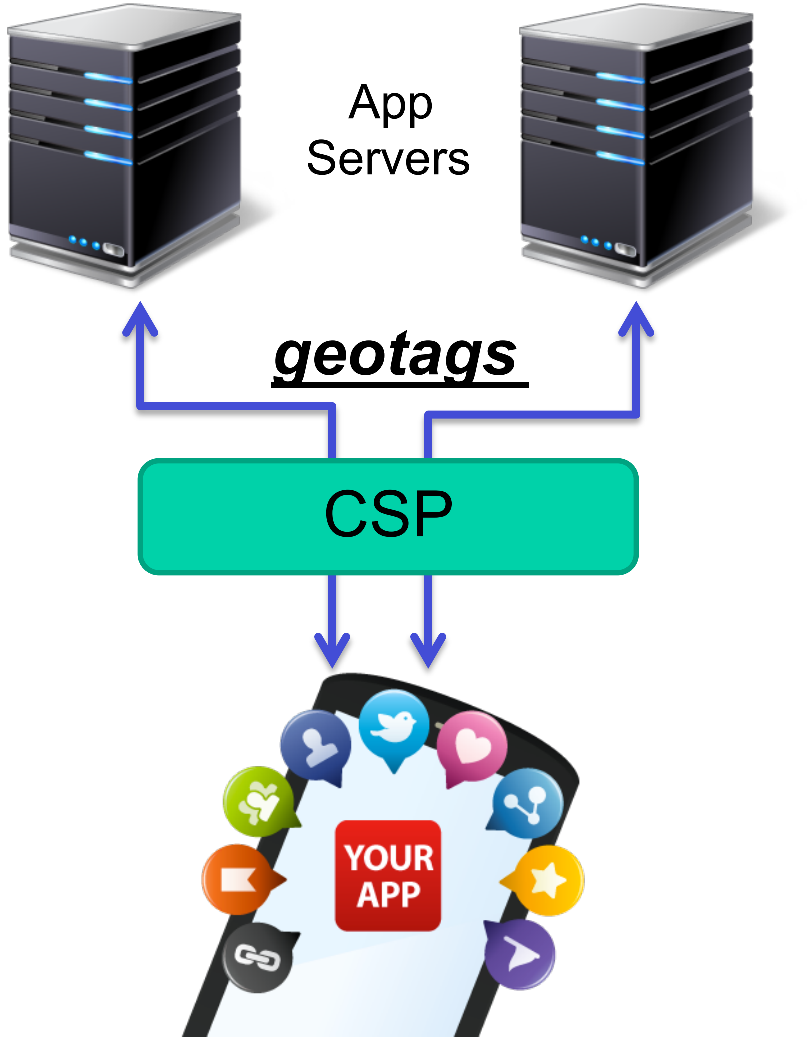
GeoEcho: Inferring User Interests from Geotag Reports in Network Traffic

Ning Xia (Northwestern University)
Stanislav Miskovic (Narus Inc.)
Mario Baldi (Narus Inc.)
Aleksandar Kuzmanovic
(Northwestern University)
Antonio Nucci (Narus Inc.)

Background



Geotag: lat/long pair

Host	HTTP requests
www.google.com	...S&**ll**=44.xxxxxx, -69.xxxxxx&…
api.twitter.com	...**lat**=39.xxxxxxx& **long**=-91.xxxxxx...
a.medialytics.com	...&**lat**=33.xx&**lon**= -78.xx&d=HTC+…

Each application has its own geotags

Motivation

- Can we collect all geotags for a single user across applications?
- What do the geotags we see actually mean?
- What can we learn about each user from their reported geogags?

- CSP can see all geotags from different applications for the same user
- A large volume of geotags can be captured from user traffic, but not all of them are user locations
- From user locations, we can learn users' real-world activities

Motivation (Cont.)

