Inferring Activities and Optimal Trips: Lessons From Singapore's National Science Experiment

Barnabé Monnot, Erik Wilhelm, Georgios Piliouras, Yuren Zhou, Daniel Dahlmeier, Hai Yun Lu and Wang Jin

Abstract The following paper presents three novel and efficient algorithms to tackle pressing questions asked by city planners as well as policy makers: Where are people starting and ending their trips? Which activities are people traveling to/from? Are they taking the most efficient route? In order to capture large-scale travel data, a novel sensor was developed by the Singapore University of Technology and Design together with industrial partners. Using computationally simple and scalable algorithms, we are able to understand the large amounts of data collected by the sensors and shed light on the three questions above.

Keywords Urban data · Large-scale experiment · Sensor data · Optimal routing · Data visualization

B. Monnot (✉) · G. Piliouras
Engineering Systems and Design, Singapore University of Technology
and Design, 8 Somapah Road, Singapore 487372, Singapore
e-mail: monnot_barnabe@mymail.sutd.edu.sg

G. Piliouras
e-mail: georgios.piliouras@gmail.com

E. Wilhelm · Y. Zhou
Engineering Product Development, Singapore University of Technology
and Design, 8 Somapah Road, Singapore 487372, Singapore
e-mail: erikwilhelm@sutd.edu.sg

Y. Zhou
e-mail: yuren_zhou@mymail.sutd.edu.sg

D. Dahlmeier · H.Y. Lu · W. Jin
SAP Singapore, 1 Create Way, #14-01/02, Singapore 138602, Singapore
e-mail: d.dahlmeier@sap.com

H.Y. Lu
e-mail: hai.yun.lu@sap.com

W. Jin
e-mail: crystal.wangjin@gmail.com

M.-A. Cardin et al. (eds.), *Complex Systems Design & Management Asia*,
Advances in Intelligent Systems and Computing 426,
DOI 10.1007/978-3-319-29643-2_19

1 Introduction

The study of movement and travel patterns for residents of mega-cities is an important and growing field. The results of this field of study are consumed by city planners, transport planners, and policy makers who must manage systems which are steadily growing in complexity as more people move into urban centers.

We look in this paper at three of the most common questions asked by these agents:

1. Where are people starting and ending their trips?
2. Which activities are people traveling to/from?
3. Are they taking the most efficient route?

To tackle these questions, we make use of data gathered by a novel sensor developed by the Singapore University of Technology and Design together with industrial partners. The main contributions described in this paper are three novel and efficient algorithms for accurately answering the above questions for very large data sets. Building origin/destination matrices have often previously been effectively accomplished with GIS methods based on GSM (cellular phone) data [1] but this approach is limited by the accuracy of the GSM localization data which is often imprecise. The algorithms presented here are computationally simple and therefore scalable. Using embedded systems for activity identification has previously been complex, memory intensive, and energy intensive; our algorithms reduce the task's complexity and accurately identify a small set of activities using little memory and energy [2]. Previous studies of route identification have the luxury of more plentiful geolocation data [3], whereas our algorithms are capable of studying route efficiency using sparse Wi-Fi geolocation data. The contributions described in this work should be useful for research and policy teams using smart-phone applications, GPS loggers, and 'Internet of Things' sensors alike to collect and study human mobility in urban centers.

2 Singapore's National Science Experiment

As part of a major Smart Nation initiative, a sensor (called the SENSg) was custom-designed by the Singapore University of Technology and Design (SUTD) together with the Delta Electronics Industrial Automation Business Group. The goal of the sensor and accompanying server infrastructure designed by the SUTD together with the A*Star IHPC group is to inspire and thereby motivate Singapore's students to pursue science and engineering related fields. The sensors gather temperature, relative humidity, light level, sound pressure level, atmospheric pressure, 9-degree of freedom motion data. Additionally, they possess a Wi-Fi radio which serves the dual purpose of scanning for MAC addresses which are used to localize the sensor nodes as well as to move sensor data to a back-end server. The sensors

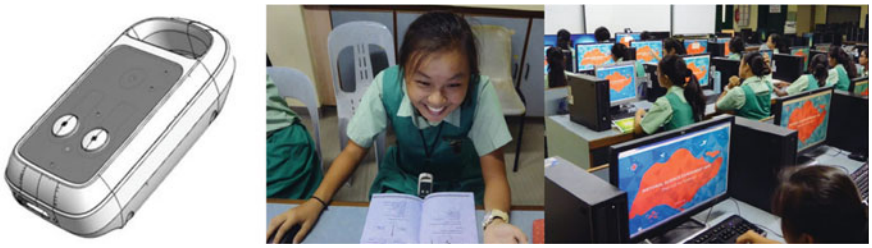


Fig. 1 The third generation SENSg device (*left*) which is worn by students on a lanyard (*center*) which allows them to explore their personal travel and environmental data (*right*)

will be deployed on a large scale from September until November 2015. The sensor shown in Fig. 1 is designed to acquire students step count, determine whether a device is located indoors or outdoors, as well as automatically identify travel mode (walking, riding a bus, riding a train etc.).

In order to validate the performance of the sensors before they are deployed, a series of three pilots were arranged at various primary and secondary schools in Singapore from June through July 2015. In total, 300 sensors were distributed and over 4.5 million unique sensor measurements were made. From this set, over 400 trips were able to be determined using the methods outlined in the remainder of this paper. The data shown in Fig. 2 are indicative of the types of data which can be acquired using the SENSg sensors. The localization of measurements is accomplished using the Skyhook API which returns geographical coordinates when it is sent a list of MAC addresses from Wi-Fi access points which were scanned by the device.

