Please Forget Where I Was Last Summer: The Privacy Risks of Public Location (Meta)Data

Kostas Drakonakis, Panagiotis Ilia, Sotiris Ioannidis, Jason Polakis

NDSS 2019 - San Diego, CA, 25 February 2019

Location (Meta)Data & Services

- Fine-grained location information collected by most modern devices
  - Smartphones
  - Wearables

- Enables a range of novel functionality
  - Additional microblogging context
  - Enhance situational awareness
  - Enrich user experience

https://www.finjanmobile.com/mobile-location-services-privacy-and-security-issues/

What about privacy?

- Pose significant privacy risks for users

- Users' key location inference can lead to:
  - Deanonymization
  - Physical threats, stalking

- Other location points can lead to:
  - User profiling
  - Inference of sensitive traits (e.g. health issues)

kostasdrk@ics.forth.gr

https://www.wired.com/story/twitter-location-data-gps-privacy/

Prior Work & Motivation

- ❖ Multiple studies on home and work inference using location data
  - ➢ Cheng et al. ICWSM ’11
  - ➢ Cho et al. KDD ‘11
  - ➢ Efstathiades et al. ASONAM ‘15
  - ➢ Hu et al. ICDMW ‘15 etc.

- ❖ Coarse granularity in their inference (e.g. zip code, city)
  - ➢ Could not highlight the true extent of the privacy risks

- ❖ Automated sensitive information inference remains unexplored

GPS Coordinates and Where to Find Them

- Our case study is on Twitter

kostasdrk@ics.forth.gr

GPS Coordinates and Where to Find Them

- Our case study is on Twitter

MOTHERBOARD

Hundreds of Bounty Hunters Had Access to AT&T, T-Mobile, and Sprint Customer Location Data for Years

Documents show that bail bond companies used a secret phone tracking service to make tens of thousands of location requests.

SHARE TWEET

In January, Motherboard revealed that AT&T, T-Mobile, and Sprint were selling their customers' real-time location data, which trickled down through a complex network of companies until eventually ending up in the hands of at least one bounty hunter. Motherboard was also able to purchase the real-time location of a T-Mobile phone on the black market from a bounty hunter source for $300. In response, telecom companies said that this abuse was a fringe case.

In reality, it was far from an isolated incident.

The New York Times

Your Apps Know Where You Were Last Night, and They're Not Keeping It Secret

Dozens of companies use smartphone locations to help advertisers and even hedge funds. They say it's anonymous, but the data shows how personal it is.

By JENNIFER VALENTINO-DeVRIES, NATASHA SINGER, MICHAEL H. KELLER and AARON KROLIK DEC. 10, 2018

The millions of dots on the map trace highways, side streets and bike trails — each one following the path of an anonymous cellphone user.

One path tracks someone from a home outside Newark to a nearby Planned Parenthood, remaining there for more than an hour. Another represents a person who travels with the mayor of New York during the day and returns to Long Island at night.

Dataset

- Twitter's public Streaming API to collect seed UIDs
  - US mainland
  - 308,593 users
- Collected each user's timeline
  - Up to 3,200 tweets
- Consider only official Twitter apps and Foursquare
  - 87,114 users with geotagging activity
  - 15,263,317 geotagged tweets

Analysis & Evaluation Datasets

- Two subsets
  - Top-6K: ~6K users with the most geotagged tweets
  - Low-10K: ~10K random users with 10 - 250 geotagged tweets