GeoEcho is designed to:

- Be fully passive and service-agnostic
- Learn users' *real-world* interests from geotags
- Be utilized by traffic observers such as CSPs
- Enable better personalized services

GeoEcho analyzes user geotags
to connect user online traffic to offline activities,
which will enable CSPs to provide better services

Dataset

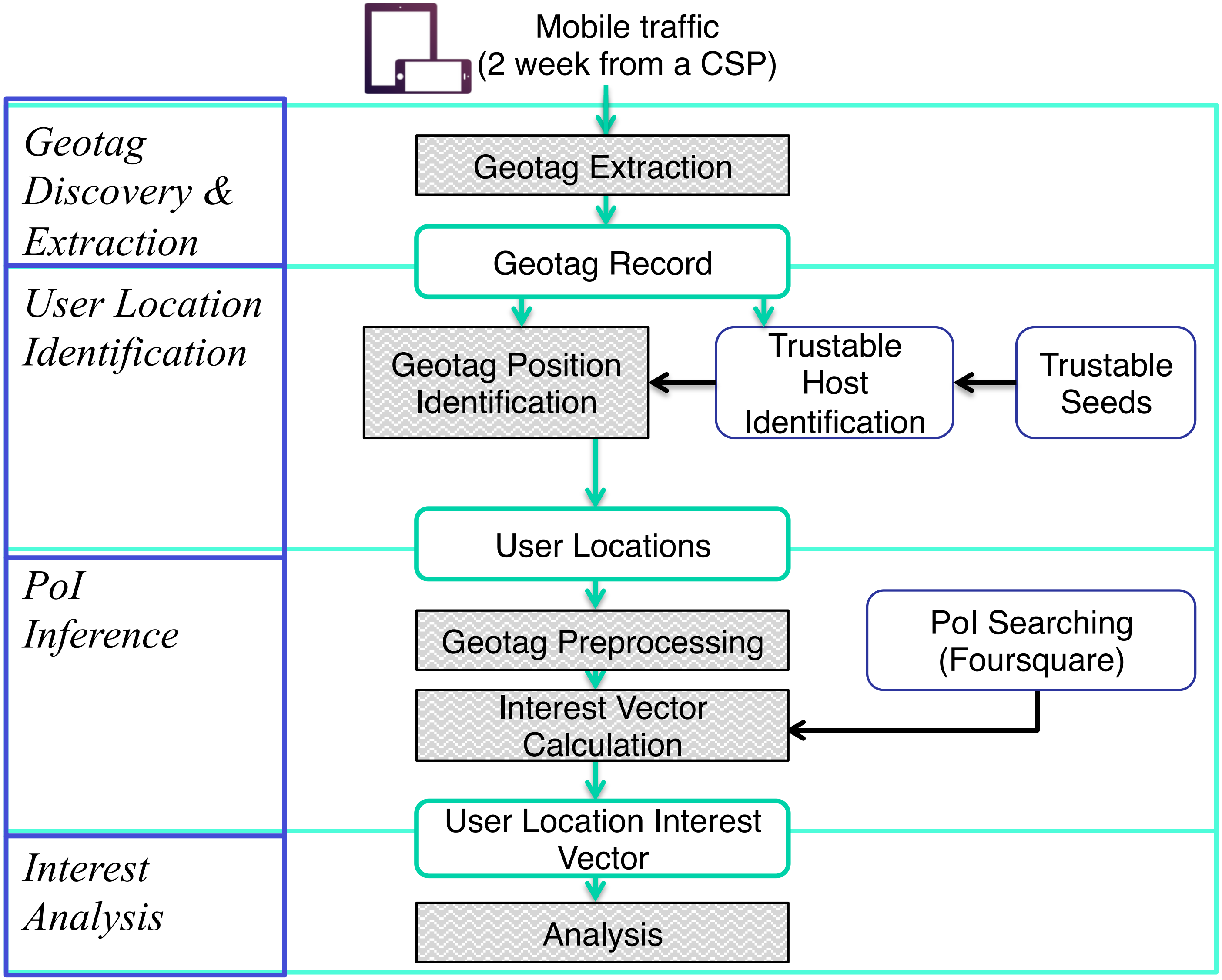
- Summary of datasets

Trace duration	2 weeks in summer 2012
Location	United States
Total user number	608,788
HTTP sessions with geotag	27,981,407
Base stations with known Coordinate	3,415

