GeoEcho: Inferring User Interests from Geotag Reports in Network Traffic

Ning Xia*, Stanislav Miskovic†, Mario Baldi†, Aleksandar Kuzmanovic*, Antonio Nucci†
*Northwestern University
ningxia2015@u.northwestern.edu, akuzma@cs.northwestern.edu
†Narus Inc.
{smiskovic, mbaldi, anucci}@narus.com

***Abstract*—Being transmitted as part of numerous Internet services, geo location data is increasingly bringing hints of people's real-world activities into Internet traffic. This paper focuses on the discovery of key properties that motivate personal activities - locational interests. We propose and design GeoEcho, a mobile traffic analysis system that extracts and analyses a wealth of latitude-longitude geotag reports with the purpose of identifying the points of interest (PoI) which people actually visit. The key challenge in such identification is that geotag reports are commonly sent arbitrarily, sparsely and without a sufficient accuracy to uniquely identify any PoI. In our analysis of a two-week trace from a large North-American cellphone operator, we show that 22% of geo reports do not even represent actual people's positions, while another 45% of the reports have low accuracy, such that they ambiguously indicate a number of potential PoIs. We devise methods that effectively identify and prune irrelevant geo information and infer personal interests of individuals. Thereby creating representative profiles of personal interests, our key results reveal that users show interest in a limited number of topics, and their interests are largely unique and stable over time. Our analysis shows a significant GeoEcho usability in various contexts ranging from generic user profile and user group analysis, to advertising and security applications.**

***Keywords*-coordinates; location; point of interest;**

I. Introduction

Location-based services are thriving in the Internet. As a result, exchange of geo-location information has become common for Internet users. Looking for whereabouts of friends, searching for nearby points of interests (PoIs) or simply checking weather, people increasingly leave geo-location footprints (latitude and longitude pairs, or "geotags") in the Internet. These footprints are a window to people's real-world activities and as such attract the significant attention of location-based service providers. In addition, such geo footprints open the door towards understanding people's real-world behavior and interests.

In this paper, we design and evaluate GeoEcho, a system that systematically extracts and purifies arbitrary geo data found in Internet traffic and utilizes such data to semantically characterize user interests in the physical world. Because GeoEcho extracts information directly from the Internet traffic, it has a much broader view in a wealth of online services, many of which regularly and frequently send user geo data. Thus, contrary to service-limited data that may work in the context of a single service or application, GeoEcho gathers data from all such services, enables us to observe an unparalleled volume of geotags and to extract user mobility and physical-world activities at unprecedented scale and quality.

At the same time, because GeoEcho is the first one to collect geo data in a completely passive fashion, *i.e.*, outside of the application or service context and without any user feedback, it must deal with highly unstructured and unverified data originating from numerous independent services and applications. Necessarily, GeoEcho faces several challenges. First, there is no unified format that could be used to extract geotags from different web services. For example, `api.twitter.com` collects user coordinates in the format of a key-value set `(lat=44.xxxx, long=-78.xxxx)`; while `mapquest.com` collects user locations with a different format (`geo=46.xxx%2c-80.xxx`). Second, an extracted geotag may not correspond to a user's location. By analyzing traffic of about 500,000 smartphone users, we discovered that a significant portion of geotags are related to "noncurrent" locations: People casually browsing remote locations at Google Maps or checking weather for a different city. Third, for user locations, the reports can have a coarse-grained accuracy. Fourth, even when the reports are fine-grained, no PoIs may exist around reported locations.

To overcome the non-current location problem, we perform a de-noising processing of the extracted geotags. In particular, we extract geotags from a keyword-based regular expression matching process as the first step of identifying user location reports. Then we devise a concept of reliable services that predominately report actual user locations and use these services to grow the set of relevant geotags. Our criteria for growing the set are based on temporal and physical proximity. We show that the proposed de-noising methodology filters 22% of geotags as irrelevant.

Next, to effectively characterize a user's interests in the physical world from the identified location geotags, we devise the *interest vector* concept. In particular, an interest vector for a user consists of statistically significant semantic features (categories and subcategories) associated with PoIs that a user encounters over longer time scales. Although GeoEcho necessarily takes coordinates as input, we intentionally use the general semantic features, not PoIs themselves, to produce a generic *location-independent privacy-preserving* user representation. We demonstrate that the interest vector is a powerful concept that comprehensively summarizes a user's mobility prop-

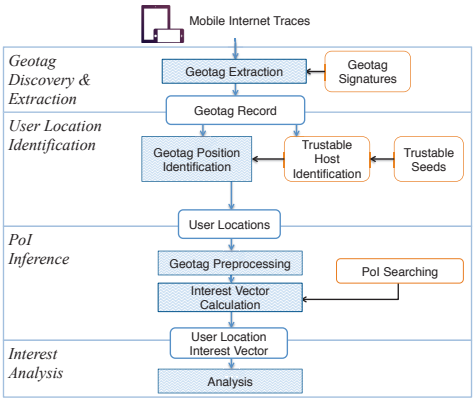


Figure 1. GeoEcho methodology overview.

erties and interests in the physical world.