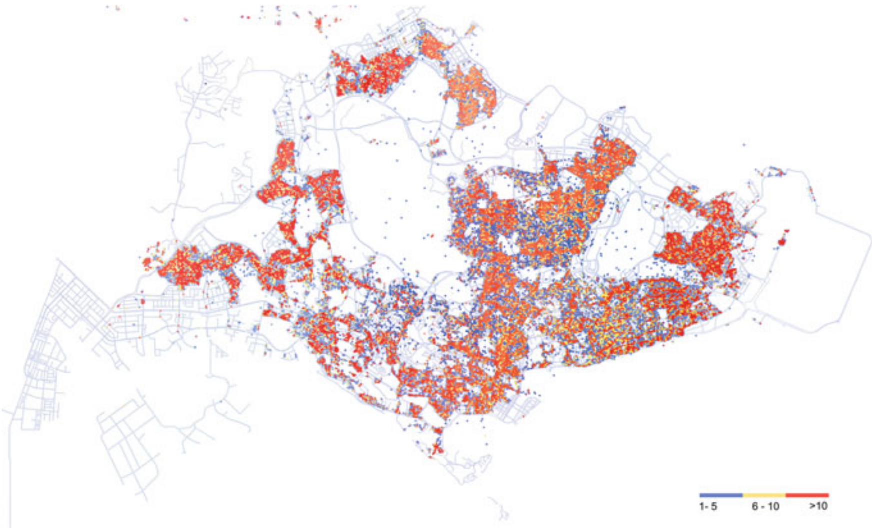


Fig. 2 Density of Wi-Fi network accesses recorded by the first phase of sensor deployment

3 State of the Art

3.1 Trip Inference

Inferring trips from geographical data has been tackled before in the literature. The first question is how to divide an agent's sequence of locations into individual trips?

Two approaches have been employed: active or passive trip reporting. In the active method, agents in the experimenting are tasked with providing the researchers a detailed account of their trips, including the nature of the trip. This allows the recuperation of verifiable information but has one major shortcoming: agents will be likely to under-report their trips, due to the cumbersome nature of logging their activities every time (a shortcoming noted by [4]).

Passive trip reporting offers a promising alternative that also scales better: only raw data such as the agent's latitude and longitude at a given moment is collected periodically; it is then the task of the researchers to process the data to break the sequence of positions into trips. The burden of verifying the agents' logs and building a database from it is avoided, which makes the method more suitable for large data sets.

Several studies have made use of passive trip reporting. In [5], the agents' cars are tracked. A trip is then defined as a sequence starting from the powering on of the car and ending when the contact is off (with additional criteria to filter out smaller trips or include stops during which the car is left on).

Schüssler and Axhausen [6] tracks the person's location instead of their car, which multiplies the post-processing difficulties due to the continuous nature of human movements. The researchers applied a smoothing method on the geographical data points to overcome the deficiencies of GPS-based data collection (listed in [7]). Trip inference is made by considering bundles of data points, i.e. logs that are geographically close to each other, and defining thresholds that discriminate between a stop within a trip or the endpoint of one. These thresholds can be based on the agent's speed (computed from the difference between two successive data points) or the time spent in the same bundle of points.

One of the most vexing problems facing researchers studying large-scale mobility patterns is the unwillingness of participants to spend their smart-phone battery energy for collecting location data from power-hungry GPS services. Several research groups (including this one) have proposed down-sampling location estimates as a solution [8, 9]. An alternative approach was taken in the design of the sensor system for the National Science Experiment which was described in detail in Sect. 2. For acquiring participant locations, the MAC addresses of surrounding Wi-Fi hotspots are scanned and recorded. This has been shown to be accurate to within 20–30 m and tends to be more effective while walking than driving [10].

3.2 Activity Matching

Trip inference and activity matching are two closely related problems. Once we have found out which trips the agent has taken, we want to know the nature of these trips. Again, we can ask the agents directly to provide information about the trips, but we expose ourselves to the same deficiencies noted before. We then seek an algorithmic way of assigning activity to a trip.

Stopher and FitzGerald [11] uses GPS data with additional information such as home and workplaces of the tracked agents to assign purpose to the trip. For moves that do not possess either or both as endpoints, they rely on land use data, i.e. geographical data giving the function of a particular piece of land (e.g., educational, hospital, road, park etc.) [12] uses both GIS land data and agent validation through a web application to confirm the activity.

In [13] theoretical frameworks are defined that allow for capturing the spatial and temporal phenomena of daily travel. Finally, a recent survey [14] focuses on the emerging trend of identifying semantically rich trajectories, instead of merely raw movement data, as the main object of interest in mobility studies. It describes a host of different techniques developed for addressing diverse issues such as creating trajectories from movement data, overlaying semantic information to trajectories and using data mining to extract higher level understanding of the characteristics of the trajectories.

