Big Data in Mobile Crowd Sensing

Sean Tan Jun Yu
School of Computer Science and Engineering

Assoc Prof Luo Jun
School of Computer Science and Engineering

Abstract **-** Big Data is a valuable commodity today, and different methods of gathering and collecting data is increasingly available. However, without the proper platforms, infrastructure and systems of collecting data found in large organisations, current methods of obtaining real-world data, such as purchasing the data or web crawling, may not be cheap or time efficient. We examine the idea of Advertising as a Platform (AaaP) as an affordable, quick and simple way for smaller organisations and researchers to collect data for mobile crowdsensing.

Based on the research and recommendation of AaaP [1] by researchers from the University of Massachusetts Amherst, we further evaluate the cost efficiency, viability and effectiveness of the proposed method in greater depth. We gathered data on browsing information (website visited, language used, location, operating system used) of people from 6 different regions - Singapore, Southeast Asia, Australia, United States, United Kingdom and Brazil, using digital advertisements placed online. Visitors' browsing data will be extracted and recorded when they visit websites containing our advertisements. Depending on whether the user grants location permissions, we use either the HTML5 Geolocation API or an IP-to-Geo lookup service to determine the visitor's location, with the Geolocation API being the more accurate option. We compare the costs, quantity and quality of data as well as analyse and interpret the data obtained and its significance. We also show the limitations and promising aspects of the study and proposed idea.

Keywords - Digital Advertising, Mobile Crowdsensing

1 INTRODUCTION

Mobile devices are becoming more pervasive and easily available throughout the world. In 2018, there are 4.93 billion mobile phone users in the world, more than two thirds of the world's population, and this number is expected to grow further [2]. Digital advertising can be used to capitalise on this, as advertisements have the full capability of a browser through JavaScript and APIs, and can gather information on the user and browsing history when the user visits a website containing the advertisement.

It is estimated that the average internet user is served 11,250 ads a month [3], because of how affordable and easy to deploy digital advertisements are. They are priced at about or less than 0.0001 USD to show a single advertisement and obtain a record. Such data is already being collected by large organisations and platforms such as Facebook and Google, but is usually not made widely available as it is private and proprietary user data. Thus we test and evaluate this method of using digital advertisements as a way for individuals and everyday people to collect large amounts of data at a fairly reasonable price, as well as analyse the data for different patterns.

In total, we gathered 5322306 records from 426377 unique devices with a minimal infrastructure on Amazon Web Services, spending less than USD 800. To evaulate the difference in effectiveness and price in different regions, we expanded the experiment to different regions and compared the results. The regions are:

- Singapore
- Southeast Asia
- Australia
- United States
- United Kingdom
- Brazil

We also use 2 different APIs for getting a user's location: HTML5 Geolocation API when available and an IP-to-geolocation API hosted on geoplugin.net.

The experiments show the promise of the method in gathering large amounts of data cheaply, but there are technical challenges, including improving the quality of the location data and more accurately identifying repeat visitors.

2 BACKGROUND

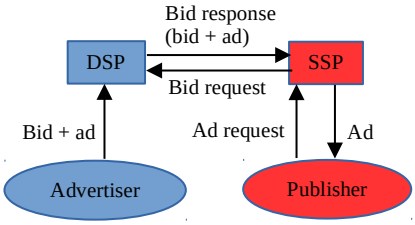

Figure 1: How SSPs and DSPs work

In a digital advertising platform, there are 2 parties involved: the publishers and the advertisers. The advertisers purchase the rights to show the advertisement on the publisher's website or platform, while the publishers sell the space on their websites to the advertisers. Supply Side Platforms (SSP) and Demand Side Platforms (DSP) perform the roles of the intermediaries Most of the transactions for digital advertising are made through Real Time Bidding

(RTB). When a user visits a publisher's website, the publisher will send a request to a SSP which will send the request to a DSP. The DSP will provide the advertisers' bids to the SSP, who will then give the space on the publisher's website to the highest bidder [4]. The highest bidder will pay the second highest bid plus 1 cent. We show this process in Figure 1.

