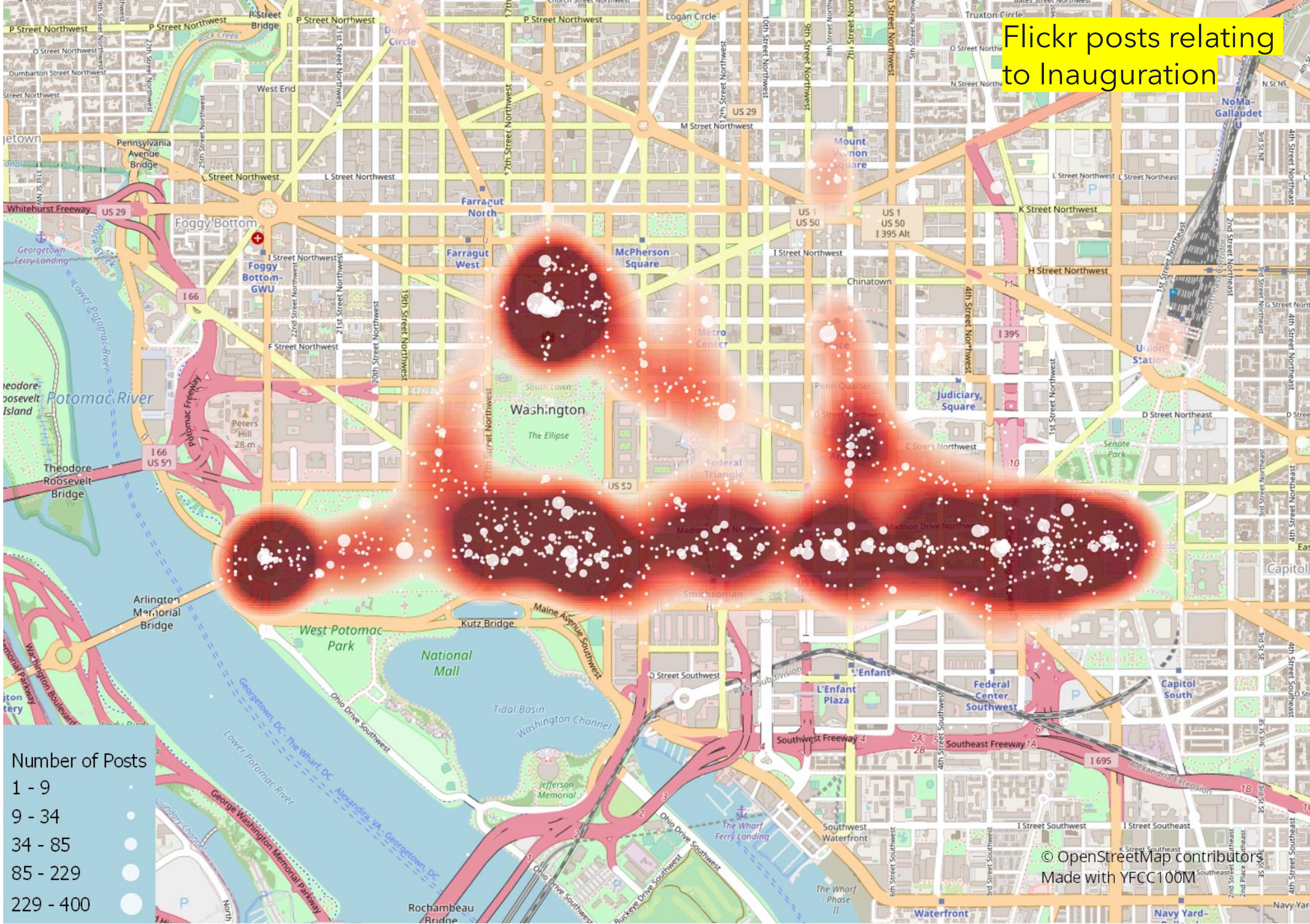


Flickr posts relating to Beach

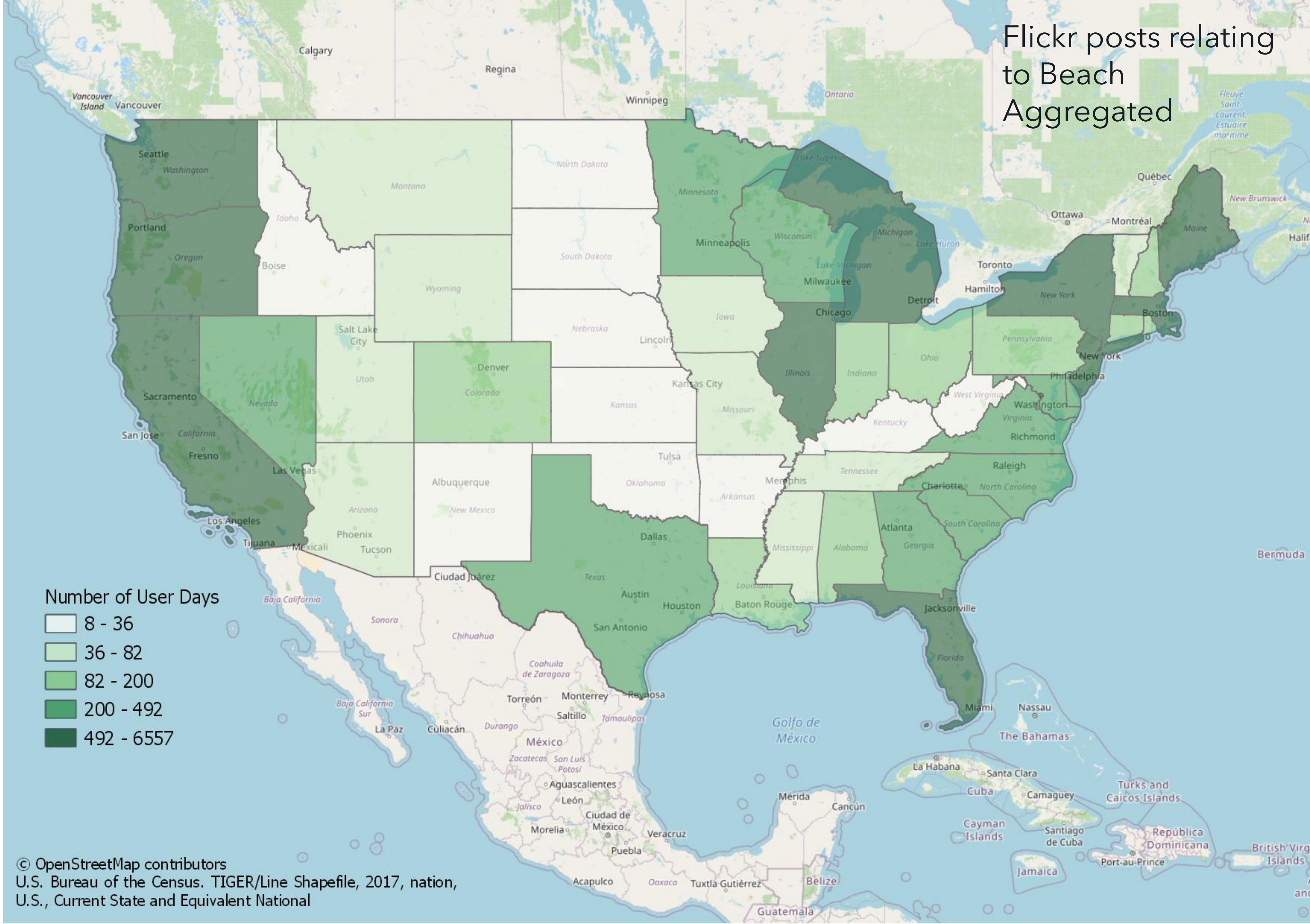
Filter by Topic



Filter by Topic



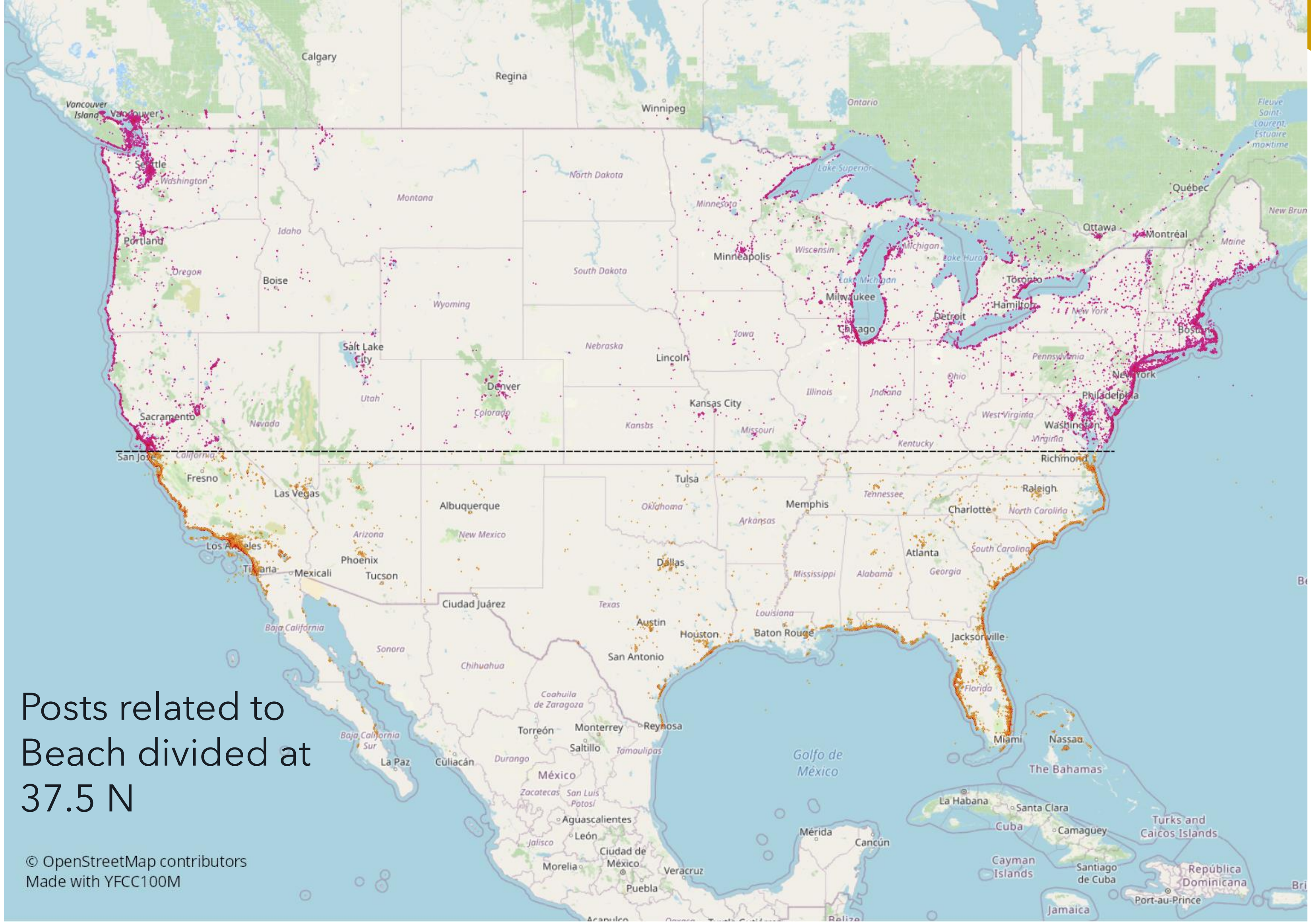
Combination of Techniques



Combination of Techniques

	Florida	California	Union	Intersection
Post Count	61726	143667	200695	4698
User Days	9351	23995	33519	173
User Count	2730	6557	8750	537

Combination of Techniques



Posts related to Beach divided at 37.5 N

	Above 37.5 N	Below 37.5 N	Union	Intersection
Unique Days	4120	4017	4560	3577
User Count	10189	8917	16636	2470

Discussion

Use Case

- “Cardinality Estimators do not Preserve Privacy” (Desfontaines et al., 2019).
- Cryptographic hash function
- Restricted API

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

Intersection Attack

- Problematic with only one user at a coordinate
- 99.6% reduction of records
- Abstract into classes

Coordinates with more than one unique user

© OpenStreetMap contributors
Made with YFCC100M

Advantages

- Reduction of data stored
- Increased processing speed
- 2% error is usually sufficient
- Flexible e.g. cryptographic hash function

Disadvantages

- Some limitation in what can be mapped
- Intersection errors
- Foresight into data generation
- No ability to disaggregate, privacy trade-off

Ease of Deployment & Recommendations

- Maintained connection to the database
- Already proposed LBSN data structure
- Set operations are quite intuitive
- Programming language library e.g. Python with DBT

Conclusions

- HLL for privacy aware analysis of spatial social media data shows promise
- The advantages of HLL allow it to be leveraged for social media analytics and big data applications
- The disadvantages are surmountable and can be further explored
- A more user-friendly data privacy library is worth exploring

Questions