4 Data Analysis

4.1 From Raw Data to Trip Recognition

4.1.1 Inferring Trips From Wi-Fi Data

The challenge of identifying travel mode from sparse Wi-Fi localization data was previously treated by this research group using a k-means clustering approach to identify home and school locations [15]. The main issue with applying clustering algorithms to this problem is that there are often substantially more points available for public spaces than for home locations leading choose bus interchanges and malls being identified as places where more time is spent, as opposed to true home locations. This paper takes an approach which considers the velocity of the person carrying the device as the primary indicator of the start or end of a trip. The dwell-time based algorithm presented here identifies static points in the localized coordinates by calculating an approximation of mean device velocity. It is less prone to mis-identification of home/school clusters because it filters for locations recorded in the late evening and during prime school hours to attempt to ensure that the POI's are correctly identified based on the probability that a point recorded at a given time corresponds to a student's location. The algorithm has the following steps:

Algorithm 1 Determining trips from dwell time

```
Require: Latitude, longitude, timestamps, stopped threshold $S_{thresh}$ e.g. 0.1m/s,
    dwell time threshold $D_{thresh}$ e.g. 240s, student at home e.g. 22h-5h, student
    at school e.g. 9h-12h
    Round latitude and longitude to 4 sig. dig. to ensure 10m maximum accuracy
    Load GIS data
    for point in lat/lon data do
        Calculate distance between each point using Haversine
        Calculate velocity between each point using $\dot{x} = \Delta$ x / $\Delta$ t
        Smooth velocity data using moving average filter (e.g. 3 windows)
        Count number of stopped points $N_{points}$
        if $\dot{x} < S_{thresh}$ and $N_{points} > D_{thresh}$ then
            Store lat/lon in list POI
    Find unique $POI$ in list POI
    Check if unique $POI$ occurred during home/school time in ts
    Assign $POI = POI_{home}$ and $POI = POI_{school}$ based on maximum occurrences
    of unique POI during home/school time
    Check that distance between $POI_{home}$ and $POI_{school}$ is >1000m
    Calculate distances between remaining unique POI
```

Source code for this algorithm in python and MATLAB can be downloaded here: SUTD Code Repository. Public sample data sets beyond the sample data contained in the repository may be requested.

When running the algorithm on a data set from a pilot test consisting of roughly 100 sensors deployed over a period of 2 days, Fig. 3 shows how the dwell-time algorithm is able to successfully identify approximately 50 % of the trips which were taken from home to school.

4.2 Matching POI's with Activities

From the previous trip analysis and temporal data, we can infer the students' homes and schools locations. The trips however may start or stop at different endpoints, such as a restaurant where students go to eat, shopping malls or commercial areas where they gather or parks and sports installation for recreational activities. We would like to be able to tell which of these activities, if any, is being undertaken by the student.

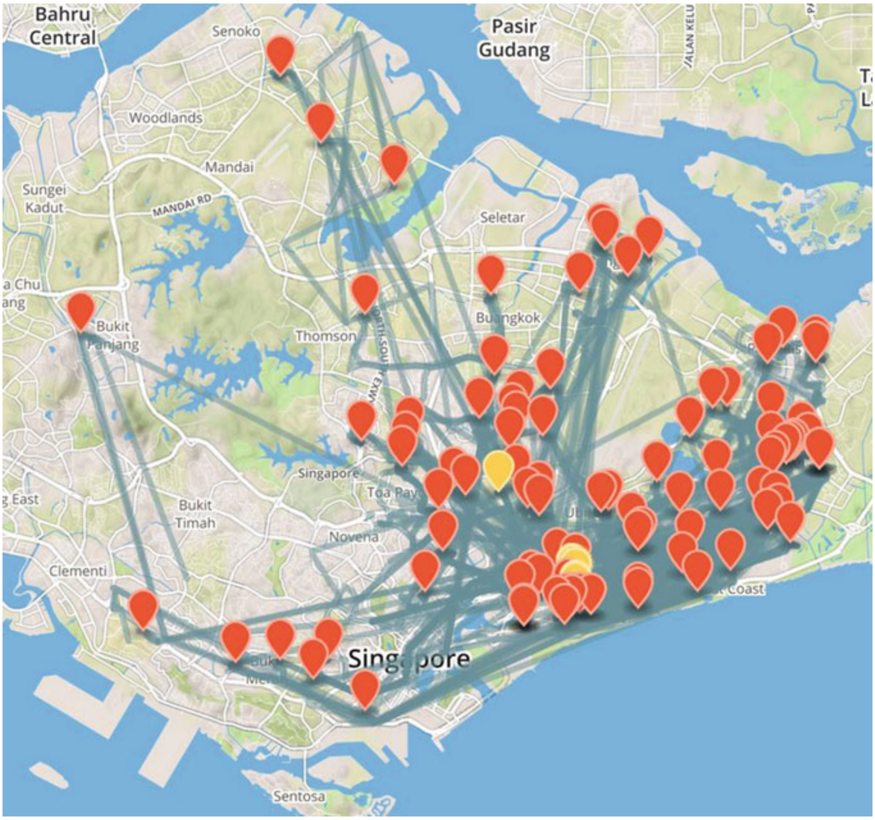


Fig. 3 Trips from home to school and back calculated using velocity dwell time (*red* home, *yellow* school)

To answer this, we obtain information from three different channels:

- Land use data
- Google Maps Places API data
- Feedback and validation from the students

The first part is to precisely identify the different POIs that the student is visiting. We define as POI an endpoint to a trip. Since we collect many of them in the first place, it is useful to identify those that are very close to one another and define them as one POI. This process reduces the redundancies.

If the trip analysis algorithm was successful, we have already identified the home as well as the school of the student. If not, it is possible to use a cruder method of counting the logs around the POIs, and decide that the one with the most logs will be the home and the one with the second most number of logs will be the school. For the other POIs, we proceed in the following way. Weights are assigned to each type of activity (food, commercial, recreation) and updated using the three channels of information. At the end of the process, we assign to the POI the activity that we find the most weighted.