Full service DSPs will rely on sales representatives and account managers that will help advertisers with the setting up of marketing campaigns and advise the customer through the process. However, using a full service DSP is costly, so we use the alternative, self service DSPs. Users of self service DSPs will have to set up the campaigns, creatives, marketing material and other details themselves, but at a much lower cost.

3 IMPLEMENTATION

We have constructed working infrastructure using nodeJS and Amazon Web Services (AWS).

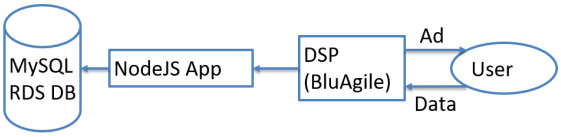

Figure 2: Architecture

An AWS MySQL Relational Database Service database was used to store the user information. This was connected to a NodeJS application created and hosted using AWS Elastic Beanstalk. A frontend script was uploaded on a DSP called BluAgile that is responsible for bidding for advertising space on the Internet and serving content to users. The frontend script that is embedded into the publisher's website will read data from the user's browser and send it to the NodeJS application which stores it in the MySQL database. The architecture is visualised in Figure 2. The following data is stored:

- Fingerprint (uses FingerprintJS2 library to identify unique users)
- Latitude
- Longitude
- Speed, Heading, Altitude (if HTML5 geolocation API is available)
- Website visited
- Browser language
- Operating System

When a user visits a website containing the advertisement, a POST request containing the browsing data will be sent to our NodeJS application, which updates the database with the incoming data. If the user continues to stay on the website, more data will be recorded and sent every 30 seconds until he leaves or refreshes the web page. This enables us to capture any possible changes in location or movement information, such as when the user is walking and using a phone, but also prevents our database from being flooded with duplicate data.

The cost of the advertisements was US$0.10 Cost Per Mille (CPM), which is the price of 1000 advertisements impressions.

There are 2 ways to obtain location information: HTML5 geolocation API and and IP-to-location lookup service. The ad will initially ask the user for permission to access the browser's location. If the user grants permission to the advertisement, HTML5 API is used which is more accurate. If permission is not given, the IP address lookup is used instead.

In order to uniquely identify a user, we use a JavaScript library called Fingerprintjs2, which takes in various browser information and creates a hash in order to differentiate between users. Such information includes timezone, user agent, screen resolution etc, and will be used to create a hash to be stored in our database.

We spent US$50 on each region of interest, and a different number of records was obtained for each region, because of varying market values of digital advertisements in different regions. The breakdown is shown in the next section. We are able to specify the region that the advertisements are targeted at, so we set up 6 separate advertising campaigns for each region and collected the data separately.

4 OBSERVATIONS AND FINDINGS

In this section, we show the analysis of the data and the insights we were able to obtain.

4.1 COST OF DATA

Table 3: Number of records obtained with US$50 spent, ranked by cost per record from lowest to highest

Rank	Region	Records Count	Cost per Record
1	Southeast Asia	2433838	0.000020544
2	United States	856651	0.000058367
3	Australia	570783	0.000087599
4	Singapore	542586	0.000092151
5	Brazil	519274	0.000096288
6	United Kingdom	399174	0.000125259
-	**Overall**	**5322306**	**0.000056367**

Table 4: Unique devices obtained with US$50 spent, ranked by cost per unique device from lowest to highest

Rank	Region	Unique Device Count	Cost per Unique Device
1	United States	279812	0.000178691
2	Southeast Asia	48815	0.001024275
3	Brazil	41124	0.001215835
4	United Kingdom	30085	0.001661958
5	Australia	16353	0.003057543
6	Singapore	10188	0.004907735
-	**Overall**	**426377**	**0.000703603**

For each region, a total of US$50 was spent on gathered data. The cost of a bid is the 2nd highest bid plus 1 cent. We can obtain the cost per record and cost per unique device in tables 3 and 4 respectively.

Each device is assigned a unique fingerprint using the library Fingerprintjs2, which pulls browser information and creates a unique hash based on the device.