- Point of Interest (PoI)
  - Used to present user interests
  - Information from foursquare API
  - 8 categories and 400 subcategories

PoI Categoreis	# of PoI subcategory	Subcategory examples
Art & entertainment	41	Art gallery, casino…
College & university	38	College gym, college stadium..
food	87	Coffee shop, Chinese restaurant..
Nightlife spots	18	Bar, night club...
Outdoors	46	Beach, ski area…
…	…	…

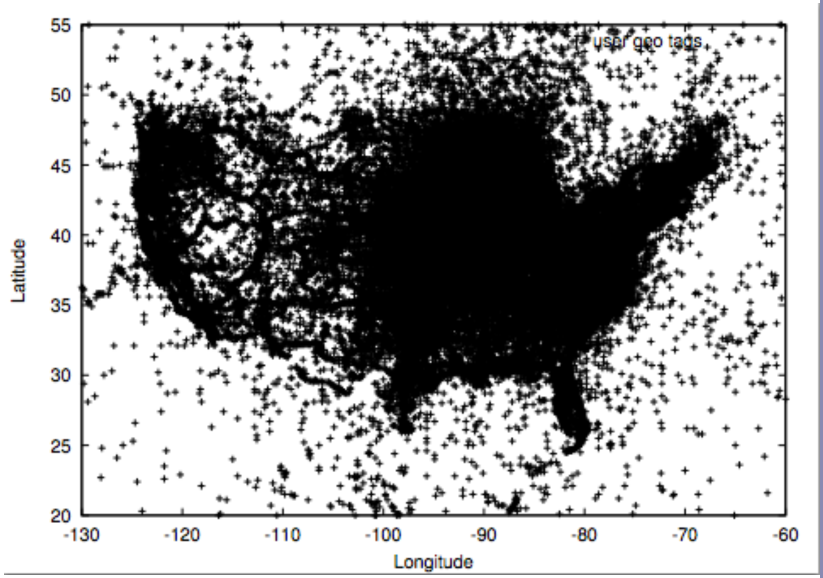
Methodology



Mobile traffic
(2 week from a CSP)
Geotag Discovery & Extraction
Geotag Extraction
Geotag Record
User Location Identification
Geotag Position Identification
Trustable Host Identification
Trustable Seeds
User Locations
PoI Inference
Geotag Preprocessing
PoI Searching (Foursquare)
Interest Vector Calculation
User Location Interest Vector
Interest Analysis
Analysis

Geotag Extraction

- Raw geotag extraction from HTTP requests:
  - 2,500 keyword based geo-signature:
    - Hostname
    - Keywords
    - Regular expression
  - 2,246 individual hosts
  - 27,981,407 geotags from HTTP sessions



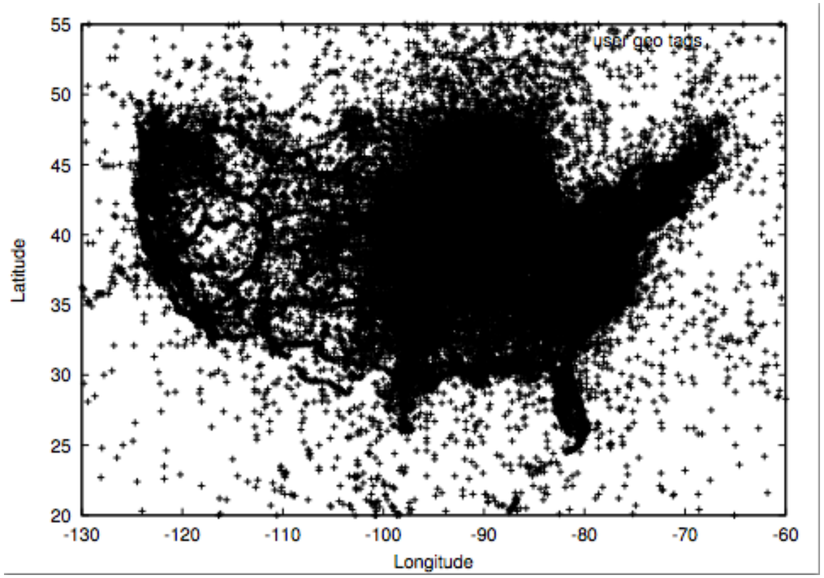
Raw geotags

The extracted geotags may not be user locations.

