PEIR, the Personal Environmental Impact Report, as a Platform for Participatory Sensing Systems Research

Min Mun, Sasank Reddy, Katie Shilton, Nathan Yau, Jeff Burke,
Deborah Estrin, Mark Hansen, Eric Howard, Ruth West, Péter Boda*
Center for Embedded Networked Sensing
University of Caliornia, Los Angeles
{bobbymun,destrin}@cs.ucla.edu,sasank@ee.ucla.edu,kshilton@ucla.edu,
{nyau,cocteau}@stat.ucla.edu, jburke@remap.ucla.edu,{ejhoward,rwest}@cens.ucla.edu
Nokia Research Center, Palo Alto*
peter.boda@nokia.com

ABSTRACT

PEIR, the Personal Environmental Impact Report, is a participatory sensing application that uses location data sampled from everyday mobile phones to calculate personalized estimates of environmental impact and exposure. It is an example of an important class of emerging mobile systems that combine the distributed processing capacity of the web with the personal reach of mobile technology. This paper documents and evaluates the running PEIR system, which includes mobile handset based GPS location data collection, and server-side processing stages such as HMM-based activity classification (to determine transportation mode); automatic location data segmentation into "trips"; lookup of traffic, weather, and other context data needed by the models; and environmental impact and exposure calculation using efficient implementations of established models. Additionally, we describe the user interface components of PEIR and present usage statistics from a two month snapshot of system use. The paper also outlines new algorithmic components developed based on experience with the system and undergoing testing for integration into PEIR, including: new map-matching and GSM-augmented activity classification techniques, and a selective hiding mechanism that generates believable proxy traces for times a user does not want their real location revealed.

Categories and Subject Descriptors

H.4.2 [**Information Systems**]: Information Systems Appliations—*types of systems, decision support*; H.5.2 [**Information Systems**]: Information Interfaces and Presentation—*User Interfaces*

General Terms

Design, Performance, Standardization

MobiSys'09, June 22–25, 2009, Kraków, Poland.

Keywords

Participatory Sensing, Location Data, Environmental Impact and Exposure, Mobile System

1. INTRODUCTION

Participatory sensing refers to the vision of distributed data collection and analysis at the personal, urban, and global scale, in which participants make key decisions about what, where, and when to sense [1]. The infrastructure formed by an installed base of well over two billion mobile phones, when combined with a cloud of supporting web services, make such adaptive, mobile, human-in-the-loop sensing systems possible. However, the existence of these capabilities does not make them immediately usable or scalable. To reveal relevant, previously unobservable implications of human activity requires the development of new types of integrative platforms. Here, we present PEIR, the Personal Environmental Impact Report (Figure 1), an example of this new class of system: It uses mobile handsets to collect and automatically upload data to server-side models that generate web-based output for each participant. This paper describes the PEIR system as originally designed and implemented, evaluates algorithmic developments motivated by its use, and outlines future work.

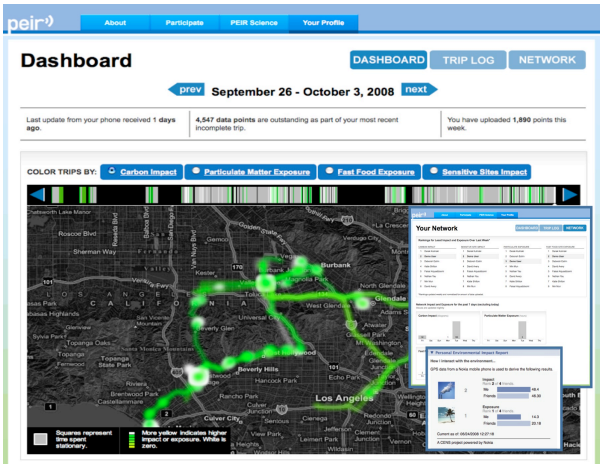

Figure 1: PEIR user interface

Personalization of Environmental Data and Models: Inspired by the Environmental Impact Reports (EIRs) required for construction and public works projects, PEIR was designed to bring specific environmental aspects of our personal lives to light so that its users can make more informed and responsible decisions. It provides web-based, personalized reports on environmental impact and exposure, currently focusing on mobility-related impacts and exposures, using only the commodity sensors built into everyday smartphones. Participants' mobile phones run custom software that uploads GPS and cell tower location traces to a private repository, where they are processed by a set of scientific models. The processing pipeline includes an activity classifier to determine whether users are still, walking, or driving[1]. It then uses both dynamic data sources, such as weather services, and more slowly changing GIS data, for example the location of schools and hospitals, or classes of food establishments to provide other inputs needed for model calculations. In a private user account, the results are presented to participants in an interactive, graphical user interface. Users can share and compare their PEIR metrics to those of other users via a network rankings web page, or via Facebook by running a custom application.

PEIR Usage Model: PEIR was created to explore how to make participatory sensing systems relevant at a large scale, through a platform that integrates mobile data collection, other real-time data sources, and models that take time and location as primary inputs. Addressing complex issues such as climate change or the effects of the built environment on obesity or wellness requires more than providing a one-time carbon calculation or online nutritional information; it requires tools that support people in gaining awareness of and weighing the costs and benefits of what they do. Like the EIR, the personal EIR promotes intelligent decision-making through review of concrete observations, comparison to similar situations and communities at a variety of scales, and consideration of patterns over time. However, unlike EIRs, which are done before major projects, PEIR operates at the resolution of the individual, runs continually, and its feedback is available on demand.

Reciprocity Between Impact and Exposure: Additionally, one quality that distinguishes PEIR from existing web-based and mobile carbon footprint calculators like Ecorio [2], or applications such as Carbon Hero [3] or UbiGreen [4], is its emphasis on how individual transportation choices simultaneously influence both environmental impact and exposure. PEIR provides users with information for two types of environmental impact, and two types of environmental exposure. The four everyday impact and exposure metrics are detailed in Section 4, but can be summarized as:

1. **Carbon Impact** is a measure of transportation-related carbon (CO2) footprint, a greenhouse gas implicated in climate change.

2. **Sensitive Site Impact** is a user's transportation related airborne particulate matter emissions (PM2.5) near sites with populations sensitive to it, such as hospitals and schools.

3. **Smog Exposure** is a user's transportation-related exposure to particulate matter emissions (also PM2.5) from other vehicles.

4. **Fast Food Exposure** is the time integral of proximity to fast-food eating establishments[2].

These metrics were selected because of their social relevance, and their ability to be customized for individual participants using time-location traces. We chose to focus on everyday transportation choices and patterns instead of the less frequent and more deliberately planned behaviors such as air travel since PM 2.5 particulates are of significant public health concern[3]. Extensions to capture decisions such as air travel through journaling or integration with online reservation systems, for example, are easy to imagine. Also, we have not focused on incorporating ambient exposure measurements from existing fixed sensor stations, but extensions to the system that would fuse that data with PEIR's current calculations are also envisioned.

1.1 Paper Organization

Section 2 gives a brief overview of the currently functioning PEIR system, including time location-trace collection from mobile handsets (2.1) and trace processing (2.2). At the heart of the latter are transportation mode activity classification and impact/exposure models: Section 3 evaluates PEIR's map-matching assisted HMM-based classifier and evaluates an enhanced activity classification technique (using GSM tower information) that may be integrated into future versions of the system. Section 4 describes the impact and exposure models that operate on the annotated time-location data. Section 5 describes PEIR's user experience and concludes with informal observations on system use during a two-month snapshot of a six-month trial. Section 6 begins to tackle privacy issues by introducing our application of participatory privacy regulation [7, 8] and describes specific mechanisms [9] planned for PEIR, namely support for "selective hiding," in which the system generates best guess proxy traces for times the user does not wish to reveal their real activity. Section 7 summarizes and concludes with observations on the running system's impact on our approach to participatory sensing systems research. Relevant related and future work is discussed within each section.

2. PEIR: SYSTEM OVERVIEW

PEIR has been running since June 2008 with thirty trial users using the system intermittently. After developing an initial end-to-end implementation, our most significant modifications to the architecture have focused on increasing calculation performance, so that users can see their impact and exposure scores within minutes of uploading their location

[1]Detection of biking and mass transit use are under development.

[2]This metric was included to demonstrate the broad applicability of the processing model used in PEIR to other geographically organized models, and because of the interest in the public health community about how exposure to certain types of foods may affect eating choices relating to obesity especially in children, where proximity to fast food has been demonstrated to correlate with increased obesity while proximity to non-fast food restaurants does not [5].

[3]A recent study by [6] of data from the 1970's - 1990's for fifty-one major metropolitan areas in the United States demonstrates that a reduction to exposure of these kinds of particulates correlates with increased life expectancy.