We can see that although Southeast Asia has by far the lowest cost per record, the cost per unique device is much higher than its cost per record. This shows that advertisements in Southeast Asia may not have as much reach as the advertisements are viewed by a limited number of people. On the other hand, the United States has a higher cost per record than Southeast Asia, but its cost per unique device is nearly ten times lower than Southeast Asia. Thus, the best region to host advertisements in vary depending on the user's needs – if the user needs a large amount of data, Southeast Asia is the most cost efficient region to collect it, but if the user also needs variety and reach to different audiences, then the United States may serve his purposes the best.

4.2 BROWSING TRAFFIC BY HOUR FOR EACH REGION

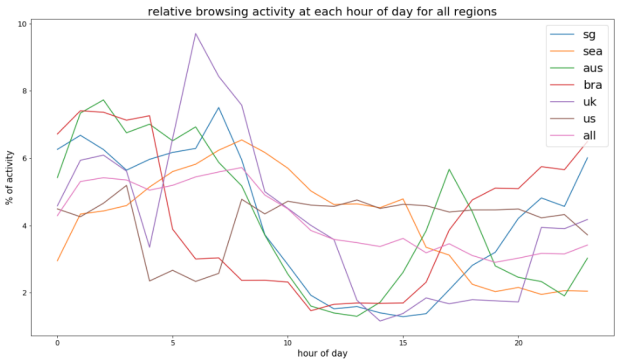

Figure 5: browsing activity of each region

The advertisements are placed on publisher's websites uniformly during the day, so the number of views on the advertisement at each hour is affected by the number of people online at the time. From figure 5, we can compare the differences in browsing traffic in different regions. The times in the graph are in each local region's timezone.

Table 6: Activity trends for each region

Region	Trend
Singapore	Activity is high in the morning. It peaks and starts to drop off at about 7am. Activity starts to pick up again at about 4pm and generally increases until the next morning.
Southeast Asia	Activity is highest from 5am to 3pm. It is much lower at night and in the early morning.
Australia	Activity is highest in the early morning before 5am, and there is also a surge in traffic at 4pm.
Brazil	Activity is high from 4pm to 5am the next morning.
United Kingdom	There is a significant surge in activity at 6 to 7am. 10% of activity comes during that time.
United States	Browsing activity in the United States is the most stable out of all the regions, which only a dip occurring between 3am to 8am.

Overall, it seems that for all regions studied, morning activity is consistently higher than afternoon and evening activity. This may be due to the nature of the publishers' websites (largely literature, movies, TV shows), rather than an indication of the actual browsing traffic at each time, and people are more likely to visit these websites at night or in the morning. The spread of types of websites will be studied in the following section.

4.3 TYPES OF WEBSITES VISITED

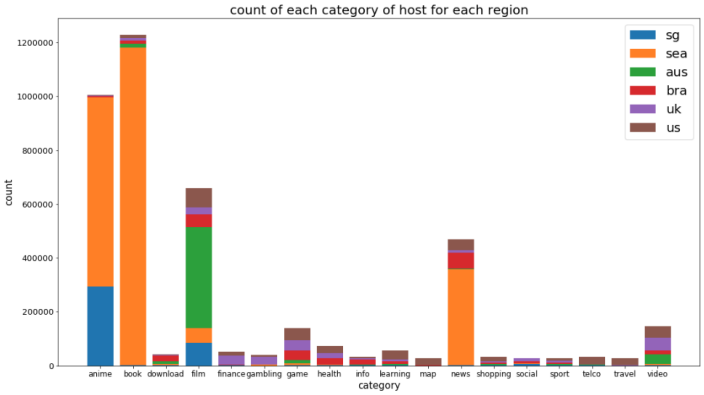

Figure 7: Approximate visit frequency for each type of website for each region

From figure 7, the most common types of website visited are literature and anime websites. This is because visitors from Southeast Asia visit those websites a lot, and because hosting ads in those countries is much cheaper, there are many records of visits on websites that Southeast Asians tend to visit.

host	counts	category
truyenfull.vn	754391	book
www.wuxiaworld.com	510451	anime
www18.soul-anime.us	421812	anime
sstruyen.com	396286	book
baykoreans.link	379371	film
www.freemalaysiatoday.com	355405	news
www.youtube.com	69644	video
fmovies.is	49331	film
www.diariodepernambuco.com.br	46414	news
i.mkvdz.com	43850	video

Figure 8: Top 10 Hosts Worldwide

Most of the traffic for Southeast Asia's book sites are because of 2 websites: truyenfull.vn and sstruyen.com, which are both Vietnamese websites, accounting for 754391 and 396286 records respectively. This possibly shows that Vietnamese speakers enjoy reading and literature in general.