As GeoEcho is designed to be fully passive and service-agnostic, the system cannot rely on feedbacks from a single service to establish explicit validation ground truth. In order to evaluate the calculated user interest vectors, we rely on geotags themselves. From the geotags, we select a ground truth data subset to build "baseline" interest vectors, then compare to the "generic" vectors of the corresponding respective users. Our results provide a strong validation of GeoEcho's indications, showing that interest vectors indeed do reveal actual user interests.

With the analysis from GeoEcho system, we have following results. (i) The cardinality of user interest vectors is small, *i.e.*, a user on average shows interest in less than 5% of all (sub)categories. (ii) A user's interest vectors is largely unique, *i.e.*, different users have very different interest vectors. (iii) A user's interest vector quickly converges towards a stable value. All these properties imply significant interest vector usability in various contexts ranging from generic user and user group analysis, to advertising and security applications.

II. GeoEcho Methodology

GeoEcho is capable of not only extracting reports of latitude-longitude geotags from network traffic, but also identifying which ones refer to locations actually visited by users and ultimately inferring from them the interests of users. This section describes GeoEcho's methodology and design as illustrated in Figure 1: (i) discovery of geotags in Internet traffic, (ii) tracking actual user locations in noisy geotag reports, (iii) identifying relevant points of interests (PoI) that users do visit, and (iv) generalizing PoI knowledge to personal interests of individual users.

A. Geotag Discovery and Extraction

Numerous web services, especially mobile apps, employ latitude-longitude coordinates (geotags) to offer location-based services. However, given that a standard for the exchange of such tags does not exist, finding them in the Internet traffic is generally challenging. One possible way of locating geotags relies on the fact that they are decimal numbers. However, only a small number of decimal numbers embedded in traffic are actually geo-coordinates and various encoding methods are deployed for them. For example, some services may report geotags as encoded strings `(44.825484%2C20.451279)`, or key-value sets `(latlon=44.825484:20.451279)`, or individual key-value pairs `(l1=44.825484 ... l2=20.451279)`.

To extract geotag reports, we first develop a keyword based process to generate geo-signatures. Geo-signatures are triples with (i) a hostname, (ii) keyword(s) to identify geotags, and (iii) a regular expression to extract coordinates. Here hostnames are identified by their fully qualified domain names found in HTTP HOST fields, instead of just the short and ambiguous names. For example, even though `maps.google.com` and `www.google.com` are both from Google, they are used for different services and they are treated as different host names. We use keywords to identify whether an HTTP request includes geotags, for example, "lat", "latlon", "geolocation", etc. If the keywords exist, regular expressions further explain how to extract a pair of decimal numbers as geo coordinates. With a given geo-signature, only one geotag can be extracted from a single HTTP request if (i) the hostname is matching; (ii) all keywords exist in the HTTP header (cookies, GET parameters); (iii) the regular expression can extract two digital numbers as latitude and longitude.

B. User Location Identification

GeoEcho focuses on actual locations of Internet users. However, not all locations extracted by our geo signatures correspond to places people have actually visited. To single out geotags that enable user localization, we introduce a concept of *geo-trustable hosts* which are known to be the receivers of geotags with actual user locations. Geo-trustable host identification starts from a set of priori known trustable hosts. Then the host set is grown based on the co-occurrence of the reports with the reports of such seed hosts. The initial set of geo-trustable hosts is based on the priori knowledge of the services which only use the user locations, reliable reference point positions (such as base stations), or other possible clues.

We define the *Trust Probability* $P_{tr}(h)$ of a candidate host h as a measure of the probability that h is a trustable host. Given a time window T_{tr}, let N be the number of distinct user terminals sending geotags to both h and trustable hosts and $M \leq N$ be the subset of the N users for whom each location reported to h is the same as the one reported to the corresponding trustable host. Then the trust probability is computed as $P_{tr}(h) = \frac{M}{N} * 100\%$.

$P_{tr}(h)$ can be recomputed at each predefined interval T_{tr}. Once $P_{tr}(h)$ is above a predefined *trust probability threshold* P_{tr}, h is promoted to be a trustable host and geolocation reports sent to it are considered to correspond to the whereabouts of the corresponding user. A shorter predefined interval T_{tr} will lead to smaller N and M, but the probability can be more accurate.

C. Inference of Visited PoIs

The next step for GeoEcho is to infer the corresponding PoI visited by the user. PoI inference is not trivial due to the ambiguity arising from user mobility, the unknown extension of a PoI around its nominal geo-coordinates, and

the often limited accuracy of geotags. First, each geotag has a certain precision which determines the report's uncertainty area that can cover a number of PoIs, *e.g.*, the coarse-grained geotag.Second, the user may only be interested in a few of the covered PoIs or none if he is just passing by. Also, users may not frequently visit the PoIs that near their work or home locations, even with frequent geotag reports. Finally, even if a PoI's nominal coordinates do not overlap with the user's position, the PoI's size might still cover the user's location.

To address the above problems, we devise a PoI scoring criteria (Algorithm 1) such that the scores reflect the likelihood of a PoI being visited. The scores take into account the user's history of reported locations as well as the several properties of geotags (precision, time of reporting, proximity to potential PoIs and the number of PoIs in the coverage area). In Algorithm 1, we define a *User PoI Vector* $P(u) = \{p_i(u)\}$ to keep track of user u's potential visits to the various PoIs. As it will be explained later, $p_i(u)$ is representative of the likelihood of the user u having visited PoI i based on his/her history of geo-location reports. PoIs can be organized in categories and subcategories, in which case some of the elements of the User PoI Vector might represent PoI categories (*e.g.*, entertainment locations) or subcategories (*e.g.*, movie theater), rather than individual PoIs.