The land use data gives us polygon representations of parks and sports installations in Singapore, which we can use to decide if a given POI is inside these polygons or not. This is a very strong clue to decide that the student is engaging in a recreational activity. However, Singapore is a very dense place that has many food outlets or stores even in parks (or the reverse: the Kallang mall has a climbing wall in the middle of the other shops). We therefore assign a consequent weight to the "recreation" activity if we find out that the POI is inside one of these parks and sports polygons yet not consequent enough to decide once and for all that the POI is recreational.

The Google Places API comes in handy to get more information about the places around a particular POI. We request the API a list of nearby places that are of a certain type (for our analysis, food, shopping and sports related places). The request returns this list with the location (latitude and longitude) and types of the places. We then update our weights once again with the results of this request. If the request has returned an item of, say, "shopping" type, we add some weight to the shopping activity, *inversely proportional to the distance to the POI*, so that closer places add more weight.

With these two channels we can already discriminate between the three activities. It follows a best guess approach that gets more confident with the number of results returned by the Google API, a technique that does not appear in the previous literature. We are able to assign the POIs with an activity and understand better the nature of the trips taken by the students. We hope a smaller amount of activities are classified as "Others" as our database of POIs and our student input grow.

To drive the point home we finally give the student the ability to cross-check the results of our analysis in the most painless way possible. Previous studies have underlined the difficulty of collecting the data from the participants in an active way, where they would be asked to provide themselves information about their trips. This usually leads to an under-reporting of the trips taken by the agent because of the cumbersome logging. By integrating in our interactive web-application a system that asks for the most uncertain POIs, say, "Is this where you get some food? Yes or No" with a visual map and position of the POI, we are able to correct deficiencies of the algorithm and even build a database of labeled examples that we can reuse in further studies. This is closer in spirit to the approach followed in [12].

We give in Algorithm 2 a pseudocode explaining how to assign activities. Source code for this algorithm in Python can be downloaded here: SUTD Code Repository.

Algorithm 2 Assigning activity to POIs

```
Require: Latitude, longitude of the nodes; GIS data of recreational surfaces
  (sports/parks).
  Load node data
  Load GIS data
  for node in node data do
      Get trips from node
      Get POIs from trips
      Eliminate redundant POIs
      for POI in obtained POIs do
          if POI was assigned home or school then
              Assign home or school to POI
          else
              if POI is inside recreational land then
                  Assign weight to recreational activity
              Load data from Google Places API
              for result in Places API response do
                  Assign weight to activity linked to result
              if Weights have been collected then
                  Decide POI type from most weighted activity
              else
                  Assign others to POI
      Build origin-destination matrix for node from assigned activities
```

5 Engaging in Activities

5.1 Activity Classification

The Singapore National Science Experiment will give access to unprecedented amounts of data collected over several days. The algorithms described before will process a large amount of it, yet they could often use a bit of guidance from the human to corroborate the results and gain valuable insights from them. The addition of a visual component to understand large data sets has been recognized in such fields as fraud detection (see [16] and the older but influential [17] for two different accounts of this). We therefore seek to empower our users, which comprise both the researchers attached to the project as well as the students carrying the sensors, by giving them access to visual representation of the data from different angles.

Earlier, we have touched upon the cross-validation made by the students participating in the experiment: after our algorithms have assigned activities to the POIs, we confirm these activities from simple questions answered by the students. This helps us build a database of the POIs which we can then use to ask ourselves the questions: How often do students take trips from school to sports? From a food place to home? Understanding these questions gives us a crucial insight in the behavior of our agents.

One way to represent this information is by using a chord diagram. The underlying data structure is a matrix which we call *origin-destination matrix*. The origin-destination matrix T is a square matrix of size the number of different activities we are able to differentiate (in our study, there are six: Home, School, Food, Commercial, Recreation and Others). $T_{i,j}$ captures the number of trips made from activity i to activity j. We compute for each student s a origin-destination matrix T^s and sum it all up to get the total trip information T of all our agents.

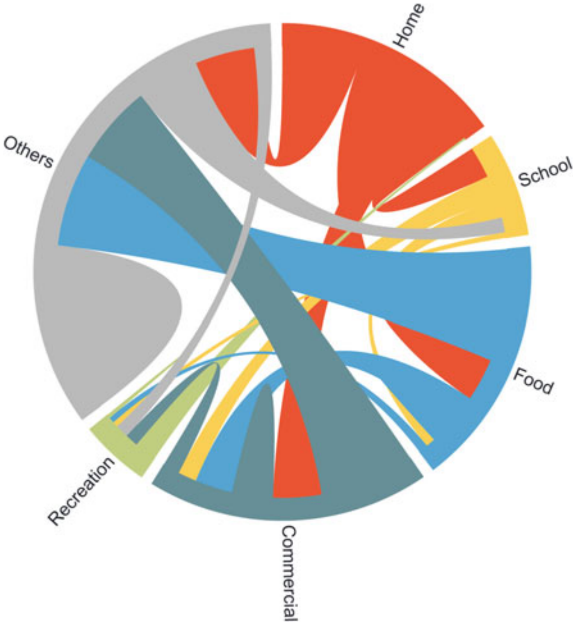
We give below an example of chord diagram obtained from the total origin-destination matrix of one pilot session. The chord diagram is generated in Javascript using the D3 library [18]:

Around the outer ring of Fig. 4 we have labels giving the different activities. This ring is divided in six unequal portions, each one with a different color corresponding to a particular activity (e.g. light green for Recreation). The size of the portion is proportional to the number of trip start points recognized as belonging to the corresponding activity, which for activity i is the number $\sum_j T_{i,j}$.