- Allows to study the differences between prolific and restrained users

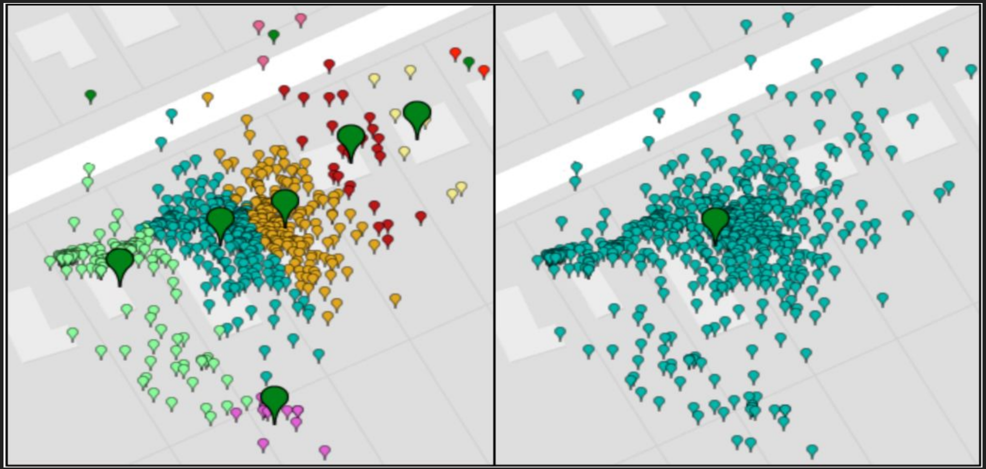
Location Clustering

- 1st level clustering
  - ArcGIS API maps coordinates to postal address
    - Cache results to reduce redundant API calls

- 2nd level clustering
  - Certain 1st-level clusters correspond to the same location
    - GPS errors
    - User leaving/arriving at location
    - Precision of geocoding API

Location Clustering

- ❖ 2nd level clustering
  - ➢ A larger cluster is surrounded by smaller ones
  - ➢ Merge secondary clusters with dominant one using DBSCAN
  - ➢ Enhances cluster’s “signal”



Ground Truth Datasets

- Manual and strict workflow to generate accurate ground truth
  - 2 independent annotators
  - Discarded ambiguous users
- Inspected clusters matching key phrases and the 10 largest clusters
  - “At home”, “This job” etc
- Final ground truth datasets:
  - Home-Top: 1,004 users (Work-Top: 298 users)
  - Home-Low: 1,043 users (Work-Low: 92 users)

Key Location Inference

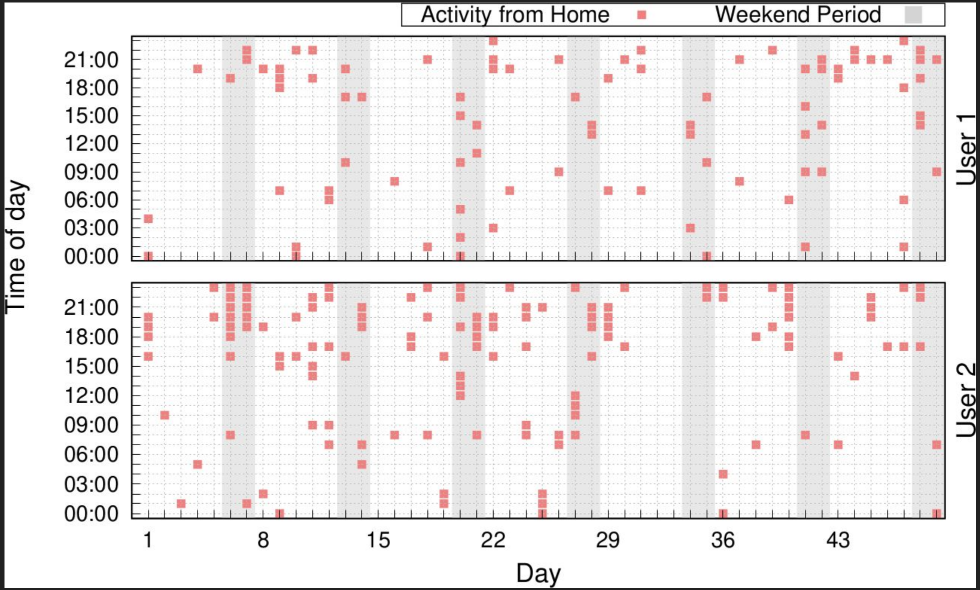
- ❖ Process spatiotemporal (meta)data
  - ➢ Social-graph and content agnostic

- ❖ Guided by common societal and legislative norms in the US and EU
  - ➢ E.g., 8 hour work shifts