Algorithm 1 User PoI Vector calculation

Require: G(Geotags for locations visited by user u), $P(u)$(User PoI Vector)
  $G' \leftarrow PreprocessGeotags(G)$
  for geotag $g_j \in G'$ **do**
    $P \leftarrow CandidatePoISelection(g_j)$
    for each PoI $i \in P$ **do**
      $p_i(u) += 1/sizeof(P)$
    end for
  end for

In the *preprocessing stage* geotags corresponding to the locations visited by the user u are preprocessed in order to avoid bias in over-scoring the user's PoI vector. As we will explain in Section V-A1, the frequency of geotag reports varies greatly; the filtering has the objective to avoid repetitive reports of the same location with high reporting rate. To this purpose, the preprocessing stage returns the set of unique geotags within a time interval T selected based on observation of geotag report intervals from the same user. The preprocessing stage also excludes geotags related to users' residences and work places, because of the possible bias for the unvisited PoIs nearby these locations. The two locations can be identified by either recurring time-of-day patterns of each user [1], [2], or time-spending probability models [3], [4], [5].

The *candidate PoI selection stage* identifies PoI candidates for the reported user location g_j according to an expanding search radius that aims to take into account the accuracy and coverage of geotags. We start to search surrounding PoIs with a smaller radius for each geotag. If no PoI is found, we increase the searching radius to the nearest PoI(s) until reach a threshold. For fine-grained geotags, the threshold will be small; otherwise all the PoI within the geotag's coverage will be considered. The granularity of geotags and PoI search sensitivity will be analyzed in Section V-B.

Notice that: First, each selected PoI around g_j is considered with equal possibility to been visited. This is to increase the likelihood of user u having visited a PoI $p_i(u)$ with more nearby geotag reports. Second, Algorithm 1 can be run on subsets of geotags G for different time durations. For example, the system might build a Monday User PoI Vector, reflecting the likelihood for a user of being at each PoIs on Mondays.

D. Interest Vector Extraction

From a given PoI vector, we further infer the user's interests. For this purpose GeoEcho generates *User Interest Vectors* from each user's PoI Vector. Algorithm 2 shows the pseudo-code for the creation of a User Interest Vector from a User PoI Vector. We first select the elements from PoI vector with a value higher than a predefined threshold to remove rarely visited PoIs, then generate normalized scores for these elements. Each element of an User Interest Vector corresponds to a PoI category or subcategory and the value of the element is the combined likelihood of all the PoIs belonging to the (sub)category that are above the given threshold. The normalization allows the comparison of User Interest Vectors corresponding to different geotag report durations (*e.g.*, daily or weekly) or belonging to different users. *Cosine similarity* [6] is used as a measure of how similar or different the user interests could be.

Algorithm 2 User Interest Vector calculation

Require: $P(u)$(User PoI Vector), L(Likelihood threshold), $I(u)$(User Interest Vector) initialized to 0.
  for each $p_i(u) \in P(u)$ **do**
    if $p_i(u) \geq L$ **then**
      $j \leftarrow (sub)category(i)$
      $I_j(u) += p_i(u)$
    end if
  end for
  $s \leftarrow \sum_j I_j(u)$
  for each $I_j(u) \in I(u)$ **do**
    $I_j(u) = I_j(u)/s$
  end for
  return $I(u)$

III. Experimental Setup

In this section, we present the features of a two-week traffic trace collected on the network of a major cellular operator in North America, and then explain how we devise the PoIs at locations reported in the trace.

The traffic traces used in the experiments reported later in this paper were captured during two different weeks (June 19 to June 25 and July 3 to July 9, 2012) from a major CSP in North America. We leverage the RADIUS [7] protocol infromation to identify all sessions of each anonymous user and associated base station. User privacy is

Table I
STATISTICAL SUMMARY OF THE TRAFFIC TRACE.

Trace duration	2 weeks in Summer 2012
User number with raw geotags	608,788
Total HTTP sessions	1,604,461,319
HTTP sessions contain geotags	27,981,407
Total Base station number	202,545
Base station with known position	3,415

Table II
POINT OF INTEREST CATEGORIES AND SUB-CATEGORIES.

PoI categories	PoI subcategory number	PoI subcategory examples
art & entertainment	41	art gallery, art gallery, casino, comedy club, ...
college & university	38	college stadium, college gym c:college cafeteria, ...
food	87	coffee shop, Indian restaurant donut shop, Chinese restaurant
nightlife spots	18	bar, pub, night club, strip club, ...
great outdoors	46	beach, farm, mountain, ski area...
professional & other places	49	dentist's office, hospital, park doctor's office
shop & services	86	bank, fish market, cloth store, shoe store, bookstore, ...
travel & transport	35	air port, rental car location but station, ...

preserved by anonymizing all user identifiers, such as cellphone numbers, email addresses and IP addresses.