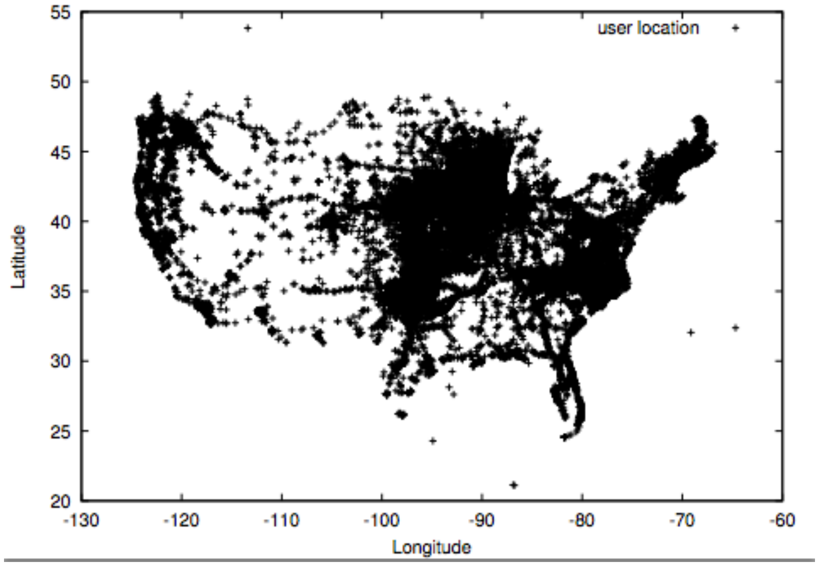
User Location Identification

How to identify user locations from reported geotags?

- Geo-trustable hosts
  - HTTP hostnames that only collect user locations
  - Identified by the nearby base stations