We can also tell that like their Southeast Asian counterparts, Singaporeans also enjoy anime and manga. On the other hand, a significant proportion of Australians enjoy watching movies and television shows online. The rest of the regions are more varied in the types of websites that they visit, and their interests are spread out more evenly.

4.4 OPERATING SYSTEMS USED FOR BROWSING

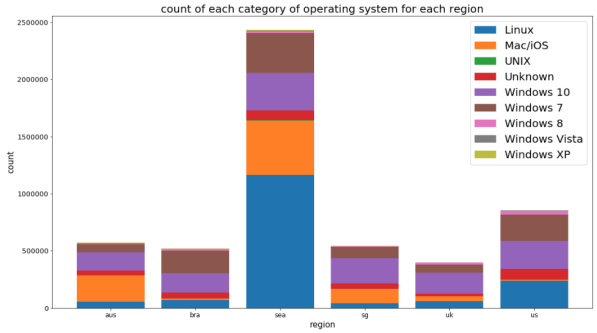

Figure 9: Approximate operating system distribution for each region

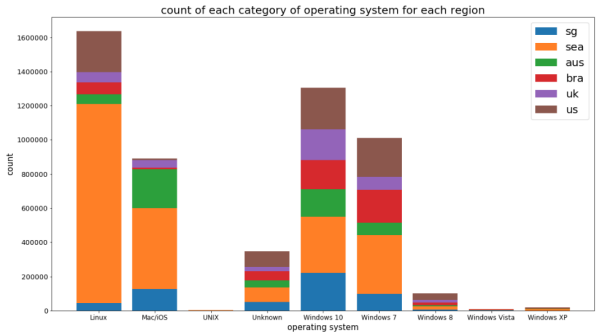

Figure 10: Regional distribution for each operating system

The most noticeable observation from figures 9 and 10 is that Linux, which is a proxy for the Android operating system, is the most popular operating system overall, at least for this study. Android is by far the most popular mobile operating system for web browsing in both Southeast Asia and the United States. On the other hand, Singapore and Australia gravitate slightly more towards Mac/iOS compared to the other regions.

Another interesting observation is the low adoption rate of Windows 8. The web traffic activity from Windows 8 is significantly less than that of its predecessor, Windows 7, and its successor, Windows 10. Also, almost half of the Windows users have not upgraded to Windows 10 yet.

4.5 NUMBER OF PEOPLE WHO GRANT HTML5 PERMISSION

When a user views a website containing our advertisement, a popup will appear, telling the user that our website wants to know the user's location. If a user grants us permission, we can access more accurate latitude and longitude information through the HTML5 Geolocation API, and we are also able to access additional variables – accuracy, altitude, speed and heading. If permission is not granted, then an IP to Geo service is used, which will have lower accuracy as IP addresses may be associated with a wrong region or a very large geographic region [5].

The following table shows the proportion of visitors that allow location permission from each region:

Table 11: Location permission grants from each region

Region	Allowed	Total	% Allowed
Singapore	80929	542586	14.9
Southeast Asia	24241	2433838	1.0
Australia	136501	570783	23.9
Brazil	12487	519274	2.4
United Kingdom	137770	399174	34.5
United States	106	856651	0.01
Overall	392034	5322306	7.4

We can observe that people in Singapore, Australia and the United Kingdom are more trusting and have a higher tendency to grant permissions. For researchers and individuals who plan to make use of advertisements to collect data may consider focusing on these regions as the quality and variety of location data that can be collected is likely to be superior to the other regions.