Table I provides a statistical summary of our traffic traces that included traffic from 608,788 users through 202,545 base stations from which geo-coordinate reports could be extracted. Overall, 27,981,407 HTTP sessions (or 1.7% of the 1,604,461,319 total) contain geotag reports.

To further understand user real-world interests from the extracted geotags, we use the Foursquare API [8] to find PoIs that might be at the geotag locations. In particular, Foursquare records the coordinates corresponding to the nominal position of each venue (i.e., a PoI for the user visiting it) with a 10 m accuracy. Each venue in the Foursquare database is associated with a category and a subcategory. Table II shows all the 8 categories and examples of 400 subcategories.

IV. USER LOCATION IDENTIFICATION

In this section, we evaluate the proposed methods for geo-location report extraction and filtering as well as study the properties of obtained location data. Specifically, we show how the proposed approach fares on the dataset introduced in the previous section with respect to two key challenges: (i) identifying geotags in unstructured traffic, and (ii) determining trustable services that persistently reveal physical user locations.

A. Extraction of Geotags

We extract geotags by employing geo-signatures as described in Section II, demonstrate their effectiveness (in terms of number of obtained geotags) and analyze the quality of the devised information. As a demonstration of the breadth of services that deploy geotags in today's Internet, we identify over 2,500 geo-signatures and 27,981,407 geotags embedded in HTTP sessions to 2,246 individual hosts within our two-week traffic trace.

Table III
PROPERTIES OF THE EXTRACTED GEO REPORTS.

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%

An important aspect of geotag quality is their accuracy. Different consumer technologies can pinpoint user location in a radius of a couple meters to hundreds of meters. In the evaluation of our network trace, we learned that Internet services predominantly employ two levels of localization accuracy: **fine-grained geotags** are reported with more than 5 decimal digits, while **coarse-grained geotags** use only 2 decimal-digits. The percentages of the extracted reports for the various accuracy levels are shown in Table III. The fact that about 40% of locations are coarse-grained makes our goals extremely challenging: for a vast number of reports, GeoEcho would have to identify, among tens or hundreds of PoIs falling within the wide are identified by the geo-coordinates, which one(s) the reporting user has actually visited.

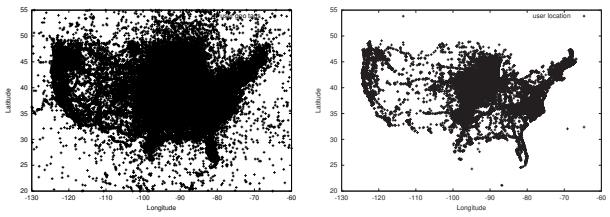
B. Geo-trustable Hosts

Geo-trustable hosts identify user locations from random geotags. As presented in Section II-B, GeoEcho leverages the initial geo-trustable host set and grows it reliably. It is possible to get the initial geo-trustable host set from priori knowledge or experiments for each potential host. To adapt different environments (different countries have different location-based services), here we illustrate how to utilize the positions of corresponding base-stations to identify and evaluate the initial geo-trustable host set.

The fact that the user must be within the coverage area of the associated base-station, enables singling out hosts that should not be included in the geo-trustable set. As the coverage range of a standard base station varies from typically 35km [9] to over 120km [10], we use smaller distances to select initial geo-trustable hosts. In particular, if the 90-th percentile of the distance between the locations reported for a host and the base-station through which the report is received is less than 25 km and the 99-th percentile is less than 90 km, the host is included in the initial set of geo-trustable hosts. With 10 initial geo-trustable hosts, we grow the set of geo-trustable hosts based on the coherence of reports with the ones sent to existing geo-trustable hosts, as discussed in Section II-B. In our data set, 35 hosts were found to be compliant in more than 90% of cases, receiving geotag reports within $10s$ of identical reports to geo-trustable hosts.

C. Effectiveness of Geo-Location Filtering

Having identified the geo-trustable hosts that reliably report user locations, we use this information for geotag filtering. As a result, we trust 21,747,858 geotags to indicate actual user locations, which is about 78% of the original set of geolocation reports. Fig 2 graphically illustrates the effectiveness of GeoEcho filtering methodology.



(a) Full set of geotags (b) User location geotags

Figure 2. Geotags before and after GeoEcho filtering.

Fig 2(a) shows the location corresponding to all geotags found in the traffic trace, which includes a fairly number of spots, for example in the ocean, where there is no network coverage. On the other hand, Fig 2(b) shows that by considering only geotags sent to geo-trustable hosts, GeoEcho selects mostly reports related to areas where the operator's largest user base is located. The remaining points quite accurately identify roaming users in Canada, Mexico and Caribbean islands.

Among 608,788 users present in our network trace, 541,568 (89%) can be localized via the filtered reports sent to trustable hosts. In this filtered set, 48 % of users are **fine-grained users**, i.e., reporting *only* fine-grained geotags, 21% are **coarse-grained users**, i.e., reporting *only* coarse-grained geolocation, and 31% are **hybrid-grained users**. In other words, about 79% of users report at least some highly accurate geo-coordinates.

V. Echo User Interests

In this section, we try to associate the geotags to PoIs by scoring the likelihood of users visiting the PoIs, as well as extrapolate that knowledge to user interests. The key problems we address are (i) scoring bias, (ii) temporal sparseness of geotags and (iii) inaccuracy of coarse-grained geotags.