From the outer ring we then have multiple chords connecting one portion to another. They hold the information given by our $T_{i,j}$ coefficients. Since the diagram can quickly become cluttered, we give the user the ability to fade out parts of it to make it more readable. This is shown in Fig. 5.

In the example above, we hovered over the Commercial activity. Look at the chord joining this activity to the Others portion, of dark blue tint. The size of the

Fig. 4 Chord diagram from total origin-destination matrix



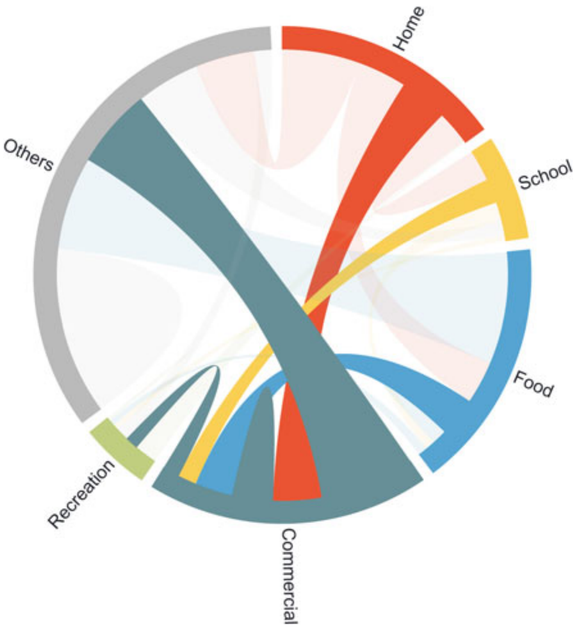


Fig. 5 Faded chord diagram from total origin-destination matrix

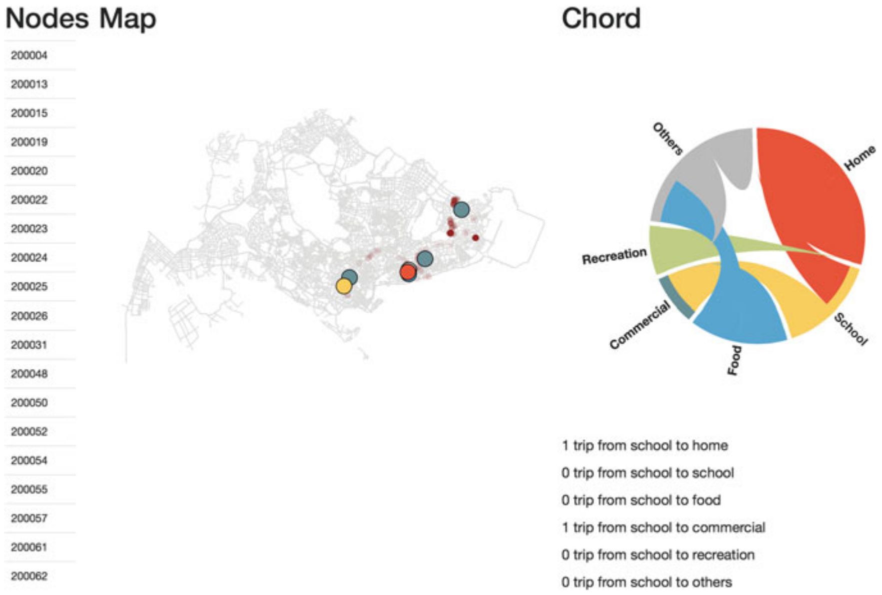
area covered by the chord on the Commercial portion is given by the number $T_{\text{commercial,others}}$, i.e. the number of trips going from a Commercial activity to one filed under “Others”. Reciprocally, the area covered by the end of the chord, in the “Others” portion of the ring, has a size corresponding to $T_{\text{others,commercial}}$.

The chord’s color is chosen to be the color of the activity that has the largest of the two chord’s endpoints attached to it. In the Commercial to Others chord, we can see that the Commercial activity has the largest endpoint, i.e. $T_{\text{commercial,others}} > T_{\text{others,commercial}}$, so we pick dark blue for the chord’s color.

Finally, for “power users” such as the researchers attached to the project, we build a complete web-app that gives access to each particular sensor data, completed by a map of Singapore with a log of the data points and the POIs (Fig. 6). To access the underlying data, we supplement the diagram with more information in a readable way, as shown below. As a portion is hovered over, we update the trips details on the right panel under the chord diagram.

On the far left is given the list of all the recorded users. Clicking on one plots the logs collected by the sensor on the map (in light red) and the POIs (filled with their assigned activity color and a black stroke). Hovering over a particular activity on the right-side chord diagram fades out the POIs not related to that activity to make the ones that are more prominent. The preliminary HTML/Javascript file used to create the visualization is given in our SUTD Code Repository.

We believe that this visual representation of the data allows a very quick understanding of the raw data logs collected from the students. A modified version of this application will be presented to the students for the “second opinion” that they are able to provide.