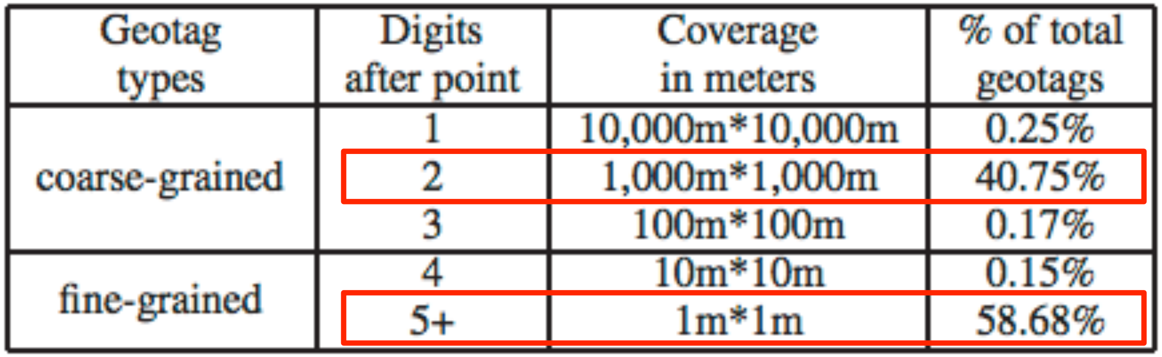
Before location identification



After location identification

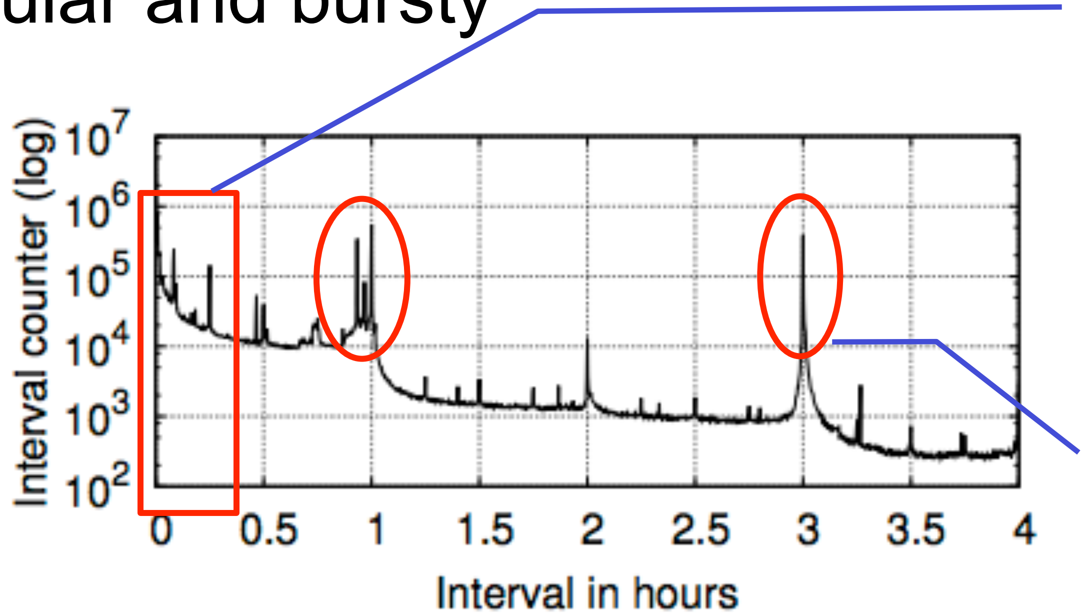
Geotag Characteristics

- Fine-grained or coarse-grained



Geotag types	Digits after point	Coverage in meters	% of total geotags
coarse-grained	1	10,000m*10,000m	0.25%
	2	1,000m*1,000m	40.75%
	3	100m*100m	0.17%
fine-grained	4	10m*10m	0.15%
	5+	1m*1m	58.68%

- Regular and bursty



Bursty because of frequent reposts

Regular geotag reports because of apps like weathers

Inferring User Interests

- User PoI Vector Calculation
  - Geotag Preprocessing:
    - Remove the geotag biases:
      - Temporal aspects
      - Locality aspects
  - Candidate PoI Selection
    - Select nearby PoIs for each geotag
      - Nearer PoIs have better chance