A. Preparing Geotags for Interest Scoring

Due to the temporal and locality bias, geotags indicating actual user locations cannot still be directly applied to association of users to PoIs or extrapolation of user interests. Here, we correct the bias from geotags for GeoEcho analytics and adjust GeoEcho's parameters to various properties of the underlying dataset.

1) Temporal Aspects: As is most of the Internet traffic, the majority of geotags are inherently bursty. The results in Figure 3 indicate a wide span of inter-arrivals: The fact that the first spike of 4,658,580 intervals within one minute is followed by random pauses (often being long), confirms the problems of reporting burstiness and sparseness. This temporal property of geo reporting introduces two key problems to GeoEcho operation: Receiving geo reports in busts may cause transient and spurious over-scoring of PoIs, while long intervals in-between the reports challenge the estimation of users' PoI visits. To adjust GeoEcho against PoI over-scoring, we normalize the rate of report processing by considering only the reports that are unique during a configurable threshold interval. For example, if a location L would be reported several times during a threshold interval T, it would be considered only once for

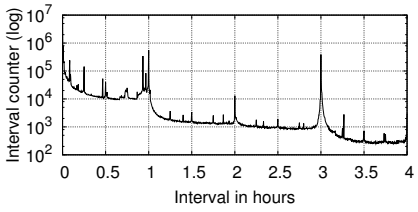
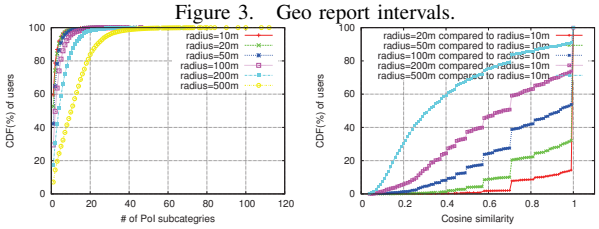


Figure 3. Geo report intervals.



(a) # of PoI subcategories with different search radii. (b) Interest vector cosine similarity with different search radii.

Figure 4. Interest vectors from different PoI search radius.

scoring. Given that the *average* inter-reporting times for most users are on timescales of hours, we set the threshold interval T to 1 hour. Consequently, 10,891,453 out of 21,847,557 reports remain feasible as "scorable" geotags for further analysis.

2) Locality Aspects: Locality of the reported data imposes two key challenges on location-to-PoI associations: (i) Including people's homes and workplaces in PoI scoring would cause huge bias, and (ii) not addressing the problem of coarse-grained reports would hinder resolution of the associations.

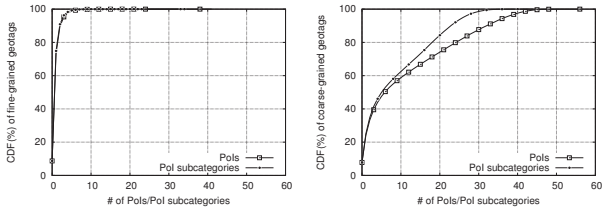
Homes and workplaces. We identify geotags pointing to people's homes and workplaces by employing a simple heuristic which looks for hours-long patterns occurring during the work- and late-night hours [11], [1]. Filtering the areas in which users spend daily most of their time between 11 AM to 4 PM (work) and 1 AM to 5 AM (home), we removed 30.7 % of geotags indicating "homes" and 23.4 % of geotags indicating "workplaces".

Coarse reporting refinement. GeoEcho tackles the problem of coarse-grained information for users by utilizing the concurrent fine-grained geotags. The system replaces coarse-grained reports with the user's previously reported fine-grain locations inside the reported accuracy range. Our statistics shows that the likelihood of error is low, given that more than 85% of cases persistently indicate only one single precise location inside the reported coarse accuracy range, while 98% of cases indicate less than 4 locations inside the range.

B. Sensitivity of Report-to-PoI Associations

Besides temporal and locality challenges, associating reported locations to PoIs may be sensitive to the selection of a PoI search radius. The radius determines the area in which GeoEcho looks for candidate PoIs around the reported locations. Given that users may send geotags while being nearby PoIs, configuring an adequate radius is highly important.

To address this problem, we analyze sensitivity to different chosen search radii of 10, 20, 50, 100, 200, 500 meters. The goal is to understand variations in the number of PoI



(a) # of PoIs/PoI subcategories covered by fine grained geotags. (b) # of PoIs/PoI subcategories covered by coarse grained geotags.

Figure 5. # of PoIs and PoI subcategories within geotag coverage.

subcategories being included in vectors. In Figure 4(a), we show the corresponding cumulative distributions of *subcategory* volumes. The results indicate that GeoEcho does not inflate subcategories significantly as long as the search radius is less than 100 meters. For sure there will be many PoIs located at the densely populated areas like downtown Manhattan. However, within a small region the PoIs tend to be similar types (e.g., restaurants). Then, even with 500 meter search radius, more than 80% of users will only be associated to less than 10 PoI subcategories. This shows that for coarse-grained geotags are still useful to track user interests for PoI (sub)categories.