Nodes Map

200004
200013
200015
200019
200020
200022
200023
200024
200025
200026
200031
200048
200050
200052
200054
200055
200057
200061
200062

Chord

1 trip from school to home
0 trip from school to school
0 trip from school to food
1 trip from school to commercial
0 trip from school to recreation
0 trip from school to others

Fig. 6 Full application

5.2 Preliminary Results

Being able to precisely match activities with frequent locations visited by the students opens the door to many interesting results. The first and most evident one is to assess the quality of public policy measures. Singapore is known to be one of the greenest capitals in South East Asia (the National Parks Board motto even states “Make Singapore our garden”) and to provide a considerable amount of sports installations around the city. Collecting data about trips and activities from middle school students allows a quick evaluation of just how much these public facilities are being used. They also provide more back-story to it: do residents have to make long trips to get to a recreational activity, and which places are the least equipped? Supplemented with mode identification to learn the way of transportation chosen by the students to reach a certain location, we can infer more information on the accessibility of these installations.

Singapore also possesses an incredibly diversified food culture, ranging from very local “hawker centres” (inexpensive food courts) to higher end mall restaurants. We can again query the sensor data to understand the behavior of Singapore students when faced with so many different food options.

So far the experiment has been run three times with a little over 300 students, each time in one school. While it is difficult to precisely answer the questions above, it is still fruitful to look at the collected data. We built up the total origin-destination matrix from the third pilot run, which is given in Fig. 4. It can already be seen that trips involving recreational activities (in our analysis, engaging in a sport or being in a park) are fewer than those with a commercial or food-related purpose.

An interesting development will be to see how precise our analysis was when feedback is received from the students, which has not been done yet. It will be a further challenge to fine-tune our algorithm given these responses from the students.

6 Choosing the Best Route

6.1 Visualizing Individual Trips

We now turn our attention to the itineraries picked by the students to join two locations. We want to know whether their choice of route is optimal, i.e. if they went by the fastest possible way from one POI to another. Since these POIs are highly familiar locations for the students (such as their home, school…), it is expected that they would have already optimized these moves as much as they can. If discrepancies are to be found, we then need to look into other data which the sensors can give us: did they pick the coldest route? The most shaded one?

To compare their trips with some real data, we use the Google Directions API to return the best possible route between two endpoints, either by car, by public transports or on foot.

The trips are then analysed using two different tools. We extend our previous application to plot students' trips compared to the Google Directions results. Some results can be seen in Figs. 7 and 8. We also develop some algorithmic procedures to remove trips that do not fit their Directions result (possibly because of a misclassified trip or because of the noise in the sensor data) and compute some statistics on the remaining ones.

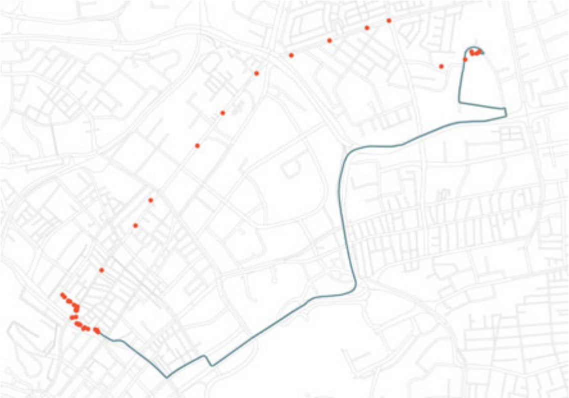


Fig. 7 Student's path (*orange dots*) and directions API result (*blue line*) for cars

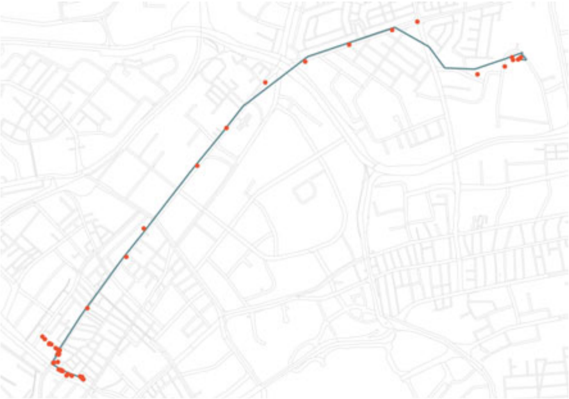


Fig. 8 Student's path (*orange dots*) and directions API result (*blue line*) for public transports

More specifically, we remove from our analysis trips that present the following characteristics:

- Trips that are too long (over an hour): these trips mostly appear due to noise in the location points that extend the trip beyond its true end.
- Trips that are not large enough (under 750 m): too short trips often get logged.
- Trips without enough data points (under 10): computing an accurate measure of the distance covered by the student is difficult if the data is too sparse.
- Trips that present too many gaps: again, we cannot have a precise distance measurement if logs disappear for too long. We use a metric based on the mean distance between two successive data points to remove these trips.

These criteria give us a coarser set of trips than that obtained for the previous activity detection analysis. One reason for that is that with no a priori on the nature of the trip, we could imagine a student going from his house to a nearby grocery store, which would be interesting to look at from the point of view of activity detection. However the best path analysis does not offer as much information for a very short trip.

In the Fig. 7 below, we plot one of the students' trip (orange dots) and compare it with the result from Google Directions (blue line). The figure is obtained by querying Google's API for the fastest way using a motorized vehicle. However, if we switch that parameter to return the fastest way by public transport, we obtain for the same trip a different picture that explains our data much better (see Fig. 8).

6.2 *Optimal Routes: Some Results*

The result presented in Figs. 7 and 8 has some interesting implications. The first and most obvious one is the possibility of using this kind of analysis for mode detection, another very active research question. Mode detection is tackled by our team using different methods, of which more information can be found in [15].