Home Inference

- ❖ Expected behavior
  - ➢ Repeated activity
  - ➢ No specific time frame

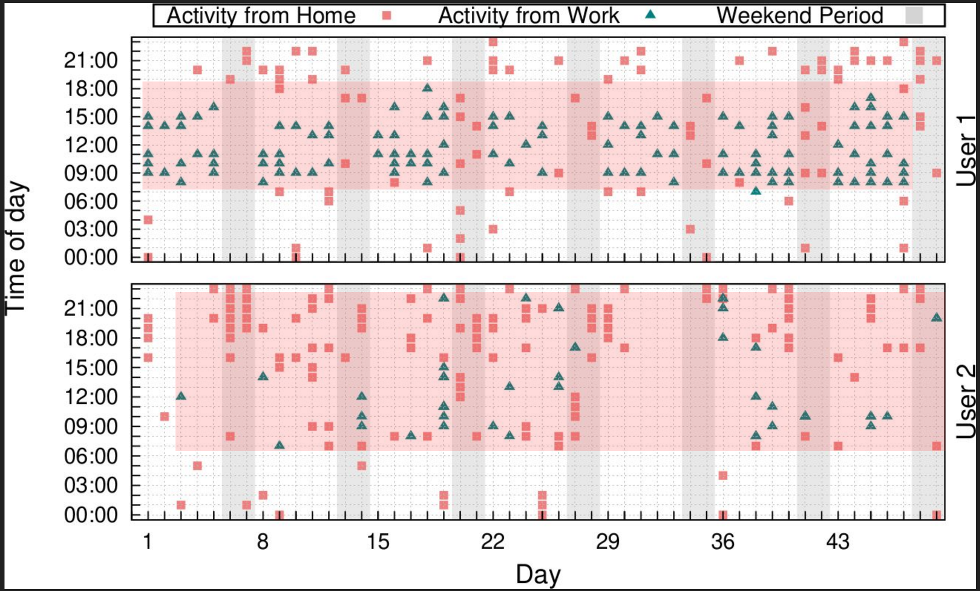
- ❖ Our heuristic
  - ➢ Only consider weekends
  - ➢ Select 5 most active clusters
  - ➢ Pick cluster with the widest time frame



Work Inference

- Expected behavior
  - Some repeated activity
  - Well defined time frame

- Our heuristic
  - Consider entire weeks
  - Select 5 most active clusters
  - Dynamically identify the dominant time frame (DTF) for each cluster
  - Pick most active cluster (entire weeks) during the DTF



Key Location Inference Evaluation

Dataset	Users	Inferred Clusters	Precision
Home-Top	1004	926	92.2%
Home-Low	1043	969	92.9%
Work-Top	298	164	55%
Work-Low	92	53	57.6%

Comparison to Prior Work

- Replicate 11 approaches for home and 2 for work inference
  - Run them on our ground truth
  - Apply 1st-level clustering on prior approaches
    - Faithful to their original design

- Outperform all prior approaches
  - Best home: 73.3% [Hu et al. ‘15], +18.9% improvement
  - Best work: 48.9% [Efstathiades et al. ‘15], +8.7% improvement

What more can we infer from a user's location history?

Identifying Highly Sensitive Places

- Identify *Potentially Sensitive Clusters* (PSCs)
  - In close proximity to sensitive venues

- Collect venue information from Foursquare
  - Within 25m from cluster’s midpoint
  - Categories pertaining to health, religion and sex/nightlife

- Determine whether the user actually visited them
  - Proximity != Visiting the venue
  - Need to increase confidence

kostasdrk@ics.forth.gr

Identifying Highly Sensitive Places

- Content-based corroboration
  - Manually compiled wordlist for each category
  - 3 most significant terms (*tf-idf*) matched against the respective wordlist
    - If there is a match, the user was likely visiting that venue

- Duration-based corroboration
  - Repetitiveness and duration of visits
  - Consider clusters with activity spanning hours or even days
  - Exclude clusters with short duration (passer-by cases)