4.6 DISTRIBUTION OF POINTS

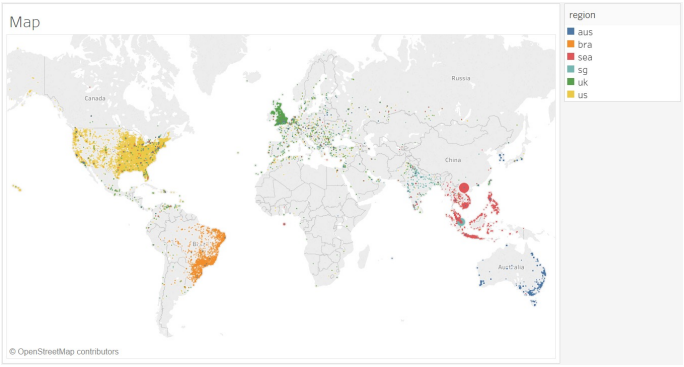

Figure 12: points associated to region of marketing campaign

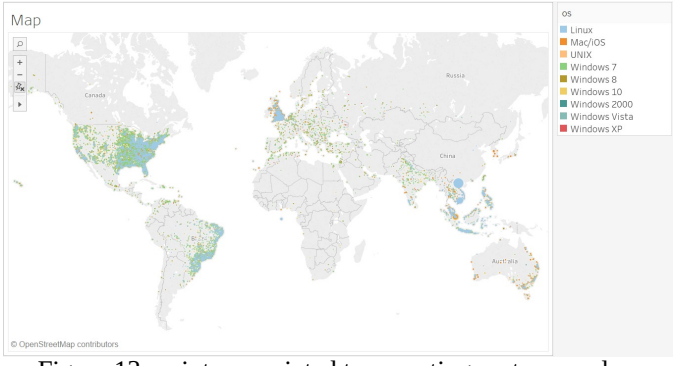

Figure 13: points associated to operating system used when advertisement was viewed

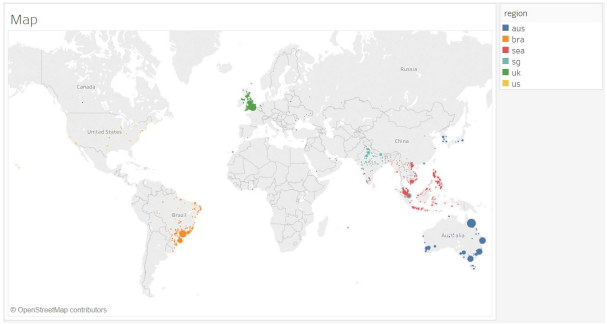

Figure 14: points that allowed location permission

We plotted all points on the world map to view the distribution of points collected.

Figure 12 shows the points associated to the region of the marketing campaign. The latitude and longitude values are captured when the user views the web page, and the associated region refers to the advertising campaign during which the point was captured. The campaign associated to the point may not match the location of the point (for example, if the region is United Kingdom but the point appears in the United States), and this may be due to the following reasons:

- The user was using a proxy or VPN, causing the latitude and longitude of his proxy server to appear instead.
- The DSP used was not able to effectively limit the marketing campaign to the region specified, so some advertisements leaked to non-targeted regions.

Figure 13 shows distribution of operating systems used when viewing the advertisement. We can observe a well spread out distribution of Android and Windows 7 across the world. Mac and iOS usage is largely centered in Australia, Singapore and parts of Southeast Asia.

In Figure 14, we show only the points where visitors allow location permission. We can see that the points are largely concentrated in Singapore, Australia and United Kingdom. Some places in Brazil also have a high number of such points, but these only make up a small portion of the total number of Brazil's points.

In the next section, we examine each region more closely to look for interesting patterns.