We can also apply this method to understand just how much more distance we have to make when we switch from a private mode of transportation to a public one. For example, we can make an informed public policy decision by deciding to add a new line of bus transportation in the places where taking one's private car is so much easier than using public transports.

This last part is better approached from an algorithmic perspective. The third pilot data set contains logs from over a hundred students, from which we identify about 300 trips. From these trips, we selected 51 of them which confirm to the set of criteria exposed in Sect. 6.1. We can then compare the distance covered by the students and the duration of their trips to the ones returned by the Directions API.

Figure 9 below charts the difference between the duration of a student's trip and the result returned by the Directions API. As we do not possess yet the mode of transportation used by the student, we select as a best guess the closest duration that the API returns to the student's. We observe in that case that most trips fall into a −20 to 20 % band of difference between the two durations. A finer analysis will be developed when the mode becomes available as an exterior measure computed from the data points.

We also count on the greater number of data points that will be collected in future runs of the experiments to erase local effects that we may be observing here, due to the fact that all students go to the same school. It will be interesting to see how the analysis will translate to more central or less dense parts of the city.

Finally, optimality also informs us on the congestion of the network for motorized trips. Maybe the user diverted from the fastest route as returned by the

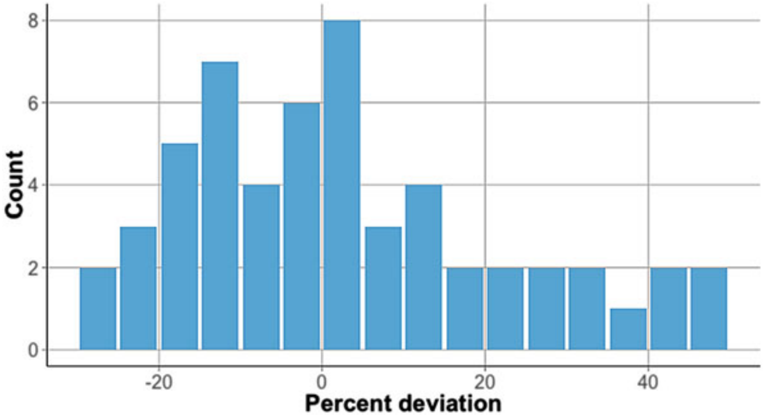


Fig. 9 Histogram showing the percentage of difference between the student's trip duration and the directions API's result

API (returned without using any information about the traffic at that time) because of the presence of many other users around him. We can then relate the making of the decision by the agent to the literature on congestion games with private information ([19] for the definition and models of congestion games). On the pedestrian level, the "cost" of taking a particular route could include such variables as the temperature (choosing to go to an air-conditioned indoors over a high temperature path outdoors), the light or noise levels (walking through a park instead of a residential area). These variables can be read in the data collected by the sensors and will be made use of in future experiments.

7 Future Work and Emerging Challenges

Singapore's National Science Experiment (NSE) opens up interesting possibilities about understanding the behavioral and social landscape of young Singaporeans. The project will involve more than 250,000 Singapore students over its 3 year lifespan who will be assigned easy-to-use devices that employ a number of sensors (location, motion, temperature, humidity as well as noise levels). Mining this data raises a number of interesting technical and analytical challenges that we briefly review next.

7.1 Big Data Algorithmic Techniques

Scaling up to high volumes of users creates new algorithmic challenges especially as each device broadcasts streams of data to a central server. Advanced algorithmic techniques such as streaming algorithms will be useful to deal with this massive data accumulation. These are types of algorithms that have limited memory available to them (significantly less than the input size) and moreover permit limited processing time per individual item but nevertheless can compute useful statistics or "sketches" of the data [20, 21]. More generally, sublinear time algorithms focus on computing functions of a target dataset while reading only a minuscule fraction of the whole input. A lot of recent work in computer science has focused on developing such techniques for a variety of algorithmic questions (see [22] and references therein). A particular notion of approximation which has been used to provide sublinear time algorithms for several problems is called property testing. These algorithms are used to decide if some object (e.g. a graph) has a "global" property (e.g., a graph is bipartite), or is far from having this property (e.g. the graph cannot be made bipartite even after removing a constant fraction of edges), while at the same time only using a small number of queries.

7.2 From Data to Stories: Understanding Real Life Social Dynamics and Networks

The ability to examine these diverse streams of sensor data gives rise to another interesting challenge. How can these data be woven together to create meaningful, high level semantic summaries about the true nature of the underlying social behavior? These data streams allow us to study two different types of correlations. The easier one is to combine the measurements of different sensors on the same device. This enables us to create more detailed and accurate descriptions about the individual behavior of each student. Critically, a different type of correlation is identifying correlations between users. In doing so, one can identify different types of groups that an individual belongs to along with the different group characteristics (e.g. size, dynamics, stability, etc.). This is a significantly more computational intensive task than the ones arising from focusing on individual users and applying sophisticated algorithmic techniques such as sublinear time algorithms will be necessary in order to deal with the emergent amount of data.

7.3 Privacy

Accessing detailed information about individuals' behavior naturally raises privacy concerns. Such privacy issues have been examined previously in the literature of semantic trajectories modeling and analysis [14]. As we move forward towards a full fledged realization of the vision of smart cities, deployed sensors will become increasingly present in the urban landscape. Any such progressive deployment should also be accompanied with the careful development of privacy preserving techniques that allow both for useful aggregation of data while minimizing the privacy loss for the individuals.