- Location metadata might disclose more than the user intended

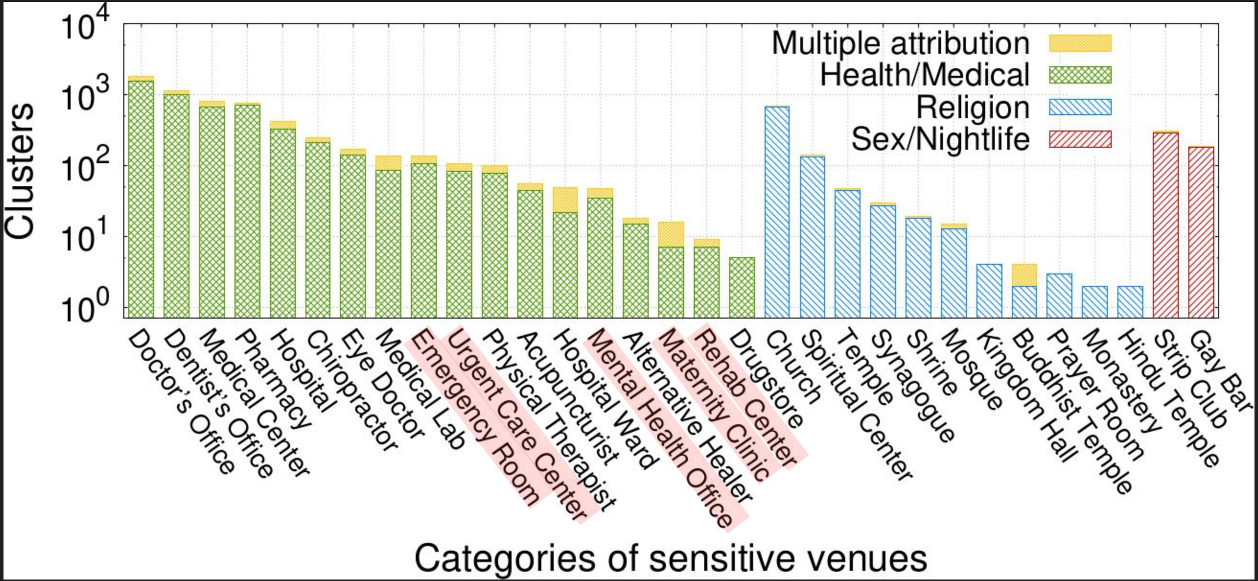
Identifying Highly Sensitive Places

Location metadata magnifies privacy loss

- Duration-based corroboration
  - Repetitiveness and duration of visits
  - Consider clusters with activity spanning hours or even days
  - Exclude clusters with short duration (passer-by cases)

- Location metadata might disclose more than the user intended

Potentially Sensitive Clusters



- 5,094 medical
- 918 religion
- 471 sex/nightlife

Content-Based Corroboration

- Ground truth users with PSCs: 1,454 (6,483 PSCs)
- Detected sensitive clusters: 545
  - Manually verified by inspecting all clusters including a wordlist term
  - Precision: 80.36%
  - Recall: 93.79%
- When applied on the main datasets:
  - Top-6K: 1,512 detected (21,863 PSCs)
  - Low-10k: 474 detected (6,918 PSCs)

Duration-Based Corroboration

- ❖ Users with DB clusters:
  - ➢ Home-Top: 691 (1,699 clusters)
  - ➢ Home-Low: 205 (276 clusters)

- ❖ ~53% and ~44% of the CB clusters also detected by the DB approach
  - ➢ Both techniques can be combined for higher confidence

- ❖ When applied on the main datasets:
  - ➢ Top-6K: 7,020 detected clusters
  - ➢ Low-10k: 2,337 detected clusters

Twitter’s Policy & Historical Data

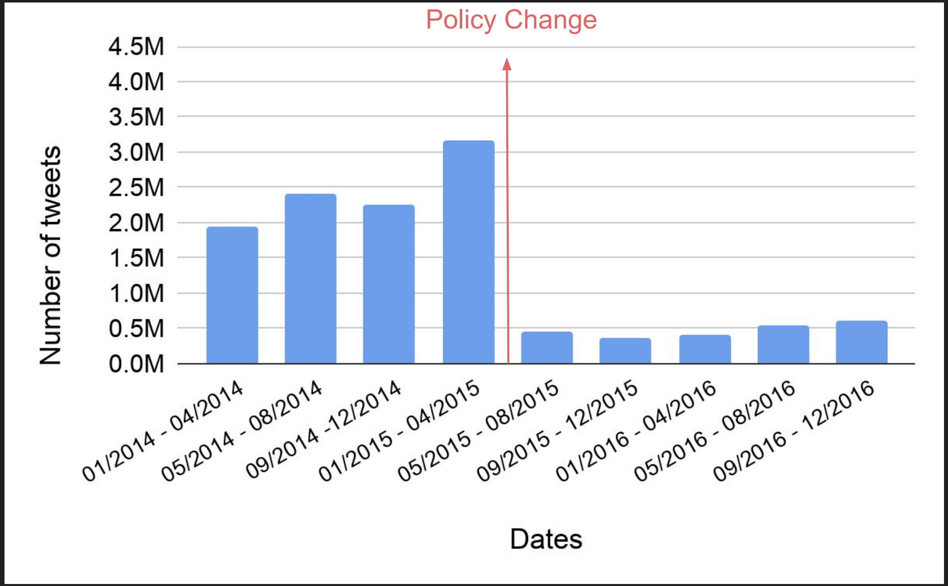
- Prior to April 2015:
  - Apps included coordinates even in coarsely tagged tweets
  - Only accessible via the API