Next, we compare the difference between interest vectors with different search radii. We use the most precise radius $r = 10m$ as the baseline for comparison. In Figure 4(b), we show cumulative distribution of cosine similarity between the respective vectors. Our results indicate that radii of $r = 20m, 50m$ still produce largely similar vectors as the baseline. Specifically, 85.2% (20 m) and 67.2% (50 m) of users remain having similar interest vectors as the baseline with expansion of the search radius.

Consequently, we configure GeoEcho with two search radii. For fine-grained geotags, we first search for PoIs within 20 meters and if nothing is found we expand the search radius to 50 meters. On the other hand, if no fine-grained information is available, we account for all PoIs within the coarse report's coverage. Applying this configuration to GeoEcho, we obtain the results in Fig 5(a) indicating that for 65.3% of fine-grained location reports the association is injective, one PoI per location. Also, over 99% of locations would have less than 10 PoIs to chose from. The corresponding results for coarse-grained reports are show in Fig 5(b).

C. Validation of Interest Vector Indications

Here, we evaluate whether GeoEcho's interest vectors indeed reveal actual user interests. As one of GeoEcho's key goals is to be fully passive and service-agnostic, the system cannot rely on any feedback from one service for establishment of explicit validation ground truth.

In order to establish the ground truth of actual user interests, we rely on geo data itself and build the following evaluation experiment. We first isolate a subset of data corresponding to unique transition from the reported locations to PoIs to user interests. Specifically, we look at fine-grained reports that have only single PoIs in range and consequently map to single PoI sub-categories, thus unambiguously indicating accurate user interests. This data subset formulates GeoEcho's validation ground truth.

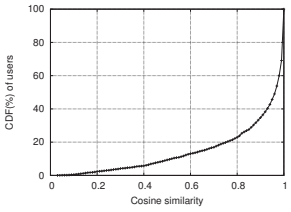


Figure 6. Cosine similarity between GeoEcho's baseline and generic vectors.

Then, we construct the "*baseline*" interest vectors for the users who report geotags from the ground truth data subset. To validate whether GeoEcho indeed points to actual user interest in general, we compare the baseline vectors with *generic* vectors of the corresponding respective users, i.e., the vectors originating from any geo reports the users transmitted. Such generic vectors embody ambiguities, having to score multiple PoIs and PoI subcategories at each visited location.

In our evaluation, we learned that 65.3% of geotags from 67.5% of users satisfy the requirement of forming baseline vectors. To compare baseline and generic GeoEcho's vectors for similarity, Figure 6 indicate that the two types of vectors are almost identical for more than 30 % of users, i.e., the cosine similarity is larger than 0.99. Also, for about 80% of users the similarity remains very high, larger than 0.8. These results provide a strong validation of GeoEcho's indications, showing that interest vectors indeed do reveal actual user interests.

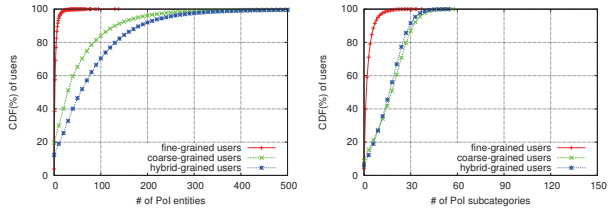
VI. User Interest Analysis

In this section, we explore how different system parameters can affect GeoEcho's indications as well as the various aspects of users interests that GeoEcho can output.

A. Geotag Granularity vs. Interests

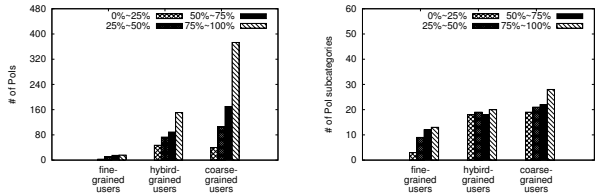
Here we evaluate how geotag granularity can affect interest vectors. Specifically, we try to learn how geotag granularity affects the number of PoIs and PoI subcategories suggested by user interest vectors. Figure 7 shows the cumulative distribution of the number of PoIs and PoI subcategories for the three representative user classes. The statistics show that 90% of fine-grained users have interests in less than 10 PoI entities and 7 PoI subcategories; 80% of hybrid-grained/ coarse-grained users have interests in less than 228/146 PoIs and 27/30 PoI subcategories. Notice that in Fig 7(a), hybrid-grained users have a larger PoI cardinality than the coarse-grained users. The reason is that the two user classes have very different underlying data content and volume of geo information. Thus, the two classes are not feasible an intuitive comparison, meaning that the hybrid-grained class should have more precise reports and consequently point to less PoIs.

The key insights are: (i) Most fine-grained users will only access a small subset of PoIs and PoI subcategories. (ii) Even for the coarse-grained users, who necessarily have a larger interest cardinality, our methodology can reduce the interest profile to about 60 PoI subcategories out of 400 in total. (iii) On average, a user shows interest in less than 5% of the 400 PoI subcategories.



(a) # of user interested PoI entities (b) # of user interested PoI subcategories

Figure 7. Interest vector size with different granularity



(a) # of user interested PoI entities (b) # of user interested subcategories

Figure 8. Interest vectors with different hourly coverage.

