Revealing the Reflexive Phone Tap

A Comprehensive Study on User's Intuition in Mobile Applications

Pengyu Wang
New York University Abu Dhabi
Abu Dhabi, UAE
phillip.w@nyu.edu

Muhammad Anas
New York University Abu Dhabi
Abu Dhabi, UAE
muhammad.anas@nyu.edu

Tianshu Wang
New York University
New York City, USA
tw2198@nyu.edu

***Abstract*—In this project, we analyzed app usage behavior from three datasets with Hadoop related big data software to find out: the category information behind the prevalent mobile apps, and typical patterns of user app usage behavior. The results could be useful to a range of segments associated with mobile applications such as Operating System Kernels, App Developers, and further research studies. We found out, from two of the datasets, Message and Utility app category are the most popular. And in one dataset, Game and Utility are the most popular. We also found out that user tend to use apps from the same category in a sequence. We also proposed that mobile os developers can use this information to optimize the os resource management. However, further research is needed for this theory.**

***Index Terms*—app usage, analytics, mobile apps, user behaviour, app usage patterns**

I. Introduction

Mobile applications have substantially revolutionized our lifestyles. From communications to entertainment, from ecommerce to utility, applications are reshaping how humans interact with each other. Yet little is known about our intuition of interacting with these applications. In this project, we look into the app usage behavior patterns by analyzing three different data sources generated by real-world users. In the first section, we will first provide a preliminary analysis of the app usage datasets. We will then link those apps to their metadata in Google Play Store, in an effort to arrive at categorical conclusions about app-user interactions. In the third section, we further explore the cross-app relationships by drawing user paths in switching the apps. Our project can contribute to the field of usage pattern modeling by 1) integrating multiple data sources collected from various locations and 2) linking sequential app usage with their respective meta-information. This study aims to look into datasets with different demographics and derive insights on multiple subjects that could further corroborate the existing studies in the emerging research field.

II. Motivation

The field of app usage analytics is a relatively nascent field. While there are much data and information available about the app usage analytics, little is done to understand the motivation and typical usage pattern behind users' app usage behavior. If done correctly, this information could be extremely valuable to the entire network chain of entities associated with mobile applications.

Through this usage, we intend to provide an overarching information that analyses user's app usage behaviours. We also aim to understand the context that lead to user switching between different applications, and observe any overlapping trends across the population of users.

The insights that this research will produce could be benefit used by various professionals. Firstly, it provides comprehensive information to the app developers regarding the app usage behaviour patterns. They can use the insights for a range of understanding, from noting which apps and the category of apps are most popular to understanding entire network effects of their different products. They can use this to develop plans of market entry and study different app categories.

Similarly, for OS developers and device performance optimizers, these results can be useful in managing resources in order to enhance user performances. Moreover, we tend to serve the marketing and advertisement platforms an understanding and knowledge of the market demands related to different apps. Considering that online ads are the one of the most efficient marketing medium today, the findings from our work aim to help in enhancement of the reach of these online campaigns.

III. Related Work

There are existing works trying to associate the context of app usage with the action. A study used Point of Interests to infer the geological location of a user to find patterns of where do they open certain apps [2]. There are also a few studies that focused on the interdependence of different apps and attempted to recommend similar apps to users [1], [3], [5]. However, all of these works explored a specific problem that could have formed correlative insights with each other. Also, some of the datasets only comes from a single city or country, therefore could not be applied to other contexts. In this study, we aim to combine different datasets with demographics across the world to develop different analytics that aims to consolidate each others' insights.

IV. Dataset

In this section, we briefly introduce the three datasets which formed the basis of the study. All of them are collected from Android users with an objective to obtain a better understanding of users' interactions with the mobile applications.

A. *Frappe*

Frappe is a dataset collected for a real-world deployment of a context-aware mobile app recommender system collected around 2015 [1]. It is used to predict the which app will be opened given the environment and the status of users. The dataset is partitioned into two tables. *Frappe.csv* contains the mobile app usage records generated by users; each row represents a single usage of an app by a user identified by user id, and it also contains the environment and background information when the action occurred. The original dataset is comprised of 96,203 records generated by 1,000 users for 4,082 unique apps. The The original schema contains contextual information that includes

- Daytime
- Weekday
- IsWeekend
- Home/Work
- Cost
- Weather
- City

While these information pertains to predicting users' intention in opening the next app, our study focused on a smaller range of features in the log that fits our purpose of investigating the app usage itself. Its schema after cleaning and a row of sample data is shown in Table 1.

USER ID	APP ID	USAGE_COUNT	COUNTRY
0	0	1	United States

TABLE I
SCHEMA FOR *frappe.csv*

The *meta.csv* contains meta information about the apps crawled from Google Play Store. Each row uniquely identifies an app and all the relevant information about it. Below lists the columns that are in schema but discarded for the purpose of this analytic.