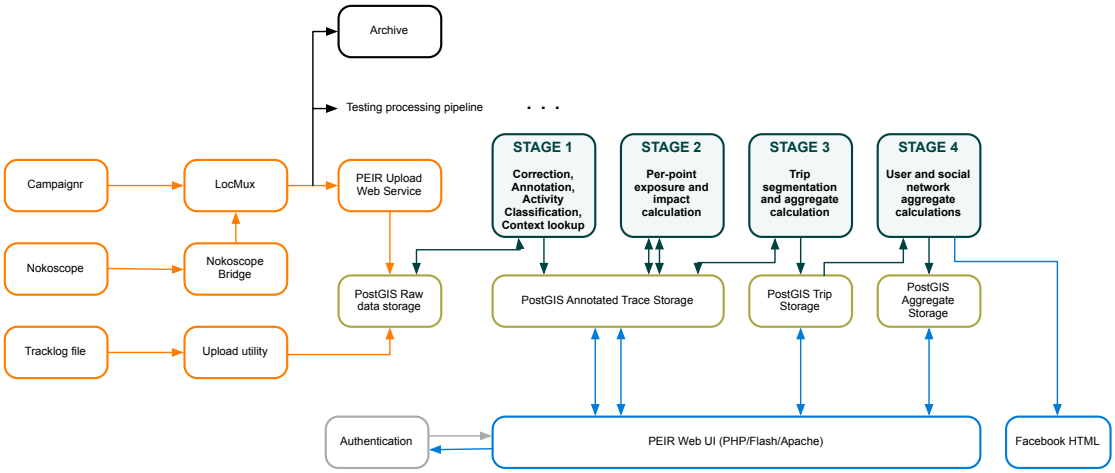

Figure 2: System architecture

data. This section describes the current architecture (Figure 2) as well as plans for improving its modularity, performance, and stability.

2.1 Location trace collection

PEIR processing operates on the time-location traces consisting of GPS records sampled approximately every thirty seconds (To reduce power and bandwidth, we experimentally selected the lowest sample rate that still resulted in good automatic classification of high speed travel by car.). In many cases, we also capture the connected cell tower, radio signal strength, and battery information for debugging and testing of future features.

The PEIR system accepts location records from mobile handsets, uniquely identified by International Mobile Equipment Identity (IMEI), that are posted in JSON format over HTTP or HTTPS, as well as upload of bulk time-location data in most tracklog formats, such as GPX, by integrating the GPSBabel software [10]. We currently support three different phone clients, two for Symbian S60 3rd edition, and one for Windows Mobile (Samsung Blackjack II)[4]. The mobile phone client used to capture location traces in most of our development and user testing has been the CENS Campaignr software, a native Symbian C++ application for Series 60 3rd edition smartphones supporting XML-configured data collection with both automatic and user-controlled sampling [11]. Using this platform we have tested PEIR with devices including the Nokia N80 (with an external GPS), N95, and E71, for over a year. More recently, we have created an experimental bridge for data from the Nokia Nokoscope client software [12], also for S60 3rd edition, which we hope will provide power consumption improvements over Campaignr. Campaignr and Nokoscope are both intended to run without user intervention; however we have found user-interpretable status display and feedback to be a consistently requested feature that we hope to implement in future versions of PEIR. Our initial development focus for the phone client has been on stability and, in the case of Campaignr, configurability for a number of different sampling tasks. Next, in addition to provide a more extensive user interface, we plan to focus on lowering power consumption and using battery level- and activity-adaptive GPS sampling such as that introduced by Nokoscope, as well as upload compression. We also plan to add device-level authentication and explore taking advantage of WiFi and Bluetooth stumbling and accelerometer data that most of these client platforms are capable of capturing, to enhance location trace and activity classification accuracy.

[4]We plan to integrate location trace collection clients for several other phone platforms in the near future.

2.2 Location trace processing

The server-side processes are implemented using Python code, shell scripts, and native/pre-compiled libraries. The PEIR web interface reads data from PostGIS [13] and is implemented in PHP and Flash served by the Apache httpd webserver, using the Wordpress blog engine [14] and Modest Maps Flash library [15]. A separate Kerberos/PKI based authentication server provides login services and stores personally identifying user information for this and other CENS urban sensing projects. The PostGIS store is on a Sun Solaris server, while processing runs on a second Linux-based server, and two separate servers for the Web front end are available to provide UI load-balancing.

Location trace data uploaded from phone clients is received by the location data multiplexing service (LocMux), a Python web service that receives and parses the JSON data, which forwards it to both testing and production servers and writes a copy to an archival store. On the production server, it is received by another web service that verifies the JSON and writes it to a PostGIS table configured with spatial indexing. The PEIR import and processing service operates on new data found in this PostGIS store in four stages.

Stage 1: Trace Correction and Annotation, Context Lookup Stage 1 of processing corrects and annotates the location data with transportation mode activity classification; i.e., it removes outliers and then attempts to determine the most likely activity for each sample - staying in one location, walking, or driving, which affects the impact/exposure models used in later processing steps. PEIR's activity classification approach uses a Hidden Markov Model that, in addition to speed, incorporates a "map-matching" technique for freeway annotation as a feature(Activity clas-

sification is described in more detail in Section 3.1). PEIR Stage 1 also looks up other contextual information, such as temperature and humidity data, needed for the model calculations. These annotations are added to each point by filling fields in the PostGIS table for processed points. Future improvements possible in this stage include better activity classification (more specific modes and higher accuracy), as well as the explicit inclusion of the uncertainty of the captured and computed data along with the primary outputs of the model outputs.

Stage 2: Exposure/Impact Calculations Stage 2 performs per-point impact and exposure processing that is described in detail in Section 4. The annotated points are used to calculate per-point values for each of the four models: (1) Transportation-related carbon impact, (2) PM2.5 exposure, (3) PM2.5 output near sensitive sites, and (4) fast food exposure. These annotations are added the PostGIS table in the same manner as those in Stage 1.

Stage 3: Trip Identification and Annotation Next, the annotated data points are segmented ("chunked") into "trips" and the four environmental impact/exposure measures are aggregated per trip (A trip is defined as a traveling from one place to the other where a user stays for more than 10 minutes.). These trips, the primary unit of data with which the users interact, and their PEIR metrics are stored in a separate PostGIS table.

Stage 4: Aggregate and social network calculation Stage 4 calculates trend and social network comparisons. These points of comparison for impact and exposure metrics support participant understanding and engagement with the system. For completed trips, a service updates trend information using the results of completed trips, calculating and storing weekly, monthly and yearly aggregates. Relative comparisons among users who are in the social network are normalized for amount of data uploaded.

3. TRANSPORTATION MODE CLASSIFICATION

Activity classification is crucial to PEIR as it enables the proper impact and exposure models to be applied automatically. We describe and evaluate a tailored map-matching technique designed to meet PEIR transportation mode activity classification needs. We detail our current algorithm and then develop and evaluate a planned enhancement, using GSM to improve upon the GPS-only classification approach by increasing the detection accuracy of surface-street driving and indoor locations.

3.1 Activity Classification

GPS data have been widely used to infer physical activities. Much of the work [16, 17, 18, 19] takes external geo-indexes such as map information to add high-level context to location data points and improve the performance of the classifiers. The work of [16, 17] uses GPS data with external knowledge about bus routes and bus stops to infer and predict a user's transportation mode such as walking, driving, or taking a bus, and in [18, 19] models of a user's activities and places from traces of GPS data and locations of restaurant, stores, and bus stops are employed for classification purposes.

For the PEIR application, it is most important to identify driving activities and less important to distinguish between staying and walking because emission values are zero in both cases. In many cases, it is easier to identify that a person is driving even when speed values are low if we make use of the fact that a GPS trace locates the user on a freeway, where the user is very unlikely to be walking. Therefore, our classifier uses freeway annotation information in addition to speed values as feature inputs to the classifier. Using the freeway annotation information obtained from our map-matching technique as part of our classifier, the accuracy increases from 40% to 82%, based on test data collected while a user was driving in heavy traffic on the highway.

3.1.1 Map-Matching Algorithm

We use a map-matching technique to find out whether a user is on a freeway[5]. Determining which road a user is on is non-trivial because individual GPS points often deviate from the physical road being traversed due to both inaccuracies in GPS measurements and the maps themselves. Map-matching techniques have been developed to improve the interpretation of GPS location data. Naive map-matching finds the nearest road segment as a correct match [20]. Although it is simple and fast, it is sensitive to the spatial road network and often fails in practice. Figure 3(top-left) shows one example of GPS data errors where the points depart from the street. In this example, the naive approach identified Mississippi Ave as the nearest road, and annotated the point as being on a surface street, while actually the person was driving on Highway 405.