4.6.1 Southeast Asia

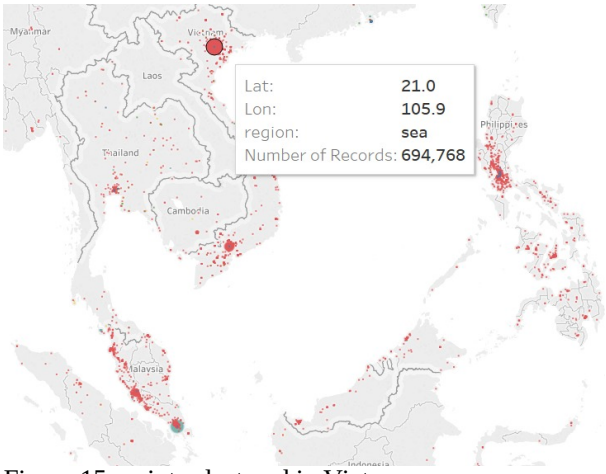

Figure 15: points clustered in Vietnam

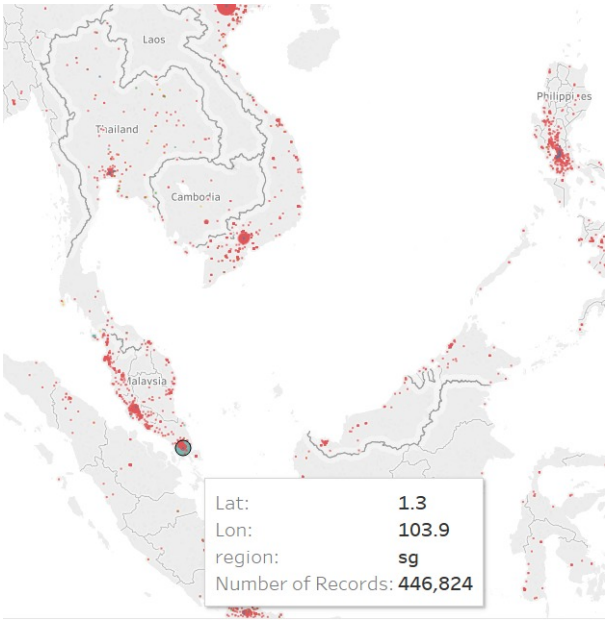

Figure 16: points clustered in Singapore

The points collected in Southeast Asia largely come from Singapore, Malaysia, Philippines and Indonesia. A strange observation is that the points in Vietnam and Singapore are concentrated in (21.0, 105.9) and (1.3, 103.9) respectively, with counts of 694768 and 446824. It is possible that it may be due to the large population concentration in those areas, but a more likely explanation is that the IP-to-Geo database lookup is not refined in those areas. As a result, any points that are in a large geographic area around those points will take on those values, as the IP-to-Geo table is not able to differentiate between the IP addresses in those areas.

4.6.2 Australia

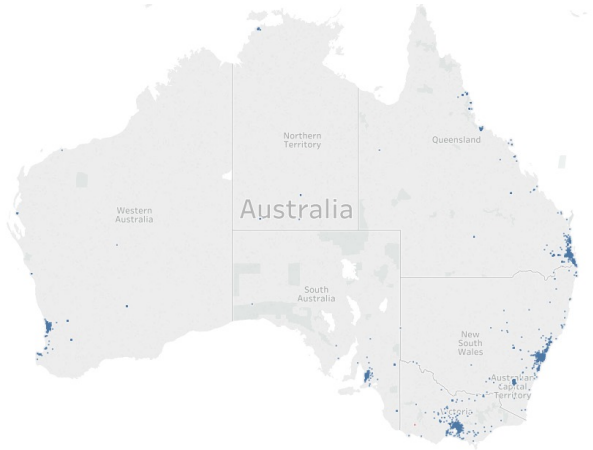

Figure 17: Points distribution in Australia

The points are clustered around high population coastal areas:

- Perth in Western Australia
- Adelaide in South Australia
- Melbourne in Victoria
- Sydney in New South Wales
- Brisbane in Queensland

There are very few points collected from the central areas in Australia, as central Australia is largely made of deserts and less inhabitable regions, so population count is lower.