- Package Name
- Downloads
- Developer
- Icon
- Language
- Description
- Short Description

After cleaning, the *meta.csv* features the schema showcased in Table 2.

APP ID	NAME	CATEGORY	PRICE	RATING
0	AnyDo	Productivity	Free	4.5

TABLE II
SCHEMA FOR *meta.csv*

B. *LSApp*

LSApp is a dataset that provides information regarding the interactions and usages of mobile applications for a sample of users. It essentially provides for each user, the app they have interacted with, along with the type of interaction (open, closed, user interaction), as well as the timestamps and the categorized session for each user. The schema is shown in Table 2.

userId	**sessionId**	**timestamp**	**appName**	**eventType**

TABLE III
CAPTION

The data dates from July 2020, as part of a research study that allowed the recording of user app usages on their phones. It consists of 292 user participants, and provides 599,635 app usage records, which encompass 87 unique apps from the sample.

C. *ISTAS*

iSTAS, like LSAPP, is a dataset that was published in a research by scholars from the same group of institutions. It was collected with an App called uSearch which reccord user app usage behaviour in a period of time after conducting a search query[4]. However, unlike LSApp, this dataset was published in 2018 an the contains directly the sequence of user app usage sequence instead of a list of open, close, and interaction with time stamp. Furthermore, the dataset utilized JSON instead of csv or other relational dataset format to accommodate the sequence list mentioned above.

The dataset was collected from 255 participants who generated 6,877 search queries and app usage sequence data[4].

The schema of the dataset is shown in table IV.

timestamp	UserID	Query	App	AppUsages
2018-04...	225	photo...	pinterest	{’Duration’: {’a..

TABLE IV
SCHEMA FOR *istas.json*

V. ANALYTICS

A. *Data Cleaning & Profiling*

While all of the three datasets are developed for app usage analytic, each of them has different schemas and their own unique purpose under the realm. In order to derive joint analytic that is meaningful, we performed the data cleaning and profiling phase to ensure the goodness of analytics and the uniformity of the insights. The specific steps undertaken for each dataset differ and will be elaborated in the following sections.

a) Frappe: For processing Frappe, we employed two different tools for data cleaning and profiling, one is the Hadoop Map Reduce and the other is Apache Spark. Spark could directly read from Hive tables and then save to it after going through transformations. After the initial cleaning described in previous section, the resulting dataset has 957 unique users with 94,389 records. We elected not to drop any rows from the *meta.csv* even though some of the app information are largely unknown. This is because there is at least one entry in *frappe.csv* that is associated with any unique app in *meta.csv*.

b) iSTAS: The JSON file was first converted to csv file with Python script to make it compatible with MapReduce. The converted csv file has the same schema as the original file. The dataset was then cleaned with MapReduce to remove irrelevant and invalid data. Since the dataset already contains a sequence of app usage; timestamp, user Id, as well as Query is irrelevant in this study. Thus those fields are removed in the cleaning step. Meanwhile, the original dataset contains record with unknown app name and was marked with ”???”, which cannot be used to form conclusion in this study. Therefore those records are also removed. As a result, the cleaned dataset has 6761 records in total. The output file of this step is named with *cleaned.txt*.

The cleaned dataset has 6761 records and has the following schema (Table V):

App	App Usages
pinterest	{’Duration’: {’a..

TABLE V
SCHEMA FOR CLEANED *istas.json*

c) LSApp: The dataset provides essential information regarding app usage behaviours and trends across a sample of users. The listing of apps used in order and identified by distinct users seemed integral for our research. Hence, first the raw data file was loaded into the Hadoop File System, after which the extra columns were dropped and only the userId and appName column is kept through the first data cleaning using MapReduce. Also, consecutive appearance of the same app record with different eventType, for example multiple consecutive opening and closing of the same app by the same user, seemed to skew our results. Therefore, for the consecutively repeating app records, only a unique record is kept and the repeating ones are dropped from the dataset. The output of this data munging is added to a RDBMS in the Hadoop architecture for analytical queries. Using Hive Queries, basic data profiling is achieved which allows us to observe the counts of usage records, unique users, and unique apps.

While the dataset provides the user app interactions in order, it doesn’t provide shorter sub-sequences of these usages which we require in order to understand the trends of switching between apps for users. Hence, on the previous output, we perform another MapReduce job to profile our data for further queries. Here, for each user, we add all the permutations of 5 consecutive app usages in a field, and hence, we have the output that records the sequence of previous 2 apps, current app, and next 2 apps for each user. This file is also added to a SQL database in Hadoop for further queries.

B. Preliminary Analysis

After the data cleaning & profiling stage, we ran basic analytics on our datasets to obtain some initial understanding of the demographics of the user we are dealing with and the outstanding features of the datasets. All of our datasets are stored in Hadoop Distributed File System separately and are imported into Hive as external tables. From there, we performed Hive QL queries for the preliminary analysis.