To address this misclassification behavior, we implemented a modified map-matching scheme, which we refer to as "Intersection-based map-matching." It finds pairs of intersecting roads that a user passes by, and then extracts the common road among subsequent intersections to determine the street on which the user was most likely to be traveling. Consider Figure 3(bottom-left)'s example, in which naive map-matching would have mistakenly selected Highway 405 as the nearest road. The intersection-based method first records that the location trace falls near an intersection between Sawtelle and National boulevards, and that the next falls near the intersection of National Blvd. and Sepulveda Blvd. The system selects National Blvd. as the street that the user is traveling on.

We detail the steps for our new algorithm as follows:

1. Find the two nearest roads for each data point.
2. If distances from a GPS data point to the two roads are less than .04 miles[6], label the GPS data point as the pre-intersection. Otherwise, add it to the buffer of data points. Continue until the next intersection point is identified, referred to as the post-intersection.
3. Compare the pre- and post-intersections and identify the road that appears in both intersections as the correct match for the buffered data points. If no common road is found, consider the subsequent GPS data points to identify an alternative post-intersection.

[5]Note that the purpose of map-matching technique usages for our application is not to get precise GPS location data but to annotate GPS traces with freeway information.
[6]We found that our map-matching method performs the best when we use 0.04 miles as a threshold value after applying the value from 0 to 0.1 miles. This value has to be evaluated further.

	Case 1	Case 2	Case 3	Case 4	Case 5	Average
Naive map-matching	76%	58%	93%	57%	56%	68%
Intersection-based	5%	83%	100%	77%	96%	72%
Intersection w/nearest road and substitution	89%	83%	100%	63%	96%	86%

Table 1: Comparison of accuracy for different map-matching approaches

Figure 3: Example of naive map-maptching failure (top-left), intersection-based map-matching (bottom-left), and intersection with nearest road and substitution (right)

After using the system for several months and examining data traces and errors, we added two mechanisms to improve our Intersection algorithm. The first mechanism considers the nearest road within .04 miles as a possible intersection point when looking for a common road segment between two consecutive intersection points. This helps us correct for the situation when the captured GPS data points are not near intersections and we therefore miss turning points that occur in between the captured GPS data points. The second mechanism considers replacing both pre- and post-intersections when there is no common road between two consecutive intersections; it does so by assigning the post-intersection as the new pre-intersection, and using a subsequent intersection as the new post-intersection. This helps us correct for erroneous identification of pre-intersections, so that the error does not propagate. We call this new algorithm, Intersection w/ nearest roads and substitution. Figure 3(right) gives an example in which this algorithm correctly locates a set of points. The previous Intersection algorithm would have tried to use pre-intersection A, B and post-intersection C, D to identify the common road and when it failed (there is no common road.) would have tried successive post-intersections until it found A, E. Therefore the algorithm would identify A as the road traveled while a user was actually driving on Highway C. Intersection w/ nearest roads and substitution instead uses A, B as pre-intersection and nearest-road C as the post-intersection. When no common road is found, it substitutes C as the new pre-intersection and C, D as the new post-intersection and thereby correctly identifies road C as the road traveled.

We evaluate our map-matching approach with a data gathered from five PEIR users over the course of two hours. To obtain ground truth information on the roads that users traveled, we visualized each individual's time-stamped traces on a map and interviewed users to correctly label them based on their recall. We included different scenarios: crossing under highways with several turning points (case 1), getting on and off highways (case 2 and 4), driving on straight paths (case 3) and switching highways (case 5). We use accuracy as our evaluation metric and accuracy is defined as the percentage of the number of correctly matched points. The result is shown in Table 1. Our improved Intersection algorithm with nearest road and substitution mechanisms added performs the best with the average accuracy of 86%. Case 1, in which there are many turns made, illustrates the situation in which these improvements have the most dramatic effect, improving map-matching accuracy from 5% to 89% for that case. In one of our five cases, Case 4, accuracy reduces from 77% to 63% because the algorithm can get confused by highway segments that lie right above surface streets.

3.1.2 Classification Model

For each GPS data point, we obtain a speed value and a freeway annotation by using the above map-matching technique. Hidden Markov Models (HMM) have been successfully used in modeling different types of time-series data, e.g. in speech recognition, gesture tracking [21, 22]. We use HMMs as our inference model to capture temporal dynamics. There are three states, "Staying in place," "Walking," and "Driving." Instead of directly using the raw feature values, we discretize values into six observations based on the distribution of feature values. The initial probability is assigned equally for the three states. Transition (the change of the states in the underlying Markov chain) and observation (how likely we observe a certain observation for each state) probability matrices are trained using a data set of sixty hours gathered from one user who also provided ground truth annotations using a journal during the course of the data collection period. The most likely sequence of hidden transportation mode states is found using the Viterbi algorithm [23].

We asked five PEIR users to annotate location traces with their transportation modes for a day; we wrote a custom python script showing multiple choices, "Stationary Indoors," "Walking Indoors," "Stationary Outdoors," "Walking Outdoors," "Driving," and users were asked to tag their location traces whenever they changed transportation modes. In total, we gathered fifty hours of data from these five users. The accuracy is defined as the percentage of the number of correctly predicted data points. As shown in Figure 4, our classifier performs fairly well, over 80% accuracy, for user 1, 4 and 5. However, it works poorly for user 2 and 3. By examining the data, we found that environmental interferences inside buildings led to lower positioning accuracy, in the order of several meters or worse. For user 2 and 3, their

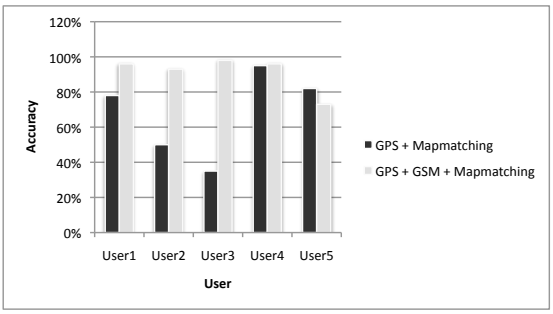

Figure 4: GPS-based versus GPS and GSM-based accuracy

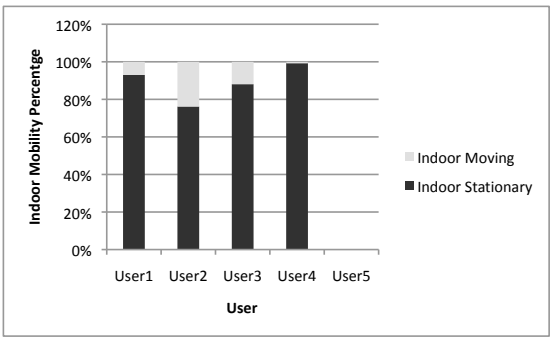

Figure 5: Indoor motion statistics

staying-in-place activity involved a lot more movements than the staying of other users (See Figure 5), introducing more errors in the GPS data; incorrect GPS coordinates with high speed values were recorded. This phenomenon caused the data to be misclassified as walking or driving. After excluding data gathered indoors, the classifier identified user 2 and 3's data 96% and 98% correctly.

3.2 Improvements by Leveraging GSM data

PEIR's models are sensitive to whether a person's activity is properly classified as driving. Classification based on GPS data alone is difficult if GPS performance is compromised by limited satellite visibility. This can result in misclassification of users when they are indoors (false positive for driving) or near tall buildings (false positives and negatives). In addition, correctly classifying slow surface street driving is difficult at any time, because map-matching cannot rule out walking as it can if the user is on a freeway. Therefore, we explore using GSM cell tower association data[7] to enhance classier performance, especially for these cases.

A GSM base station is typically equipped with a number of directional antennas that define sectors of coverage or cells, each of which has uniquely identifiable cell ID by combination of "Country Code," "Network Code," "Area Code" and "Cell ID." Information from the cell ID provides a rough indication of a person's position [24]. Features derived from this information, such as the number of changes in the associated cell IDs for certain duration, can help the classifier identify activities correctly when GPS data is noisy.

[7]We focus on what can be achieved with only GPS and GSM as they are already widely available, enabling PEIR to be deployed now without waiting for custom sensors.

The benefits of GSM data afforded by high coverage and availability on mobile phones have been recognized earlier. Google MyLocation [24] calculates a user's position relative to the unique identifications and footprints of nearby cell towers to find a user's approximate location. The work of [25, 26] uses neighboring cells in addition to the current serving cell from mobile phones to recognize high-level properties of user mobility. Unfortunately, many mobile phones do not provide access to the list of multiple cell towers in range (i.e. mobiles phones based on Symbian OS which constitute 67% of the "smart mobile device" market only reveal the connected cell ID information [27]). But we can still take advantage of the connected cell ID information for transportation mode classification.