- Since April 2015:
  - Privacy-respecting policy
  - Users must opt-in to add precise location information

- This historical data remains publicly accessible through the API

kostasdrk@ics.forth.gr

User Behavior Through Time



- Significant decrease in geotagged tweets after April 2015

Impact of Historical Data

Dataset	Date	Users	Homes	Coverage
Home-Top	Release	602	333	35.96%
Home-Top	+4 Weeks	155	68	7.34%
Home-Low	Release	394	239	24.66%
Home-Low	+4 Weeks	116	62	6.39%

- 15.43% and 11.12% of users had geotagged tweets 4 weeks later
- Precision drops to 43.87% and 53.44%

kostasdrk@ics.forth.gr

Takeaways

- ❖ Designed novel techniques to infer:
  - ➢ Users' key locations, with high precision and granularity
  - ➢ Users' sensitive information
- ❖ Implemented *LPAuditor*, a composite system that automates these attacks
- ❖ Highlighted the true extent of the privacy risks due to (public) location metadata
- ❖ Provided an extensive, comparative evaluation to prior approaches
- ❖ Revealed and studied the impact of Twitter's past invasive policy

Thank you!

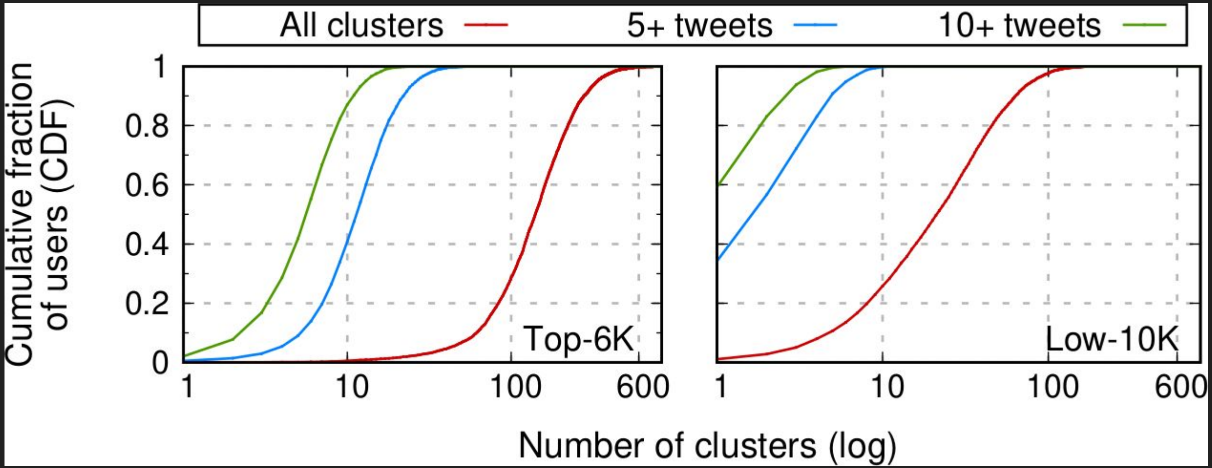
https://www.cs.uic.edu/~location-inference/

kostasdrk@ics.forth.gr

Contributions

- Techniques for inferring user home & work locations
  - High accuracy
  - Fine granularity (postal address)
- Novel approaches for inferring sensitive user information
- Design *LPAuditor*, a system that automates the attacks
- Investigate Twitter's past invasive policy and how it impacts users