Acknowledgments This work was supported by the Singaporean National Research Foundation (NRF) and the SUTD International Design Center (IDC). Production of the sensors was possible due to strong support from Delta Electronics DRC, IABG, and Taoyuan Factory 2.

References

1. Caceres, N., Wideberg, J.P., Benitez, F.G.: Deriving origin destination data from a mobile phone network. Intell. Transp. Syst. IET **1**(1), 15–26 (2007)
2. Choudhury, T., Consolvo, S., Harrison, B., Hightower, J., Lamarca, A., LeGrand, L., Rahimi, A., Adam Rea, G., Bordello, B.H., et al.: The mobile sensing platform: an embedded activity recognition system. IEEE Pervasive Comput. **7**(2), 32–41 (2008)
3. Cottrill, C., Pereira, F., Zhao, F., Dias, I., Lim, H., Ben-Akiva, M., Zegras, P.: Future mobility survey: experience in developing a smartphone-based travel survey in Singapore. Transp. Res. Rec.: J. Transp. Res. Board **2354**, 59–67 (2013)

4. Du, J., Aultman-Hall, L.: Increasing the accuracy of trip rate information from passive multi-day gps travel datasets: automatic trip end identification issues. Transp. Res. Part A: Policy Pract. **41**(3), 220–232 (2007)
5. Axhausen, K.W., Schönfelder, S., Wolf, J., Oliveira, M., Samaga, U.: 80 weeks of gps-traces: approaches to enriching the trip information (2003)
6. Schüssler, N., Axhausen, K.W.: Identifying trips and activities and their characteristics from gps raw data without further information (2008)
7. Jun, J., Guensler, R., Ogle, J.: Smoothing methods to minimize impact of global positioning system random error on travel distance, speed, and acceleration profile estimates. Transp. Res. Rec.: J. Transp. Res. Board **1972**, 141–150 (2006)
8. Jariyasunant, J., Sengupta, R., Walker, J.: Overcoming battery life problems of smartphones when creating automated travel diaries. In: Proceedings of the 13th International Conference on Travel Behavior Research (2012)
9. Kumar, S., Paefgen, J., Wilhelm, E., Sarma, S.E.: Integrating on-board diagnostics speed data with sparse gps measurements for vehicle trajectory estimation. In: 2013 Proceedings of SICE Annual Conference (SICE), pp. 2302–2308. IEEE (2013)
10. Tsui, A.W., Lin, W.-C., Chen, W.-J., Huang, P., Chu, H.-H.: Accuracy performance analysis between war driving and war walking in metropolitan wi-fi localization. IEEE Trans. Mobile Comput. **9**(11), 1551–1562 (2010)
11. Stopher, P.R., FitzGerald, C.: Processing gps data from travel surveys
12. Bohte, W., Maat, K.: Deriving and validating trip purposes and travel modes for multi-day gps-based travel surveys: a large-scale application in the Netherlands. Transp. Res. Part C: Emerg. Technol. **17**(3), 285–297 (2009)
13. Schönfelder, S., Axhausen, K.W.: Urban Rhythms and Travel Behaviour: Spatial and Temporal Phenomena of Daily Travel. Ashgate Publishing Ltd. (2010)
14. Parent, C., Spaccapietra, S., Renso, C., Andrienko, G., Andrienko, N., B, V., Damiani, M.L., Gkoulalas-Divanis, A., Macedo, J., Pelekis, N., et al.: Semantic trajectories modeling and analysis. ACM Comput. Surv. (CSUR) **45**(4), 42 (2013)
15. Zhang, N., Kee, J., Loh, G., Tippenhauer, N., Wilhelm, E., Zhou, Y.: Sensg: large-scale deployment of wearable sensors for trip and transport mode logging. Submitted to Transportation Research Board Annual Meeting 2016 (2016)
16. Chang, R., Lee, A., Ghoniem, M., Kosara, R., Ribarsky, W., Yang, J., Suma, E., Ziemkiewicz, C., Kern, D., Sudjianto, A.: Scalable and interactive visual analysis of financial wire transactions for fraud detection. Inf. Vis. **7**(1), 63–76 (2008)
17. Cox, K.C., Eick, S.G., Wills, G.J., Brachman, R.J.: Brief application description; visual data mining: recognizing telephone calling fraud. Data Mining Knowl. Discov. **1**(2), 225–231 (1997)
18. Bostock, M., Ogievetsky, V., Heer, J.: D^3 data-driven documents. IEEE Trans. Vis. Comput. Graph. **17**(12), 2301–2309 (2011)
19. Nisan, N., Roughgarden, T., Tardos, E., Vazirani, V.V.: Algorithmic Game Theory, vol. 1. Cambridge University Press, Cambridge
20. Alon, N., Matias, Y., Szegedy, M.: The space complexity of approximating the frequency moments. In: Proceedings of the Twenty-Eighth Annual ACM Symposium on Theory of Computing, STOC ‘96, pp. 20–29. ACM, New York (1996)
21. Babcock, B., Babu, S., Datar, M., Motwani, R., Widom, J.: Models and issues in data stream systems. In: Proceedings of the Twenty-first ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, PODS ‘02, pp. 1–16. ACM, New York (2002)
22. Rubinfeld, R., Shapira, A.: Sublinear time algorithms. SIAM J. Discret. Math. **25**(4), 1562–1588 (2011)