PoI vector calculation formalizes the PoI selection

Inferring User Interests

- Geotag Preprocessing

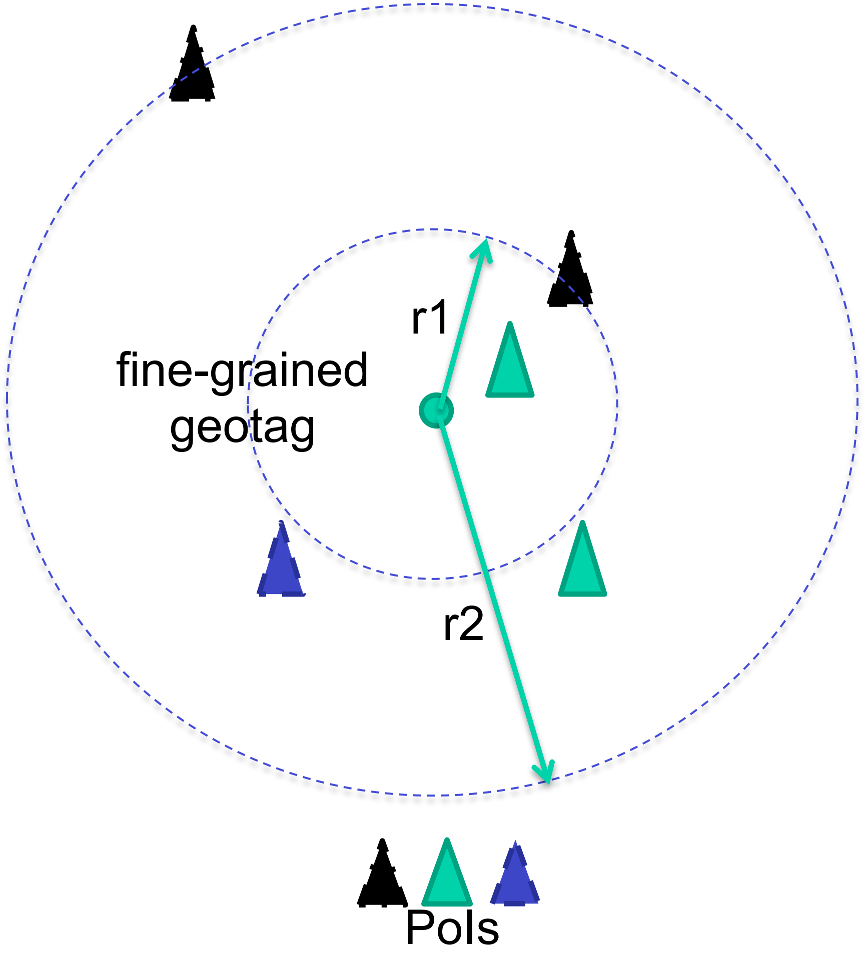
Geotag Biases

- Geotag are not regular in time
- More geotags around home or work place
- Coarse-grained geotags will cover too many PoIs

- Group geotags into hours: the same geotag will be considered once within each hour
- Remove home and work places: 30.7% geotags removed
- Refine coarse-grained geotags: coarse-grained geotags are replaced by inside fine-grained geotags

Inferring User Interests

- Candidate PoI Selection



Fine-grained geotags:

- Different PoI search radii
- r1 (20m) < r2 (50m)

Coarse-grained geotags:

- About 500m*500m coverage
- Consider all covered PoI

All selected PoIs from the same geotag are considered with equal user interest.

Inferring User Interests

- User Interest Vector Calculation
  - Calculate user interest vectors on different time scales (daily, month, etc.)
  - Normalize the selected PoIs into vectors to enable comparison between different different users.

PoI Category	PoI Subcategory	Interest Score
food	coffee_shop	0.05
food	chinese_restaurant	0.15
college	gym	0.25
college	stadium	0.2
college	library	0.3
nightlife	bar	0.05

An example of user interest score

User interest vector calculation formalizes the user interests from the user PoI vector for further analysis/comparison