Features	Window Size(Seconds)
Total Traveling Distance	60
Average Speed Differences	120
Average Speed	180
Average Traveling Distance	240
Number of Unique Cell IDs	150,300
Number of Cell ID Changes	240
Freeway Annotation	1

Table 2: Selected features for GPS and GSM classifier

Our GSM-enhanced classifier considers previous data points as well as the current data point to compute its features. In this way, we extract meaningful features from the associated cell information. The GSM enhanced classifier improvements are important for the PEIR application because they help to avoid classifying noisy indoor GPS readings as driving, and help to correctly classify slow driving on surface streets as driving instead of as walking.

Each data point used by this algorithm consists of GPS coordinates, speed, associated cell ID and timestamp. The seven selected features listed in Table 2 are computed for each data point. To identify the most effective window size for each feature we evaluated values from 5 to 300. We adopted the C.4.5 Decision Tree as our inference model for initial offline analysis because of its simplicity [28].

To better understand the performance results of the models, the same data set used in evaluating our running classifier in Section 3.1.2 is used to test the proposed GPS and GSM based classifier. The results are shown in Figure 4. The classifier achieves the overall accuracy of 91% and works well when users are on surface streets and highways, as well as indoors. The techniques evaluated here will require further analysis to assess their implications for system performance. Also, in addition to using GPS and GSM features [29, 30, 26], future activity classification techniques could leverage additional on-board sensor types (e.g., accelerometers and WiFi [31, 32, 33]), as well as additional GIS and map data (e.g. bus routes [16, 17, 18, 19]).

4. NEAR REAL-TIME MODELING OF EXPOSURE AND IMPACT

Server-side PEIR analysis culminates with the computation of a series of metrics that describe both the impacts that a participant has on the communities they travel through during their day, as well as the environmental hazards they

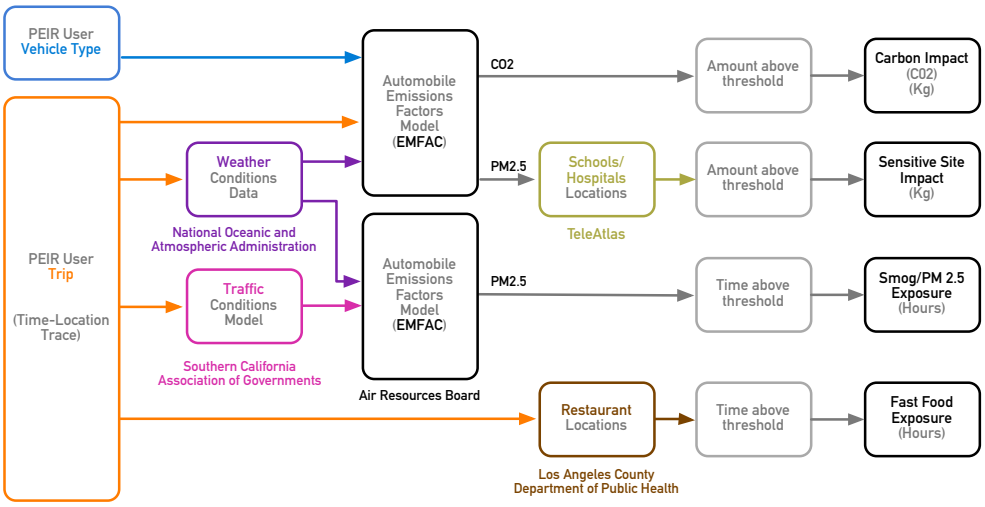

Figure 6: Data flow diagram

encounter. Admittedly, these are fairly broad terms, and in our first implementation of the system, we focused on four specific measures. In some sense, our choices (and our focus on transportation in general) were inspired by our experiences living in Los Angeles. While other groups, possibly in different parts of the country, might have different concerns, the PEIR system has been designed around the broader concept that location-traces, decorated with some estimate of participant activities and local "context," can be combined with impact and exposure models to create new interpretations of a participant's choices. In this section, we describe the computation of the four metrics we selected for the first implementation of PEIR: Carbon dioxide emissions, PM 2.5 emissions near sensitive sites, PM 2.5 exposure and fast food exposure. The data flow for these processes is shown as Figure 6.

4.1 Modeling assets

A number of data sources and modeling "primitives" are used repeatedly in the PEIR system. Most of these sources are currently updated on relatively long timescales (possibly 6 months to a year or more) because that is what is available and most geographically scalable. However, the system can draw on more dynamic inputs as they become available.

4.1.1 Roads and specific locations

Maps of the local roadways (exposed through PostGIS) are critical for PEIR. StreetPro, part of the ESRI 2006 Maps and Data collection, provided these [34]. In addition to aiding activity classification, the nearness of a participant to a freeway, say, is the critical parameter in computing their exposure to PM 2.5. We also depend on StreetPro for the location of sensitive sites within Los Angeles County, specifically schools and hospitals. Finally, from the Los Angeles Department of Public Health, we obtained a list of all the fast food restaurants in Los Angeles County (with the exception of Pasadena, Vernon and Long Beach) [35]. To ease computation, a "buffer" polygon is computed for each point (each hospital, school, and fast food establishment) so that later PEIR processing can easily establish whether a participant is traveling nearby (for sensitive sites the buffer is 200m, while for fast food it is a quarter of a mile - see the text below.). Again, these point data (hospitals, schools, fast food establishments) reflect the interests of the PEIR team; there is nothing in our architecture that would preclude other choices.

4.1.2 Weather

With our emphasis on transportation and airborne pollutants, our impacts and exposures necessarily depend on the local weather conditions (specifically, temperature and relative humidity). Through the Meteorological Assimilation Data Ingest System (MADIS), the National Oceanographic and Atmospheric Association (NOAA) maintains a real-time database of observations from weather stations worldwide [36]. These data are culled hourly for stations in Los Angeles County. To tie these data to a participant's location, we create another table that associates Zipcode Tabulated Areas (ZCTAs) [37] with the five nearest weather stations (nearest to the centroid of the ZCTA; note that any two MADIS weather stations are separated by several ZCTAs.). This table was formed using ArcGIS [34]. Future versions should investigate spatial mappings that are more uniform or reflective of micro-climates than ZCTAs.

4.1.3 Traffic conditions

Real-time traffic measurements are available in some major metropolitan areas like Los Angeles. However, we wanted PEIR's approach to be scalable to cities without such infrastructure. So, we instead developed our first implementation around a traffic flow model from the Southern California Association of Governments (SCAG) [38]. Originally designed for transportation planning, this model reports bidirectional traffic flow information (the number of vehicles per hour, their types and their speeds) for all road segments in Southern California. These data are further stratified into six time frames (morning and evening rush hours, midday and night) across six county regions. Eventually, the complete "model" (table) is made visible to PEIR by exporting it from ArcGIS to PostGIS.

4.1.4 Vehicle emissions

The final ingredient in our asset list has to do vehicle performance. To support its planning process, the California Air Resources Board (CARB) has developed the Emissions Factors Model (EMFAC) [39]. EMFAC, a FORTRAN program, computes vehicle emissions based on current weather conditions (temperature and relative humidity), and the speed and type of vehicle (the former derived from a participant's trace, and the latter collected at sign-up). While EMFAC produces a number of estimates, we rely on PM 2.5 and carbon dioxide emissions. During our initial development of the PEIR system, EMFAC proved to be a bottleneck. To speed computation and to "open" the model, we developed an approximation to EMFAC, computed via a functional ANOVA model [40]. Through phases of repeated fitting and testing, we created a tensor-product spline model that both allowed for fast computation as well as a view into the dependence of the emissions outputs as functions of weather and vehicle characteristics. We used the statistical computing environment R [41] to fit the functional ANOVA model and stored the result as a Python object; in this way we can update the model as needed and not interfere with the running system.

These four data and processing components are the basis for the impact and exposure computations in PEIR. Note that a participant can easily interrogate each of the first three pieces with a GIS platform. By design, the functional ANOVA model approximating EMFAC is also easily viewed, emissions being represented as a combination of several curves and surfaces, each a function of the four input variables. As a group, these four can also benefit from participant contributions. With an expanded participation model, members of the PEIR community could add new sensitive sites and new points of interest; contribute to finer-scale citizen-supported weather monitoring; correct SCAG estimates with actual traffic observations [42]; and replace generic EMFAC profiles with the specifics of their own vehicles. These additions are a key part of PEIR's future.

4.2 Computations

The four components listed above are called on at several points during the PEIR processing pipeline. The impact and exposure measures are computed record-by-record, where each record consists of a participant's system ID, their vehicle type, their current location and speed (derived from their GPS), and an activity class.