4.6.3 United Kingdom

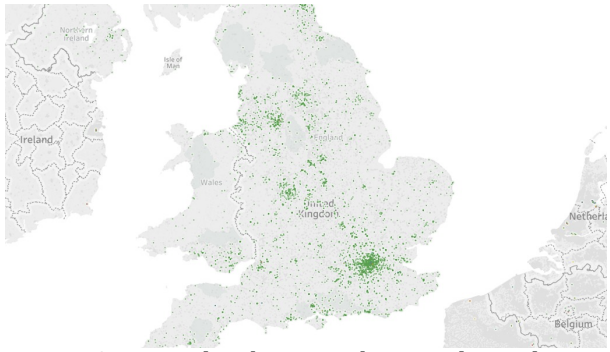

Figure 18: Points distribution in the United Kingdom

There is an even spread of points across England, and a lot of the points are from London and the surrounding regions.

There are fewer points in the other areas of the United Kingdom (Wales, Ireland, Scotland).

4.6.4 United States

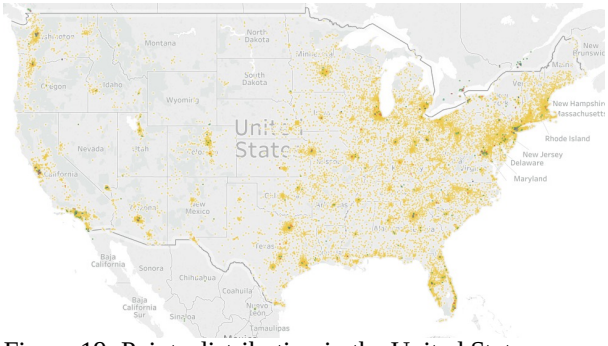

Figure 19: Points distribution in the United States

Most of the points are found in Western United States, especially close to the Northern part of the East Coast. There are also a small number of points in California, Oregon and Washington on the West Coast.

4.6.5 Brazil

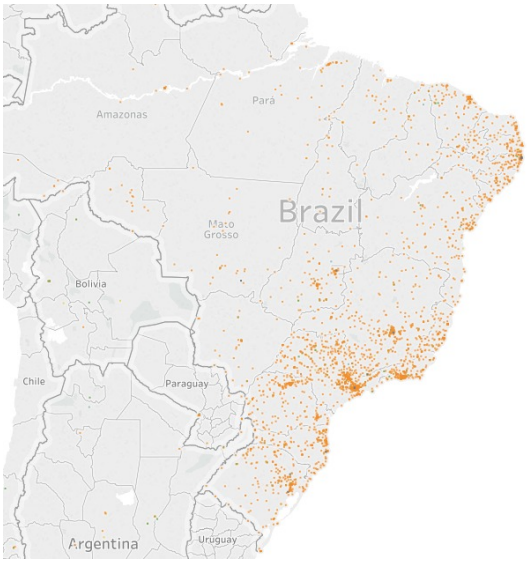

Figure 20: Points distribution in Brazil

Most of the points are found in the Eastern part of Brazil near the coastline. Many of the points are concentrated in Sao Paulo and Rio De Janeiro, 2 major cities in Brazil. There are much fewer points in Western Brazil as that is where the Amazon Rainforest is, and population density is lower.

5 CHALLENGES

We have previously examined the usefulness of using digital advertisements for gathering data and its accuracy. In this section, we examine the challenges that it faces and potentially stops it from being a reliable source of cost efficient information.

5.1 IDENTIFYING USERS

Initially, we wanted to use cookies, but cookies were found to be unreliable as people can clear cookies or browse privately to disable cookies. A study by ComScore found that about 3 in every 10 Internet users delete their cookies in a month, with an average deletion frequency of about 4 times per month [6]. Therefore we wanted to identify users based on something that changes less frequently.

We decided to use a library called Fingerprintjs2, which uses unique, long-term variables that change less frequently compared to cookies. These variables include screen resolution, language, operating system, etc. In an experiment to determine the uniqueness of browsers, it was found that 83.6% of browsers had an instantaneously unique fingerprint, and if Adobe Flash or a Java Virtual Machine was enabled, this number increases further to 94.2%. The study also showed that if we picked a browser at random, at best 1 in 286777 other browsers will share its fingerprint [7]. This enables us to uniquely identify users accurately and without as many restrictions as cookies. However, it is still impossible to tell anonymous visitors apart 100% of the time.