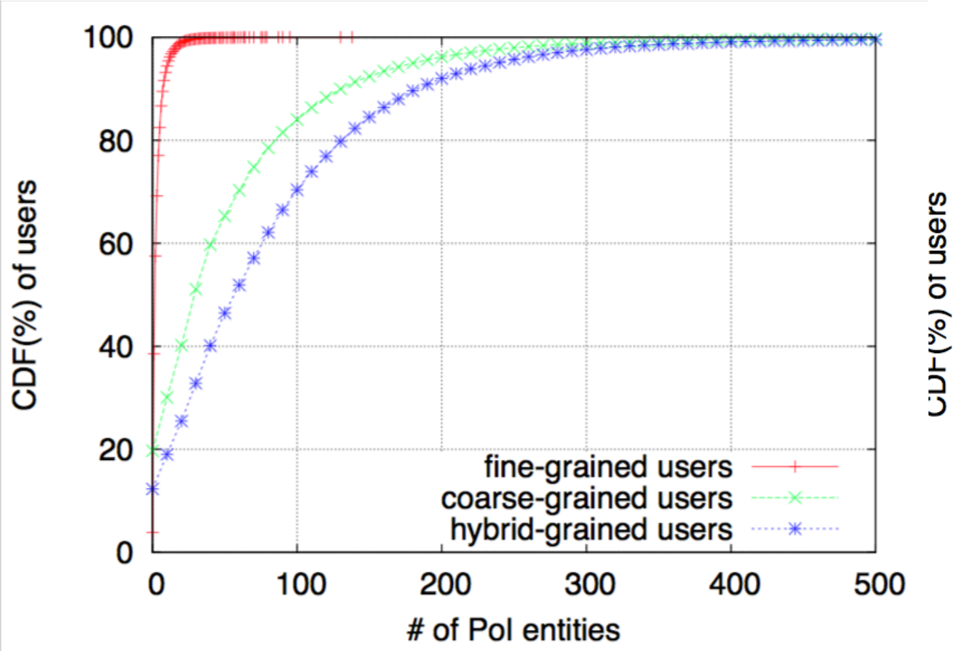
User Interests Analysis

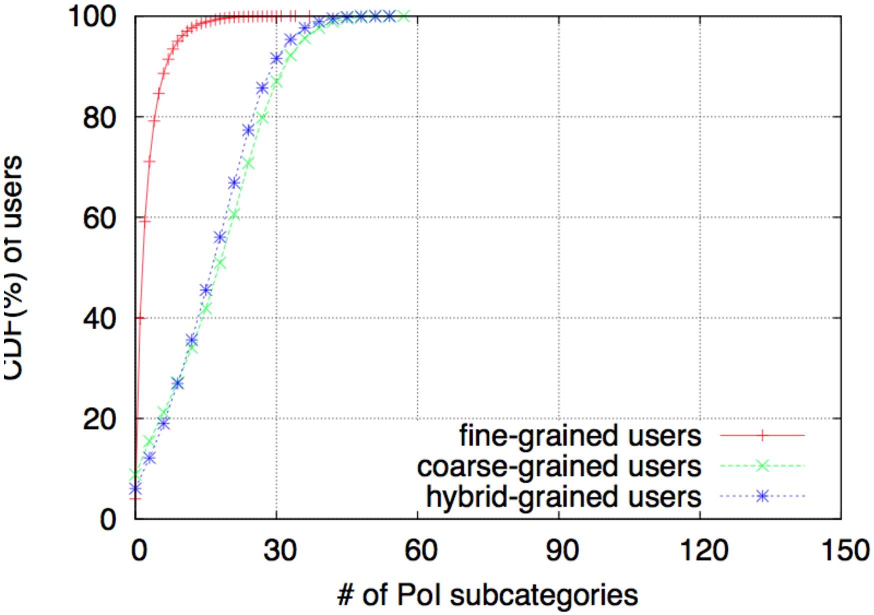
- With User Interest Vectors:
  - Can we learn how many PoIs are interested in?
  - Can we predict user movement by different time?
  - Can we group different users with similar interests?

With user interest vectors, traffic observes such as CSPs can learn many details of end users and are possible to provide better services like recommendations and advertising

User Interests Analysis

- User Interest Vectors:
  - PoIs can be used to present user real-world interests