4.2.1 Impacts: Carbon dioxide and PM 2.5

Carbon dioxide emissions are the simplest to explain. For those records classified with the activity driving, we add current temperature and humidity at the participants' location (data from the closest weather station that reported data in the last hour; closest among the five associated with the ZCTA of the participant's current location). We then apply the functional ANOVA approximation to EMFAC to estimate current carbon dioxide emissions. Technically, EMFAC computes an emissions factor in units of grams per mile; this is translated into grams using the participant's speed and an estimate of the amount of time traveling at that speed (computed as the $(t_j - t_{j-1})/2$where t_j is the time associated with the jth record). PM 2.5 emissions near sensitive sites work in the same way, except that we accumulate emissions (again, in units of grams) only when the participant is within 200m of a sensitive site[8]. The buffer zones around sensitive sites are pre-computed, and so this step involves a single spatial intersect query(A buffer of 200m was chosen after reviewing literature on the rate at which pollutant levels drop off with distance form the source [43].).

4.2.2 Exposures: Fast food and PM 2.5

Fast food exposure also makes use of a spatial intersect query using the participant's current location. In this case, the buffer is a quarter mile and was determined based, in part, on ease of access. The records of a trip are then annotated with a flag indicating whether a fast food establishment was nearby. We then accumulate the total amount of time near a fast food restaurant per trip.

The computation of PM 2.5 exposure is broadly similar in that we need to determine how close participants are to know hazards, in this case vehicle emissions. For each road segment, the SCAG model is used to estimate the number, type and speed of vehicles present. Given a PEIR record, then, we combine the SCAG model, data on the local weather conditions and EMFAC predictions to estimate a participant's rate of exposure to PM 2.5. Air quality indices do not quote emissions rates, but instead work with concentrations of pollutants. For interpretability, we further transform our estimate into a concentration by taking into account the volume of air over a road segment. In this way, we can determine the amount of time per trip that a participant spends in a hazardous condition (a concentration of 0.112716 $\mu g/m^3$) [43].

Note that while in some sense exposure and impact are dual calculations, we quote metrics that are in units of grams for impact (quantity of pollutants) and time for exposure (time spent in a hazardous condition). Our choice for the impact metric made PEIR results comparable to existing carbon calculators; while our choice for the exposure metric was more directly interpretable for the PEIR user. As with other parts of the modeling system, there is nothing fixed about these choices. Ultimately, participants should be able to shape the output in the format they are interested in. We present these metrics as a proof of concept.

5. USER ENGAGEMENT

While environmental impact and exposure values may not be immediately understandable, location (from which those values are derived) is already part of people's everyday vocabulary. We designed the PEIR user interface with this in mind, using a map-based visualization as its foundation. Just as the traces that underlie PEIR metrics provide powerful, intuitive insight, that same location information is highly sensitive and potentially of great privacy concern. Our current system takes privacy seriously, and as we offer more features based on location data and PEIR model outputs, we are developing techniques that provide users with full control over who sees their data, data deletion and sharing, and transparency of data processing. This section describes the spatially-oriented user experience, explores methods to share PEIR outputs with others, and describes observations learned from monitoring usage of the system. Privacy protecting techniques are further explored in Section 6.

[8]Currently, our system only considers a small number of sensitive sites (hospitals, schools), and so the impact measure is often zero.

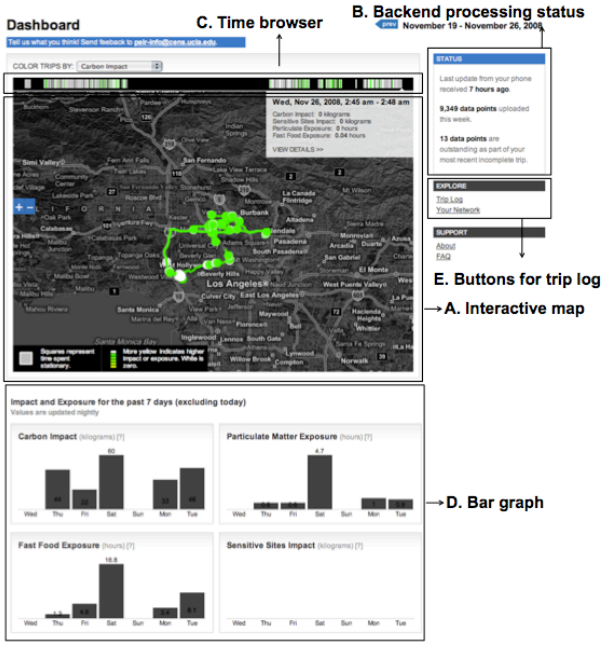

Figure 7: Dashboard view

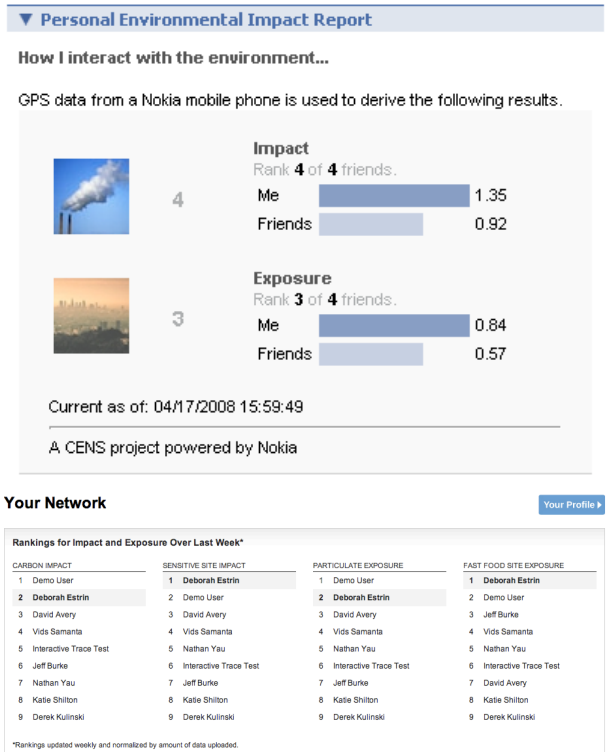

Figure 8: Facebook application(top), PEIR network view(bottom)

5.1 PEIR metric legibility

PEIR users see a dashboard of activity (Figure 7.A) for the week as well as current upload and backend processing status (Figure 7.B) at initial login. We use an interactive map to visualize location traces color-coded by level of impact or exposure. The more intensely colored a trace, the higher the impact or exposure metric is. Users are able to select among PEIR's available models - carbon impact, particulate matter exposure, sensitive sites impact, and fast food exposure - to color traces by the metric of interest.Traces are highlighted as the user scrolls over them on the map, and trip details (e.g. trip type, impact estimates) are displayed when the user clicks on the corresponding trip. Trips can be browsed by time via the time browser (Figure 7.C) on top of the map. Color-coding corresponds to that of the map. Bar graphs (Figure 7.D) supplement the map and provide a snapshot of the week's impact and exposure in a day-by-day breakdown. This dashboard is meant as an overview for the user. From the dashboard, the user can look at previous weeks, or move on(Figure 7.E) to extended details for an individual trip. Details can also be reached from the PEIR trip log, which is simply a list of all trips a user has made that can be filtered and sorted by date, trip type, and impact and exposure estimates.

To support deeper legibility of PEIR data provenance, the UI offers a breakdown of how each trip's impact and exposure values were estimated; it shows the user that it is not simply the act of driving that causes impact and exposure. The trip breakdown shows, for example, percentage of time spent on the freeway, activity classification, and the weather at the time and place the trip took place.

5.2 User Publishing/Sharing

The dashboard and trip details show information on an individual's impact and exposure, but how do users learn how they relate to others? Comparing themselves to peers, coworkers, or friends can help users determine where their carbon consumption or pollution exposure ranks, and perhaps provide some incentive to reduce.

The PEIR publishing and sharing functions provide context for users to interpret and compare their results. Users share their impact and exposure via a Facebook application. The Facebook application shows impact and exposure for both the user and the average values for Facebook friends who have also installed the application (Figure 8(top)). Green icons of trees appear if impact and exposure are low relative to friends, and smokey and smoggy icons appear if impact and exposure are high. Users can also see their rank among friends. Within the PEIR pages, users can see the same rankings among friends (Figure8(bottom)), as well as a weekly snapshot displayed in bar graphs (not shown in figure) similar to that of the user's individual dashboard.