B. Geo-Reporting Intervals vs. Interests

Here, we are interested in understanding whether more frequent reports would significantly inflate vector cardinality, *i.e.*., the number of indicated PoIs and PoI subcategories. To evaluate the relationship between the reporting intervals and interest-vector sizes, we separate users into 4 test groups exhibiting different *hourly coverages* of geotags (0%-25%, 25%-50%, 50%- 75%, and 75%-100%). Here *hourly coverages* is referred as the percentage of hourly intervals in which users report locations. Populations of users in each group are: fine-grained user class (255471, 2164, 118, 16), hybrid-grained user class (108547, 6323, 166, 9) and coarse-grained user class (166738, 1967, 44, 5). Necessarily, the higher hourly coverage corresponds to a higher geotag reporting frequency and fewer users who would report the data at corresponding time intervals approaching the hourly coverage continuity. For each group, we calculate average number of PoIs and subcategories, as shown in Fig 8.

Our key insights from Figure 8 are: (i) As expectedly, higher geotag reporting frequency leads to more PoIs and subcategories. Indeed, a larger number of geotag reports will generally point to more visited PoIs. (ii) Unexpectedly, the higher geotag reporting frequency does *not* significantly affect PoI subcategories, as shown in Fig 8(b). The average increase in the number of PoI subcategories for the three classes of users in Figure 8(b) is 1.7, 0.7, and 3.0, respectively. This means that although higher geotag reporting frequency leads to more possible PoI entities, the user interests expressed in terms of PoI subcategories remains almost the same. (iii) The number of PoIs and PoI subcategories is still limited, even for users with highly frequent geotags. In particular, among 400 of PoI subcategories, GeoEcho shows that fine-grained users will only have interests in no more than 10 subcategories. Even for users with coarse-grained geotag reports, the number is limited to 30. In conclusion, a user's interest vector quickly converges towards a stable value. At the same time, the size of the interest vector is generally small.

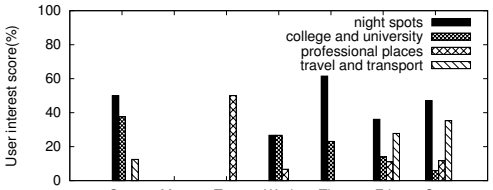


Figure 9. Daily interest patterns for a single observed user.

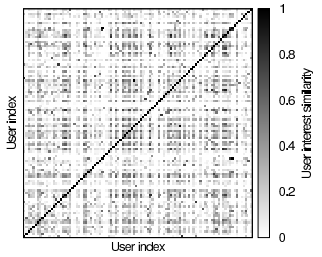


Figure 10. Interest vector comparison in a 100-user population.

C. Interest Profiling

Here, we show several outputs of the GeoEcho system that may be useful in a number of real-world applications, such as advertising, social networking or security.

Single User Interest Patterns. GeoEcho is capable to profile diurnal patterns of interests observed for individuals. In Fig 9, we shows a random user's interests scores with four selected categories. Apart from noticing that the user has an academic affiliation, GeoEcho reveals a number of other interest patterns: (i) The user prefers to travel on weekends. (ii) Apart from being affiliated to an academic institution, the user spends time at another professional institution on Tuesdays, Wednesdays, Fridays and Saturdays. (iii) The user refers from using mobile device on Mondays and Tuesdays at an academic institution.

Profiling Interests of User Groups. We employ GeoEcho to evaluate similarity of interests in user populations. In our experiment, we observe a population of 100 randomly selected users and compare the users' interests learned after processing the two weeks of our dataset. The results shown in Figure 10 compare interests of each user pair. Our key insights are: (i) Although users have a relatively small interest cardinality, the interests are fairly diversified: Individual users' interests are equally likely to exhibit any similarity score when compared to other users. (ii) Identifying groups of users with similar interests (at various levels of similarity) is simple with GeoEcho. The shaded regions in Figure 10 readily indicate such clusters. (iii) It's also easy to spot outliers, i.e., users who share very little interests with general population. Although the users have generally limited interests, our experiment spotted several users who shared absolutely no interests with the rest of the population.

VII. Related Work

A number of authors have been addressing several aspects that are relevant to geo location information. First, there is a wealth of work that leverages specific localization services or the ownership of geo-reporting infrastructure. Shaw et al.[12] tune the Foursquare's algorithms to pinpoint the exact PoIs in order to provide relevant

check-in suggestions to their users. A similar problem was addressed in [13], where the authors owned the mobile devices and software reporting geo locations. Also, leveraging Foursquare and similar services, a number of other authors [14], [15], [11], [16], [17], [18], [19], [20], [21], [22] addressed various problems of determining the exact places at which users spend most of their time, predicting PoI visits, inferring peoples' social interactions or modeling human mobility. Such research only gleans into a comparatively small subset of geo data available to GeoEcho in today's Internet. Another important difference is that such research already had geo information (and corresponding exact PoIs) available from the underlying services. Also, GeoEcho shows that a much more insightful dataset can be obtained by learning the semantics of geotags from a wealth of Internet services and knowing how to filter actual user locations.