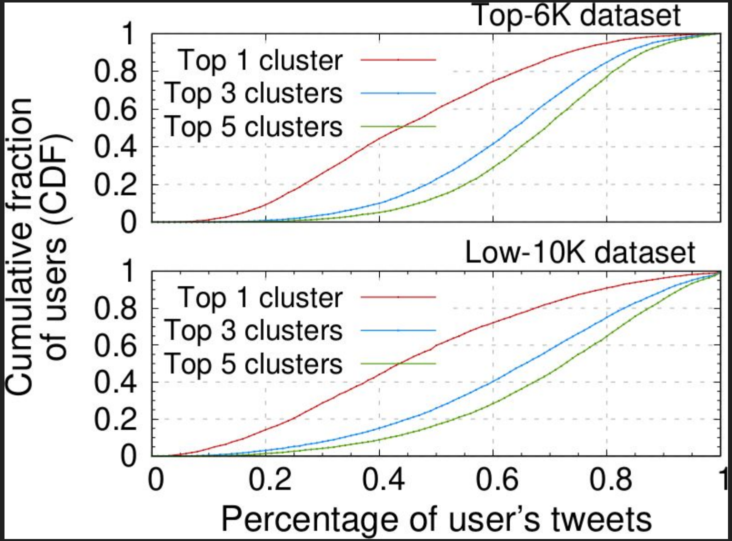
Number of Clusters



- ~28% have less than 100 clusters
- 50% have more than 140 clusters

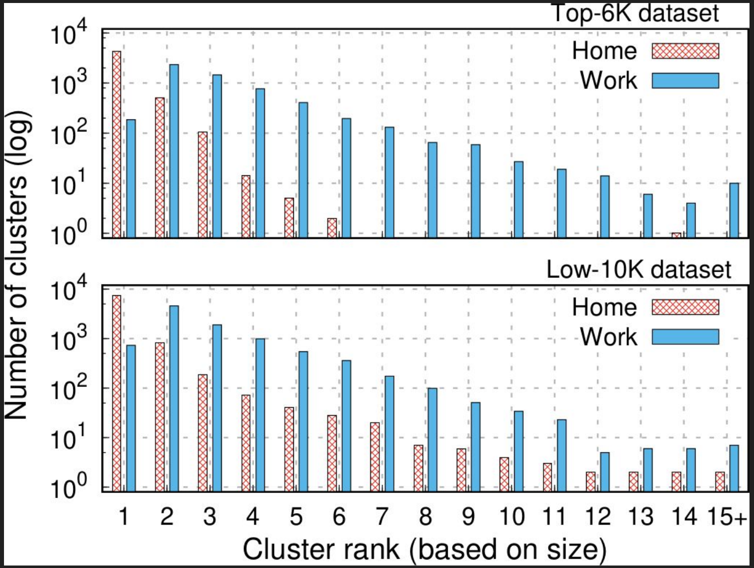
- ~11% have less than 6 clusters
- 50% have more than 21 clusters

Tweets from Top Clusters



- ❖ ~40% of the users, have more than half of their tweets in the top cluster
- ❖ ~48% have more than 70% of their tweets in their top 5 clusters