Because location data is particularly sensitive, PEIR defaults to sharing only aggregate impact and exposure data. Both user profiles and the Facebook application share and compare daily impact and exposure numbers without revealing any location data. Future improvements to PEIR will also enable users to share location data with people they trust. Giving users the option to share designated routes with specific people could encourage discovery of new routes or participation in workplace competitions. UI features will enable users to select routes to share with specified individuals. Selective, opt-in sharing of sensitive location data helps users retain control over exposure. Additionally, logging and displaying outsider access to user data can give participants feedback on how their data is used. Displaying map or trip diary access ("John Smith has accessed your trip diary 2

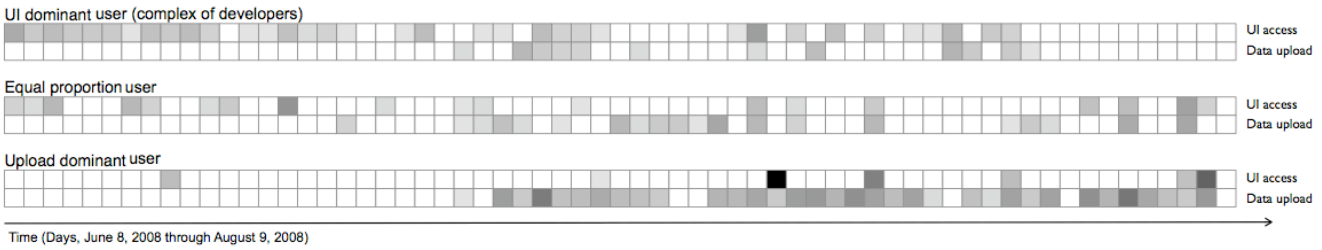

Figure 9: Usage patterns, detailed view. Darker boxes signify higher daily counts.

times this week.") alerts users as to who is viewing their data, and can help users hold friends accountable for shared location data [44].

5.3 Usage Statistics

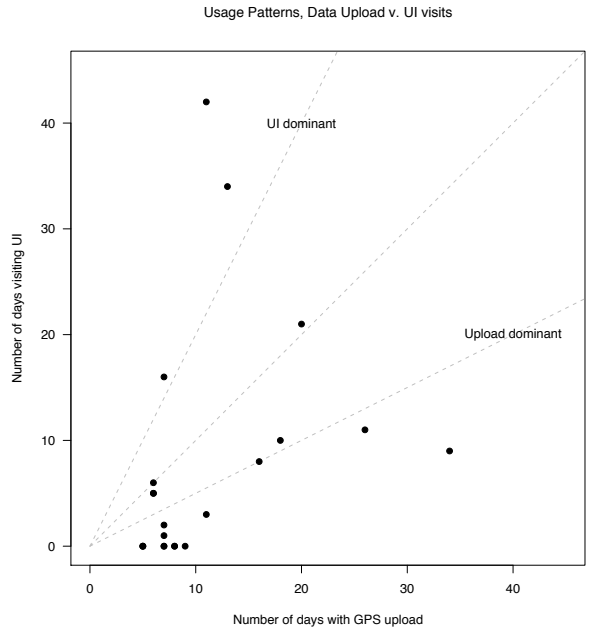

Figure 10: Usage patterns of UI views and data upload

The PEIR system using this interface has been running in "pilot production mode" since June 8, 2008 and as of November 28, 2008 has logged over four million individual GPS points grouped into over 20,000 separate trips. While the project site launched on June 8, both the underlying data processing (trip chunking, activity classification) as well as the user interface continued to undergo periodic upgrades. Many of these system changes make it hard to analyze usage patterns across the entire time period. For example, as our trip chunking technology changed, so too did the estimated duration of each trip (for the most part shifting from shorter to longer grips as the algorithm was made less sensitive to GPS noise). In a similar way, the basic structure of the user interface, the placement of the central map-based trip explorer, and the addition of a simple trip tagging module, changed how participants examine their data (and their reason for doing so; trip tagging was introduced, for example, to help diagnose and correct errors in the activity classification scheme.). Finally, our group of dedicated participants is relatively small (on the order of 20 or so of the 30 total distinct users) and made up of people with quite different motivations for examining their data. While most of our participants approached PEIR from the standpoint of personal reflection with an interest in understanding their transportation choices, many were also system designers, tasked with assessing and improving the various processing components that constitute PEIR. The casual testing and monitoring by these participants is immediately obvious in most data summaries of system usage (plots describing the frequency of data uploads or UI requests, say). In addition, we can see the effect of external campaigns, periods of focused data collection, that impact all of the participants. Therefore, any PEIR usage summary is necessarily equal parts proof of concept (the system is alive and able to handle contributions from an expanding group of participants.) and part investigation into broad modes of interaction (such as patterns of upload versus UI access).

Given all these caveats, we selected a two-month period immediately following our initial launch (June 8, 2008 through August 9, 2008) when the processing and interface designs were fairly stable. In this window, just under 1300 trips were recorded with a median duration of about 25 minutes. In Figure 10, we plot the number of days each active participant (someone with at least five days of trip uploads, trips lasting 15 minutes or longer) examined their data by loading the PEIR UI, versus the number of days they uploaded data during our two month window. We have added dashed lines through the origin with slopes one, a half and two(indicating users who uploaded and examined data at roughly the same rate, and those that were UI or upload "dominant"). From this plot, we can identify two extremes where participants contributed trips or visited the UI in much greater proportion. Here we see the roles of our participants shaping usage; the person with over 40 days of UI visits was actually more of a complex than an individual, with several PEIR team members logging in daily as this participant (with their permission) to check the system's uploads(Our logging facility cannot distinguish between these scenarios.). We expect that the upload-dominant usage pattern will be more common when PEIR is open to the general public, since data upload can be automated.

In Figure 9, we examine a participant's data for each of the three usage patterns more closely, in particular the interplay between upload and visual analysis via the UI. The UI dominant participant is the "complex" as we indicated before, the equal proportion participant had 20 days of UI access and 20 days of data upload, and the upload dominant

participant had over 30 days of trips uploaded and 6 days of UI visits. The rows are grouped into pairs where the lower represents upload activities and the upper indicates UI accesses for each of the three participants. Furthermore, each column represents a day starting from June 8 and ending with August 9. The cells are shaded according to the number of trips uploaded on the indicated day or the number of (unique) trips viewed through the UI on that day (these counts range from zero to 62; a linear grayscale is applied to the square-root of these counts.).

While we hesitate to infer general PEIR usage patterns from this fairly small sample of participants, we can draw some modest conclusions. First, PEIR is a functioning platform, easily capable of handling the demands of this small community, and currently being scaled support groups of roughly 100. Next, the interplay between data upload and UI access for our "equal proportion" participant is consistent with the experiences of several members of our group; data collection can continue relatively easily in the background, with only periodic visits to the UI to assess overall exposures and impacts. Interestingly, uploads that occur after a gap of a few days are usually accompanied by a same-day visit to the UI. Again, this pattern makes sense given the kind of analysis PEIR provides; the more (temporally) distant a trip, the harder it is to examine (and instead an overall commuting pattern becomes relevant.). As we scale PEIR to process GPS traces from a larger group of participants, we are architecting a more complete UI and upload tracking system. Future reports will allow us to detail the typical age of trips analyzed in the UI, the interplay between the Facebook application and our GUI, and system-wide "challenges" issued by participants to lower their impacts. Having presented the spatially oriented user interface for PEIR, in the next section we address the critical privacy issues that arise from the capture, storage, and sharing of the personal location traces that underlie PEIR.

6. PARTICIPATORY PRIVACY REGULATION

Although capturing location traces and displaying inferences based on them are critical to PEIR's purpose, both are potentially invasive. Shared or stolen data on individuals' routes and routines could compromise their safety. Granular records of participants' actions prevent plausible deniability, compromising behavior we use to smooth social relationships. Location tracking can also create chilling effects on legal but socially stigmatized activities [45, 46, 47].

These risks involve complicated tradeoffs. Are the benefits of using PEIR worth the risks of exposure? The answer will vary depending upon individual preferences and situational factors (exposure of what, and to whom). System legibility principles discussed above, such as displaying and explaining all data stored in the system, already support users in understanding the scope of data collection. In addition, discussion forums can encourage users to voice their privacy concerns, and learn about the concerns of others, while system alerts can remind users of pervasive data collection, sharing choices, and data retention windows. Below, we explore more involved techniques for participatory privacy regulation: selective hiding, which attempts to return the capability for plausible deniability, as well as selective deletion and retention rule sets.