Thus, over short periods of time, where the user is unlikely to have a different fingerprint, this method of identifying users is fairly reliable at differentiating and identifying repeat visitors. However, over long periods of time, when it is more likely that the user changes a certain aspect of his browser that will affect the fingerprint, the reliability of this fingerprinting method may decrease.

5.2 LIMITED ACCURACY

As seen in section 4.6, IP-to-Geo services are lacking in accuracy, especially in certain countries. This is highlighted in Singapore and Vietnam, where 694768 and 446824 points are from the exact same latitude and longitude, due to ambiguity in the IP address that the IP-to-Geo service is unable to interpret.

Because of the limited accuracy of IP-to-Geo services in certain countries, and with the low rate of permissions for access to location being granted, this method of collecting data may be ineffective in some countries like Singapore and Vietnam, especially if location data (latitude and longitude) is crucial. Before making use of this method, users should ensure that IP-to-Geo services are reliable for their targeted regions.

5.3 INFORMATION THAT DOES NOT IN BELONG THE REGION

As seen in Figure 12, the region associated with the marketing campaigns and the actual location of the browsing traffic may not always match. As a result, unwanted or less valuable data may be collected, which is a waste of resources and money, as cost will be incurred and further effort may be required to remove these records. From figure 12, data from the United States may be collected even when the targeted region is specified as the United Kingdom, which is evident from the green points that appear in the United States.

This challenge requires further research, as we are unable to ascertain the reason for this phenomena. An attempt with other DSPs to check for such an overlap will help us to find out if the problem is DSP related. However, the number of these records is minimal and will not be an extremely large problem in terms of cost or effort, but it is a potential area of refinement.

6 CONCLUSION

The method yields interesting and useful information in terms of the locations and browsing habits of people, at a very affordable cost of about 0.000056367 per record. In countries like the United States, we are able to obtain fine grained, varied information in many different states with a comprehensive reach to many different devices. However, in other countries like Singapore and Vietnam, results are less ideal as IP-to-Geo lookup services are unable to accurately pinpoint visitors' locations using only IP address. It is important for potential users of the method to ensure that the method is effective in the region, such as through an initial small scale test, as well as use an appropriate DSP for the region, before implementing anything concrete. When well implemented in a region that is able to provide accurate data such as Brazil, Australia and the United States at a low cost, this method is effective in gathering data at scale.

ACKNOWLEDGMENT

We wish to acknowledge the funding support for this project from Nanyang Technological University under the Undergraduate Research Experience on CAmpus (URECA) programme.

REFERENCES

[1] Mark D. Corner, Brian N. Levine, Omar Ismail, and Angela Upreti, "Advertising-based Measurement: A Platform of 7 Billion Mobile Devices", n.d.; https://people.cs.umass.edu/~mcorner/papers/mobicom17.pdf)

[2] Statista.com, "Mobile phone users worldwide 2013-2019", n.d.; https://www.statista.com/statistics/274774/forecast-of-mobile-phone-users-worldwide/

[3] Christopher Elliott, "Yes, There Are Too Many Ads Online. Yes, You Can Stop Them. Here's How.", Feb 09 2017; https://www.huffingtonpost.com/entry/yes-there-are-too-many-ads-online-yes-you-can-stop_us_589b888de4b02bbb1816c297

[4] Igor Pavlov, "Real Time Bidding. What is RTB and how it works?", Jul 5 2017; https://medium.com/admachine/real-time-bidding-what-is-rtb-and-how-it-works-86463667df89

[5] Bradley Mitchell, "Does IP Address Location (Geolocation) Really Work?", Mar 18 2018; https://www.lifewire.com/does-ip-address-geolocation-really-work-818154

[6] Gian M. Fulgoni, "Cookie Deletion Rates and the Impact on Unique Visitor Counts", 17 Apr 2007; https://www.comscore.com/ita/Insights/Blog/Cookie-Deletion-Rates-and-the-Impact-on-Unique-Visitor-Counts

[7] Peter Eckersley, "How Unique is Your Web Browser?", n.d.; https://panopticlick.eff.org/static/browser-uniqueness.pdf