Key Location Inference - Main Datasets

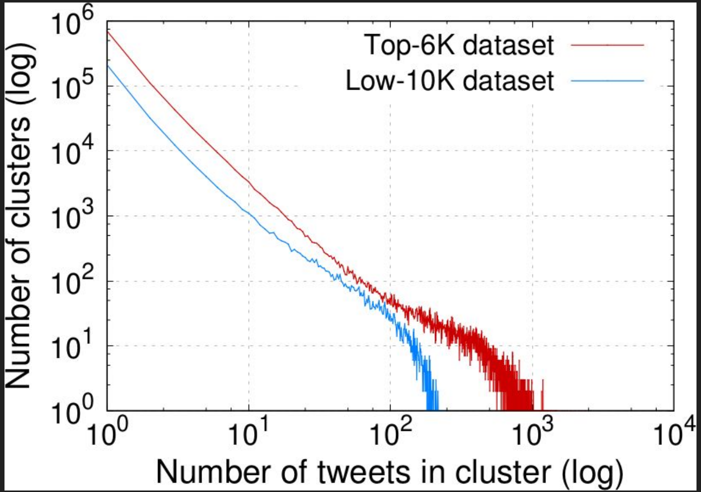


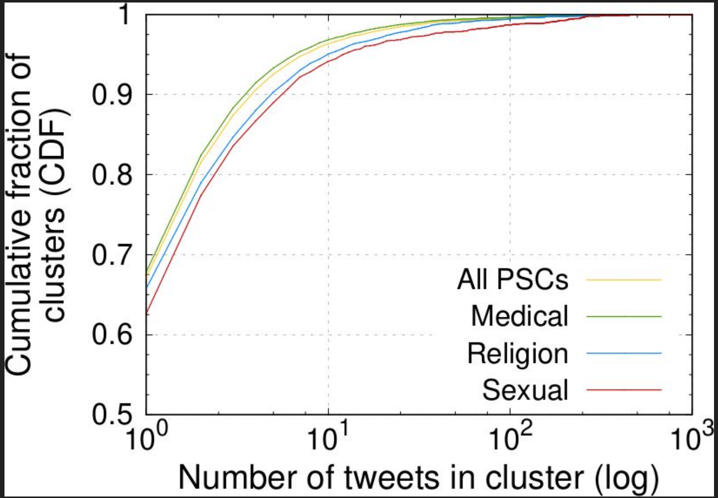
- ❖ The inferred clusters' rank distribution matches our groundtruth evaluation

Comparison to Prior Work - Analytics

		Heuristic Description	Dataset Top	Dataset Low	Proposed by
Home	1	Cluster with the highest number of tweets	72.3%	67.8%	[19], [20], [34], [39]
	2	Most tweets between 20:00-8:00	72.1%	66.4%	[45]
	3	Most tweets between 24:00-7:00	69.3%	54.7%	[34]
	4	Last destination of the day (before 3am)	73.3%	64.8%	[34], [39]
	5	Last destination of the day (w/o days with tweets between 24:00-7:00)	71.4%	64.4%	[34]
	6	Weighted PageRank for destinations	44.1%	26.4%	[34]
	7	Weighted PageRank for origins	37.5%	20.9%	[34]
	8	Most popular cluster in terms of unique days, during the *Rest* (2:00-7:59) and *Leisure* (19:00-01:59) time frames	73.1%	64.9%	[25]
	9	WMFV (best reported time frame: 24:00-5:59)	65%	50.9%	[43]
	10	W-MEAN (best reported time frame: 24:00-5:59)	0.6%	14.7%	[43]
	11	W-MEDIAN (best reported time frame: 23:00-5:59)	15.6%	24.5%	[43]
	12	LPAuditor's Home detection without 2^{nd} level clustering	73.7%	69.3%	this paper
	13	**LPAuditor's Home detection**	**92.2%**	**92.9%**	**this paper**
Work	14	Most popular cluster in terms of unique days, during the *Active* time frame (e.g., working hours, 08:00-18:59)	33.2%	48.9%	[25]
	15	Cluster with the second highest number of tweets	18.5%	22.8%	-
	16	LPAuditor's Work detection without 2^{nd} level clustering	32.2%	30.4%	this paper
	17	**LPAuditor's Work detection**	**55%**	**57.6%**	**this paper**

Clusters’ Size





- Power-law distribution
- Smaller clusters are important from a privacy perspective

- ~67% of PSCs have a single tweet
- Only ~4% have 10 or more

Content-Based Corroboration - Analytics

	Home-Top	Home-Low	Total
Users in Dataset	1,004	1,043	2,047
PSCs	5,393	1,090	6,483
Users w/ PSCs	938	516	1,454
Guessed Clusters (CB)	464	81	545
Users w/ CB Clusters	328	72	400
True Positive (TP)	368	70	438
False Positive (FP)	96	11	107
False Negative (FN)	25	4	29
Precision (TP/TP+FP)	79.31%	86.41%	80.36%
Recall (TP/TP+FN)	93.63%	94.59%	93.79%
F-Score	85.87%	90.31%	86.55%