6.1 Selecting Sharing and Hiding

As mentioned in Section 5.2, the PEIR system will enable users to share their location traces with people they trust. But, in certain situations, individuals might not want to share all portions of their trace. A common privacy filter might be to hide a particular trip to a location (such as visit to the hospital, a certain store, or a particular restaurant). However, simply removing the trip is suspicious - the lack of data may raise attention to the space/time being protected. Thus, we propose a new approach, selective hiding, that replaces a location trace segment to a particular significant destination with a trace that is most closely related to the original in terms of model output equivalency and is based on historical information of the user's likely movements. This approach has the following objectives:

- Privacy enhancement: Increase the user's sense of privacy when sharing a substituted-trace.
- Application output equivalency: The substitute trace results in minimal changes to the PEIR metrics.
- Believability: The substitute trace should be credible to the people with whom the user shares his/her data.

This selective hiding approach lets a user choose sensitive destinations and the algorithm generates candidate substitute traces that do not contain the indicated destination but which do generate similar PEIR metric outputs. The system adjusts the candidate substitute traces to recreate the PEIR metric outputs by time shifting, and increasing or decreasing the time duration for activities. It then selects the best-fit substitute and notifies a user with the changed routes. The current system runs offline and semi-automatically, but will be integrated into the running system in the future.

To better understand the performance of our technique, we also implement and evaluate two previously proposed countermeasure methods: spatial rounding and noise addition. Spatial rounding is the process of changing the original location to one that is coarser [48]. An example of spatial rounding is the process of rounding or snapping locations to larger granularity zone [49]. We snap specific latitude and longitude to the northwest point on a square grid with spacing $0.1 \sim 10000$ meters. Noise addition deals the process of distorting the original location value by a certain amount. For each point, we generate a noise vector with a random uniform distribution over (0,Π) and a Gaussian-distributed magnitude from N(0, σ) [49]. We use ten different σ values from 50 to 500 meters in increments of 50 meters.

To show the effectiveness of the three measures in enhancing privacy (in this case hiding a trip to a particular significant destination), we first show how much data corruption results when prior art counter measures are used to hide the person's destination. We used a data set gathered from three users for one month and identified their significant destinations from original and obfuscated traces. We found that in order to protect significant destinations, the spatial rounding/noise addition has to be significant (at least 100 meter addition), which conforms with the results shown in [49]. Second, we take a particular user's trace in PEIR and apply the three methods to hide sensitive destinations. The scenario is one in which the user wants to hide his/her visit to a particular restaurant (C) on Friday evening by modifying the original trace, A, B, A, C, A. Using the prior

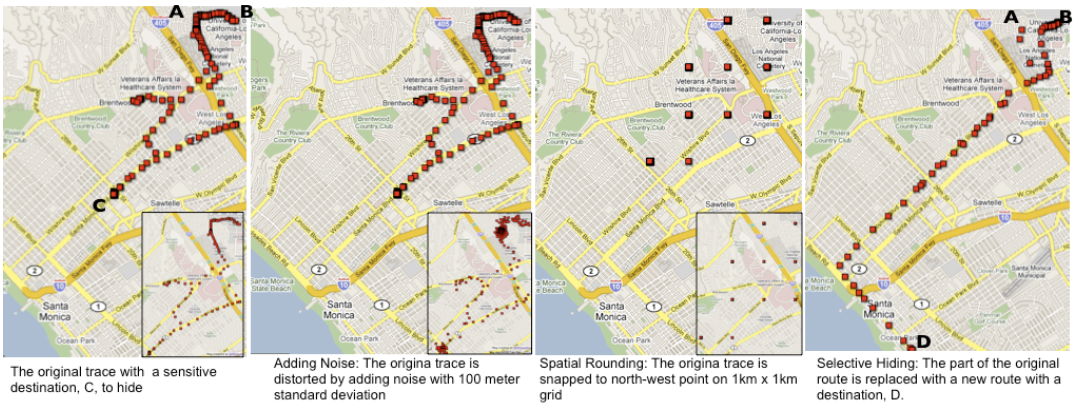

Figure 11: The effect of adding noise, spatial rounding and selective hiding in protecting privacy

art countermeasures, fifty and 100 meter noise is added, and 100 meter and 1km grid values are applied to the original trace. In comparison, our selective hiding method enables the user to select substitute traces from previous trips that start from either A or B, and end at A in order to fill the gap in the original trace. The substitute is selected such that the substitution produces similar impact and exposure metrics. Because most people have different mobility patterns on weekdays and weekends, in this example the search is based on previous trips only on Friday evening. This method allows us to find "realistic" route candidates. A new destination, denoted by D in Figure 11, is chosen and a new trace, A, B, D and A, is generated to replace the original one. The time and date of the route candidate are shifted so that the new trace fits into the original ones. The duration of driving from D to A is reduced slightly by increasing speed values to decrease carbon impact values. The duration of staying at D increases in order to increase a particulate exposure value. We could increase the duration for A or B, but D is chosen since it is the most polluted.

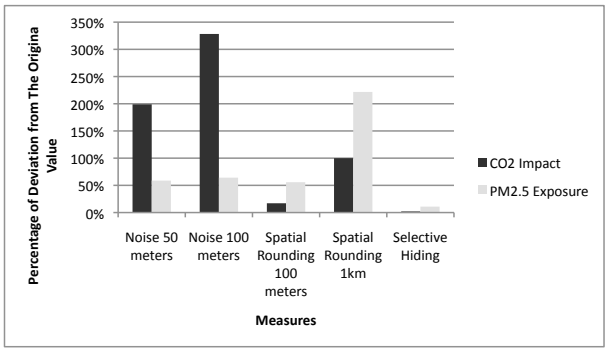

Figure 12: Comparison of selective hiding techniques

Figure 11 shows how the original trace changes after applying the three counter measures - prior art and our mechanism. Although location traces become more vague after adding noise and rounding, the movement pattern remains the same and sensitive information is not necessarily hidden such that the user may not feel comfortable sharing the modified traces. On the other hand, the new location traces using our method show a different, yet credible movement pattern while producing similar application outputs. Our countermeasure technique generates substitute location data that, unlike prior art, still generates believable output from location-based models. This further helps avoid detection efforts that would use anomalous model outputs as indicators of faked data. Figure 12 shows how much PEIR metrics deviate from the original impact and exposure values when we applied prior art counter measures relative to our path substitution approach. Adding noise disperses points and increase speed values, which results in higher carbon emission values. Spatial rounding to points on a grid causes several points to go to one particular point, which increases "staying points." In the case of spatial rounding with 1km size grid, the carbon emission values become zero. That is, the high degree of corruption required to preserve privacy could make location-based services like PEIR unusable by altering application outputs. However, selective hiding with substitute path segments produced nearly the same PEIR model output met.

6.2 Selective deletion and retention

Another important privacy issue is control of data retention. As users collect data, they create an exhaustive database of their routines and locations. Collected over months or years, this data provides an intimate portrait of individuals' lives that could be subject to theft or subpoena. Allowing users to delete their data can prevent some of the privacy harms that stem from long-term retention [50, 51, 52].

PEIR retains aggregate calculations of impact and exposure indefinitely, so that users may compare their impact and exposure over months or years. But the system will default to deleting all location information after six months unless users specify otherwise. Users will be able to alter their profile preferences to retain location data for shorter or longer periods. Users can also delete specific routes or locations in their trip diaries. Deleting individual trips permanently deletes the associated GPS coordinates from the database. We are also exploring approaches to permanent deletion from backups [53]. PEIR does not recalculate aggregate impact and exposure, however, so the impact of deleted trips remains reflected in PEIR totals.

7. CONCLUSIONS

PEIR exemplifies an emerging class of adaptive, human-in-the-loop sensing systems that combine the distributed processing of the web with the personal reach of mobile technology to engage people in exploring the previously unobservable relationships of their actions to the world around them. PEIR automatically segments each user's location data into trips, and generates impact and exposures for these journeys. To help users build personal understanding of the estimates, the system also provides trends over time, access to intermediate calculations, and comparisons among users in the same Facebook social network. This paper outlined PEIR's existing architecture and implementation details, and discussed system lessons learned and enhancements already underway, for a version in use by thirty users over six months in 2008, collecting approximately four million GPS records.

PEIR has been an excellent experimental platform that motivated new algorithms to provide well performing activity classification, runtime model computation, and selective sharing. Future work on the PEIR system will focus on scalability, stability, performance, and usability. In order to scale the server to handle thousands of real time users and future application enhancements we must define more modular interfaces for incorporating data inputs and models. Future users will exist around the world, and so we must make it easier to integrate with local models and data sets relevant to PEIR inferences. As we do so we face a tension between greater inference accuracy, through the use of local air quality resources for example, and that of greater scalability, by keeping the inferences based on more broadly applicable models as inputs. Similarly, our activity classification must be extended to accommodate modalities common in other locations such as cycling, bus, train, and subway. Sustained usability of the system will be greatly enhanced through the introduction of goal setting and feedback. This calls for new features on both the server side and the handheld device.