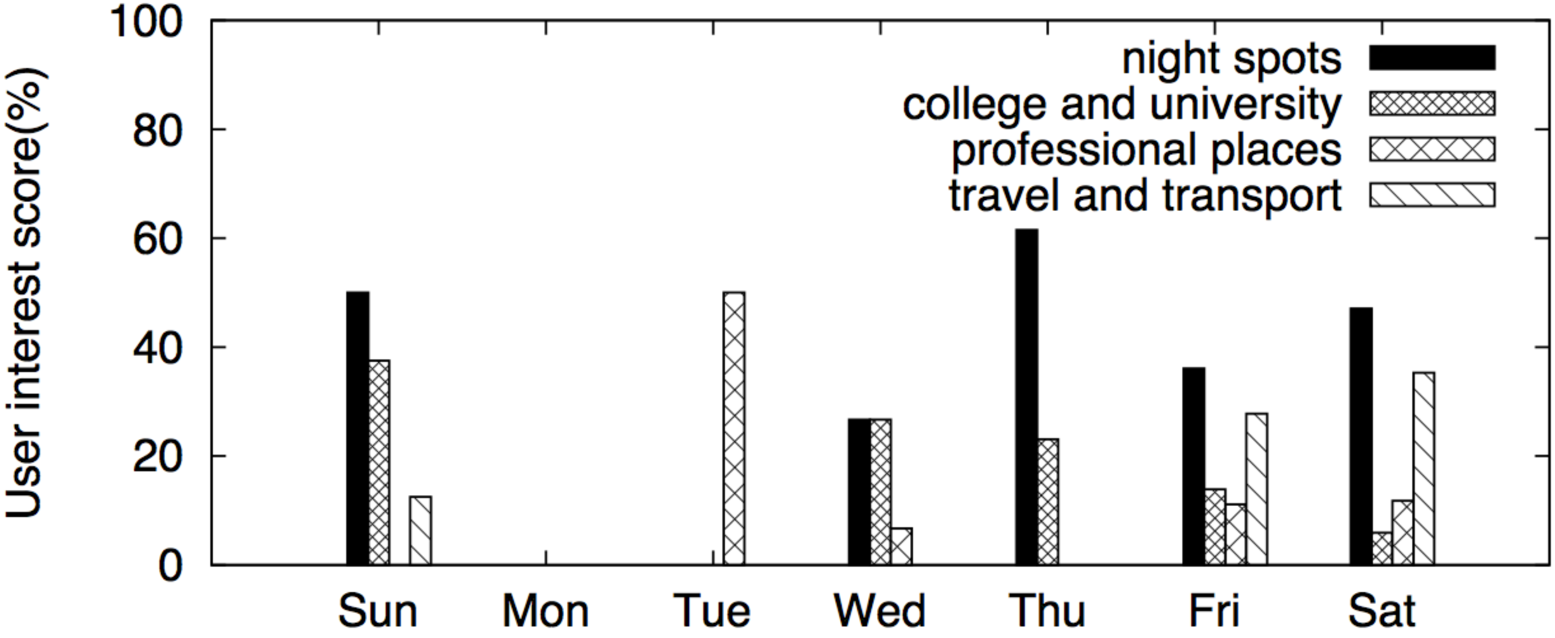




The cardinality of user interest vectors is small (among 400 of them)

User Interests Analysis

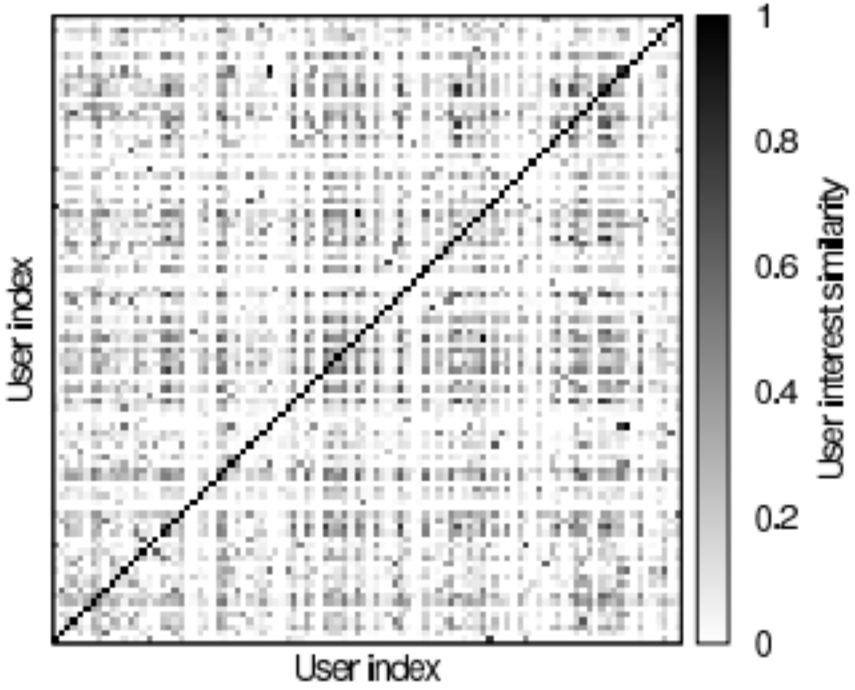
- User Interest Patterns:



User interest vector can be calculated on different time duration (daily/monthly/yearly) to learn user interest patterns

User Interests Analysis

- User Interest Uniqueness



Similarity of PoI interests from 100 random users

The user interest vectors are largely unique

Summary and Conclusions

- Methodology:
  - Extract user coordinates to get user locations
  - Define and calculate user interest vectors
  - Connect online traffic to *offline physical activities*
- Geotag characteristics
  - Noisy, irregular and bursty
- User interests:
  - Cardinality is small
  - User interests are largely unique

GeoEcho will generate formalized user interest vectors, which can be calculated on different time duration.

CSPs can use such interest vectors to provide better personalized services, such as advertising, recommendation, etc.

GeoEcho: Inferring User Interests from Geotag Reports in Network Traffic

Thanks!