kostasdrk@ics.forth.gr

Duration-Based Corroboration - Analytics

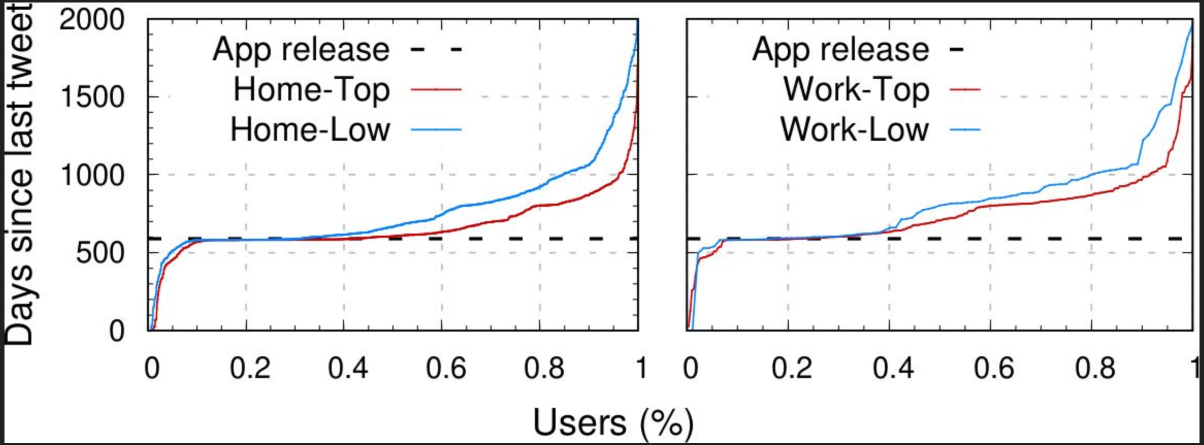
	Home-Top	Home-Low	Top-6K	Low-10K
Visited Clusters (DB)	1,699	276	7,020	2,337
• Medical	1,307	194	5,193	1,626
• Religion	245	56	1,176	493
• Sex/nightlife	147	26	651	218
Users w/ DB Clusters	691	205	3,012	1,672
Common CB/DB Clusters	53.44%	44.44%	53.9%	47.25%
Users w/ CB/DB Clusters	86.89%	59.72%	86.26%	65.88%

User Behavior Through Time

Dataset	Before 4/2015	After 4/2015
All tweets	24.98%	1.35%
Coarse-grained tweets	99.9%	2.85%

❖ 35-fold reduction in geotagged tweets

Impact of Historical Data

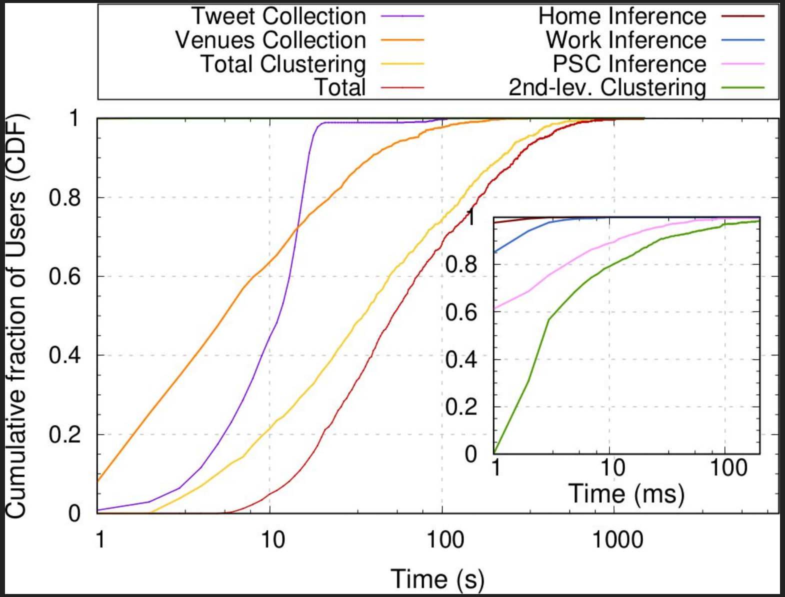


- ~56% and ~68% posted last from home right before the release dates
- Few users kept posting geotagged tweets afterwards

Performance Evaluation

- Randomly selected 1k users
- Tweet collection in less than 20s for 98% of users
- Venue collection up to 6s for half the users
- Clustering up to 35s for half the users

- Total time
  - Less than 52s for half the users
  - 95% of users can be processed within 6 minutes



Future work

- Tune our approaches on areas with different societal and legislative norms
- Apply on different data sources (e.g. wearables)
- Investigate differences in rural vs urban areas
- Explore the more recent *POI* tag and how it can be exploited to infer sensitive user information

kostasdrk@ics.forth.gr