8. ACKNOWLEDGMENTS

This work is supported by Nokia Research Center and National Science Foundation (CCR-0120778, CNS-0627084). Hardware is provided by Nokia and Sun Microsystems. Creating PEIR was a collaborative effort involving: Faisal Alquaddoomi, Betta Dawson, Hossein Falaki, Jeff Goldman, Richard Guy, August Joki, Isaac Kim, Joe Kim, Olmo Maldonado, Alex McElroy, Vinayak Naik, Nicolai Petersen, Jason Ryder, Alexis Steiner, Henry Tirri, Calvin Wang, Karen Weeks.

9. REFERENCES

[1] J. Burke, D. Estrin, M. Hansen, A. Parker, N. Ramanathan, S. Reddy, and M.B. Srivastava. Participatory sensing. In *ACM Sensys WSW Workshop*, 2006.

[2] Ecorio. http://ecorio.org, 2008.

[3] Carbondiem. http://carbonhero.com, 2008.

[4] Ubigreen. http://dub.washington.edu/ubigreen, 2008.

[5] E.Moretti J.Currie, S.DellaVigna and V.Pathania. The effect of fast food restaurants on obesity. Working Paper 14721, National Bureau of Economic Research, February 2009.

[6] CA. Pope III, M. Ezzati, DW. Dockery. Fine-Particulate Air Pollution and Life Expectancy in the United States. New England Journal of Medicine, 360(4):376–386, 2009.

[7] K. Shilton, J. Burke, D. Estrin, M. Hansen, and M. Srivastava. Participatory Privacy in Urban Sensing. *MODUS*, 2008.

[8] K. Shilton, J. Burke, D. Estrin, and M. Hansen. Privacy and Participation in Urban Sensing. *Interaction*, 21:235–272, 2008.

[9] K. Shilton, J. Burke, D. Estrin, M. Hansen, and M. Srivastava. Achieving Participatory Privacy Regulation: Guidelines for Urban Sensing. *CENS, UCLA*, 2008.

[10] Gpsbabel. http://gpsbabel.sourceforge.net, 2008.

[11] A. Joki, J. Burke, and D. Estrin. Campaignr-a framework for participatory data collection on mobile phones. Technical report, CENS, UCLA, 2007.

[12] Nokia nokoscope. http://alpha.nokoscope.com, 2008.

[13] Postgis. http://postgis.refractions.net, 2008.

[14] Wordpress. http://wordpress.org, 2008.

[15] Modest maps. http://modestmaps.com, 2008.

[16] D.J. Patterson, L. Liao, D. Fox, and H. Kautz. Inferring High-Level Behavior from Low-Level Sensors. *Lecture Notes in Computer Science*, pages 73–89, 2003.

[17] L. Liao, D.J. Patterson, D. Fox, and H. Kautz. Learning and inferring transportation routines. *Artificial Intelligence*, 171(5-6):311–331, 2007.

[18] L. Liao, D. Fox, and H. Kautz. Extracting Places and Activities from GPS Traces Using Hierarchical Conditional Random Fields. *The International Journal of Robotics Research*, 26(1):119, 2007.

[19] L. Liao, D. Fox, and H. Kautz. Location-based activity recognition using relational Markov networks. In *Proc. of the International Joint Conference on Artificial Intelligence (IJCAI)*, 2005.

[20] M.A. Quddus, W.Y. Ochieng, and R.B. Noland. Current map-matching algorithms for transport applications: State-of-the art and future research directions. *Transportation Research Part C*, 15(5):312–328, 2007.

[21] LR Rabiner. A tutorial on hidden Markov models and selected applications inspeech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[22] G. Rigoll, A. Kosmala, and S. Eickeler. High Performance Real-Time Gesture Recognition Using Hidden Markov Models. *LNCS*, pages 69–80, 1998.

[23] GD Forney Jr. The viterbi algorithm. *Proceedings of the IEEE*, 61(3):268–278, 1973.

[24] Google maps with my location. http://google.com/mobile/gmm/mylocation/, 2008.

[25] I. Anderson and H. Muller. Practical activity recognition using GSM data. Technical report, CSTR-06-016, Computer Science, University of Bristol, 2006.

[26] T. Sohn, A. Varshavsky, A. LaMarca, M.Y. Chen, T. Choudhury, I. Smith, S. Consolvo, J. Hightower, W.G. Griswold, and E. de Lara. Mobility Detection Using Everyday GSM Traces. *LNCS*, 4206:212, 2006.

[27] 64 million smart phones shipped worldwide in 2006. http://canalys.com/pr/2007/r2007024.htm, 2006.

[28] J.R. Quinlan. *C4. 5: Programs for Machine Learning*. Morgan Kaufmann, 1993.

[29] S. Consolvo, D.W. McDonald, T. Toscos, M.Y. Chen, J. Froehlich, B. Harrison, P. Klasnja, A. LaMarca, L. LeGrand, R. Libby, et al. Activity sensing in the wild: a field trial of ubifit garden. 2008.

[30] Y. Zheng, L. Liu, L. Wang, and X. Xie. Learning transportation mode from raw gps data for geographic applications on the web. 2008.

[31] S. Reddy, J. Burke, D. Estrin, M. Hansen, and M. Srivastava. Determining Transportation Mode On Mobile Phones.

[32] E. Miluzzo, N.D. Lane, S.B. Eisenman, and A.T. Campbell. CenceMe-Injecting Sensing Presence into Social Networking Applications. *LNCS*, 4793:1, 2007.

[33] M. Mun, D. Estrin, J. Burke, and M. Hansen. Parsimonious Mobility Classification using GSM and WiFi Traces, 2008.

[34] ESRI. ArcGIS 8. *ESRI, Redlands, CA*.

[35] La county department of public health facility list. http://publichealth.lacounty.gov/rating/, 2008.

[36] MF Barth, PA Miller, and AE MacDonald. MADIS: The Meteorological Assimilation Data Ingest System. In *Symp. on Observations, Data Assimilation, and Probabilistic Prediction*, pages 20–25, 2002.

[37] U.s. census bureau zip code tabulation areas. http://census.gov/geo/ZCTA/zcta.html, 2008.

[38] Southern california association of governments traffic flow model. http://scag.ca.gov, 2008.

[39] CAR. Board. EMFAC: Emissions Factor Model, 2000.

[40] C.J. Stone, M. Hansen, C. Kooperberg, and Y.K. Truong. Polynomial splines and their tensor products in extended linear modeling. *Ann. Statist*, 1997.

[41] R. Ihaka and R. Gentleman. R: A Language for Data Analysis and Graphics. *Journal of Computational Graphical Statistics*, 5:299–314, 1996.

[42] S. Amin, S. Andrews, S. Apte, J. Arnold, J. Ban, M. Benko, A.M. Bayen, B. Chiou, C. Claudel, C. Claudel, et al. Mobile century-using GPS mobile phones as traffic sensors: a field experiment. In *World congress on ITS*, pages 16–20, 2008.

[43] P. Ong, M. Graham, and D. Houston. Policy and Programmatic Importance of Spatial Alignment of Data Sources, 2006.

[44] D.J. Weitzner, H Abelson, T. Berners-Lee, J. Feigenbaum, J. Hendler, and G.J. Sussman. Information Accountability. *Comm. of the ACM*, 51(6):82–87, 2008.

[45] H. Nissenbaum. Protecting Privacy in an Information Age: The Problem of Privacy in Public. *Law and Philosophy*, 17(5):559–596, 1998.

[46] J.E. COHEN. Privacy, Visibility, Transparency, and Exposure. *Univ. of Chicago Law Review*, 75(1), 2008.

[47] L. Palen and P. Dourish. Unpacking Privacy for a Networked World. In *CHI'03*, 2003.

[48] R. Agrawal and R. Srikant. Privacy-preserving data mining. *ACM SIGMOD Record*, 29(2):439–450, 2000.

[49] J. Krumm. Inference Attacks on Location Tracks. *Lecture Notes in Computer Science*, 4480:127, 2007.

[50] L. Bannon. Forgetting as a feature, not a bug: the duality of memory and implications for ubiquitous computing. *CoDesign*, 2(1):3–15, 2006.

[51] J.F. Blanchette and D.G. Johnson. Data Retention and the Panoptic Society: The Social Benefits of Forgetfulness. *The Information Society*, 18(1):33–45, 2002.

[52] V. Mayer-Schoenberger. Useful Void: The Art of Forgetting in the Age of Ubiquitous Computing. 2007.

[53] R. Perlman. The Ephemerizer: Making Data Disappear. *Info. System Security*, 1(1):51–68, 2005.