A major concern for any research related to user locations is privacy. The authors of [1], [3], [4] showed that having hold of even anonymized call record details enables inference of user identities as well as the users' most frequently visited locations. Moreover, the authors of [23], [5] shows that having even a vague knowledge of person's home or work addresses can identify the person and personal beliefs, preferences and behavioral aspects. While the threat of misusing any data produced by GeoEcho is inevitable, we try to anonymize our output. Specifically, our output could be just anonymous interest vectors of people in a certain geographic area. On the other hand, GeoEcho does not contribute to leakage of private information, given that it only extracts data from the existing internet traffic exchanged in clear text.

VIII. Summary and Conclusions

In this paper, we designed and evaluated the first completely passive system that extracts knowledge of users' interests in the physical-world from the generic and highly unstructured Internet traffic.

Methodology. We introduce GeoEcho, a traffic analysis system that extracts and analyses a wealth of latitude-longitude geotag reports with the purpose of identifying the points of interest which people are interested in. Specifically, GeoEcho extracts and de-noises coordinates to get user positions, and then calculate user interest vectors from surrounding PoIs.

Characteristics of Geotags. From GeoEcho, we analyze the characteristics of user geotag reports. Our insights are following: (i) User coordinate reports are noisy, as 22% raw geotags do not represent actual people's positions. (ii) Although users may have bursty reports and regular reports, geotag reports from most users are still sparse. (iii) Our experiments show that various scoring and vector creation methods can reliably reduce the identification ambiguity to one PoI in 65.3% of fine-grained cases.

User interest vectors. We evaluate the accuracy of user interest vectors with selected geotags as ground truth. Our further analysis shows: (i) the cardinality of user interest vectors is small, (ii) the interest vectors are largely unique, (iii) interest vectors are consistent and stable over time, can quickly converge towards a stable value.

References

[1] H. Zang and J. Bolot, "Anonymization of Location Data Does Not Work A Large-Scale Measurement Study," in *MOBICOM '11*.

[2] I. Trestian, S. Ranjan, A. Kuzmanovic, and A. Nucci, "Measuring serendipity: connecting people, locations and interests in a mobile 3G network," in *IMC '09*.

[3] S. Isaacman, R. Becher, R. Caceres, S. Kobourov, M. Martonosi, J. Rowland, and A. Varshavsky, "Identifying important places in people's lives from cellular network data," in *Pervasive '11*.

[4] S. Isaacman, R. Becker, R. Cáceres, M. Martonosi, J. Rowland, A. Varshavsky, and W. Willinger, "Human mobility modeling at metropolitan scales," in *MobiSys '12*.

[5] P. Golle and K. Partridge, "On the Anonymity of Home/Work Location Pairs," in *Pervasive '09*.

[6] "Cosine similarity." http://en.wikipedia.org/wiki/Cosine_similarity/.

[7] C. Rigney, S. Willens, A. Rubens, and W. Simpson, "Remote Authentication Dial In User Service (RADIUS)," Jun. 2000, Internet RFC 2865.

[8] "Foursquare venues api." https://developer.foursquare.com/overview/venues.

[9] "Femtocell." http://en.wikipedia.org/wiki/Femtocell.

[10] "Channel card for extending coverage area of base station." http://www.google.com/patents/US6647266.

[11] E. Cho, S. A. Myers, and J. Leskovec, "Friendship and Mobility: User Movement in Location-Based Social Networks," in *SIGKDD '11*.

[12] B. Shaw, J. Shea, S. Sinha, and A. Hogue, "Learning to rank for spatiotemporal search," in *WSDM '13*.

[13] Y. Zheng, Q. Li, Y. Chen, X. Xie, and W.-Y. Ma, "Understanding Mobility Based on GPS Data," in *the 10th International Conference on Ubiquitous Computing '08*.

[14] Z. Cheng, J. Caverlee, K. Lee, and D. Z. Sui, "Exploring millions of footprints in location sharing services," in *AAAI ICWSM '11*.

[15] N. Eagle and A. Sandy Pentland, "Reality mining: sensing complex social systems," *Personal and Ubiquitous Computing*, vol. 10, no. 4, pp. 255–268, Nov. 2005.

[16] A. Sadilek, H. Kautz, and J. P. Bigham, "Finding your friends and following them to where you are," in *WSDM '12*.

[17] A. Noulas, S. Scellato, and N. Lathia, "Mining User Mobility Features for Next Place Prediction in Location-based Services," in *Data Mining (ICDM) '12*.

[18] S. Scellato, A. Noulas, and C. Mascolo, "Exploiting place features in link prediction on location-based social networks," in *KDD '11*.

[19] T. H. N. Vu, K. H. Ryu, and N. Park, "A method for predicting future location of mobile user for location-based services system," *Computers & Industrial Engineering*, vol. 57, no. 1, pp. 91–105, Aug. 2009.

[20] I. Trestian, K. Huguenin, S. Ling, and A. Kuzmanovic, "Understanding human movement semantics: a point of interest based approach," in *WWW '12*.

[21] Y. Zheng, L. Zhang, Z. Ma, X. Xie, and W. Ma, "Recommending friends and locations based on individual location history," *ACM Transaction on the Web*, 2011.

[22] Y. Zheng and X. Xie, " Learning travel recommendations from user-generated GPS traces," *ACM Transaction on Intelligent Systems and Technology*, 2011.

[23] S. B. Wicker, "A little too smart - The loss of location privacy in the cellular age," 2011.