1) Frappe: In this stage, we used Apache Hive and Spark to conduct some preliminary analysis on the dataset. For *frappe.csv*, we first analyzed the demographics of the user population participating the dataset. They comes from 80 different countries around the world, and 23,911 of them, the largest number of all countries, comes from United States of America. This does mean that we do not have sufficient users for certain countries to derive any statistically significant countries. Kuwait, Monaco, Andorra, Belarus, Thailand, Iran, Estonia, Latvia, Guatemala, and China all have less than or equal to 10 users. The user are mainly located in Western Europe, North and Central America, and a few countries in Asia.

There is much information to be extracted from *meta.csv* as a standalone analysis. Google Inc. is found to be the most prolific developer Furthermore, we counted the average number of unique apps that a user uses. Since the data is only collected for a duration of more than half an year, we could empirically interpret this data as the average number of frequently used apps for a single user, which is 19.69.

We also counted the number of apps belonging to each category. The results showed that 470 apps are labeled as Tools, 259 apps are labeled as Productivity, and 240 apps are labeled as Arcade & Action. Shopping, Books & References, and Education are the three categories with the least apps.

2) LSApp: Once we had the required schema for our analysis on LSApp dataset, a range of insights regarding app usage behaviours were found, leveraging the Hive QL. To understand an overview of usage patterns across our sample, we first found the overall most used apps across the users. The results of this are provided in the Fig. 3. The findings provide the insight that the most popular 10 apps are used more than the total usage of all other apps on the store. Further, we see that not only Google and Google Chrome are the most popular 2 apps, but they are about half as much used as all other apps (besides the first 10) on the store. The insights are particularly useful to understand the overall traffic on the mobile applications.

To understand the user market of each app, we delve further. We want to observe which apps are the most popular for each individual user, and then aggregate to find which apps are the highest used apps for most number of users. The results provided in Fig. 4 show that for how many (or percentage) of users, a given app is their most popular app individually. We can see that Google is the highest used app for 1 in 3 users, and ranking at third, Facebook is the highest used app for 7.2% of the users. The results provide a comprehensive way to understand the market of the app usages, as moving further from just the app usage traffics, it allows us to understand which app is the most used, or go to app, for most numbers of users. Hence, it provides an understanding of the market reach of the apps in terms of users.

3) iSTAS:

a) User demographic information: User demographic information was not included in the dataset file. However, as noted in the metadata file of this dataset, *README.md*, the participants of this study contains 59% female, 50% aged between 25 and 34 with various educational background[4]. This dataset doesn't contain any contry and location information.

b) App average usage time analyzation: We first want to figure out which app is the most popular by average usage time in this dataset with MapReduce. With the cleaned data mentioned above, the app usage sequence was segmented. After that, for each app name mentioned, the map code will output app name as key and usage time as integer value. The reduce code can then sum up and find the average usage time for each app name key. After that, the result was imported into hive database to find out the top 20 app name sorted from the most to the least by average usage time.

The result is visualized in Fig. 3.

We can then analyze the category of those apps by looking up information online. The result is shown in Fig.4.

We can then concatenate the category information by count. And have the result ranked from the most to the least: System-software, Game, Cashback-App, Utility-App, Web-browser, Cryptocurrency, and Reading. However, there're two points that should be noted: First, since system-software is considered necessary for mobile OS functioning and not typically produced by third party developers, we can easily explain the reason why the mentioning count in the dataset is high. Second, the cashback app, Yoolotto's average usage time, 63962539 ms, when converted to hours, is about 17.77 hours. This usage time is abnormal to ordinary users when considering each app usage sequence is recorded in a 24 hours period. Also, for cashback app, it not uncommon for users to setup "mobile farm" to run the app as long as possible to abuse the system to earn money. The similar situation also happens in other cashback app and inflated the average usage time for the whole category. Thus we should also ignore this category.

As a result, the most popular app categories are: Game, Utility-App, Web-browser, Cryptocurrency, and Reading.

c) App usage sequence popularity analyzation: The cleaned data gathered from the cleaning step is inputed into MapReduce program. Then, the mapper code will parse the app usage sequence and map all app usage sequence triplet slice as key and number 1 as value. Since the order of the sequence is critical for this dataset, the program will not list all combinations in the sequence. For example, if a app usage sequence is "A, B, C, D", the map code will output "A, B, C" and "B, C, D" as keys and 1s as values. Then the outputed key-value pair was feed into the reduce code and was summed up as "app sequence" + "sum".

After this step, the output data was then imported into hive. A query code was then executed to find out the most popular app usage sequence splice.

The result is shown in Fig.5. As we can see: 1. Google app has been mentionex multiple times in the output. And 2. Apps in the same category are often mentioned in the same sequence record.

C. App Category Analysis

App Category analysis is mainly completed on the Frappe dataset, since Frappe provided meta information about the app that is already linked with the user-generated records.

First, we ran a query to determine the most frequent apps being used across the datasets. In Frappe, we found out that Facebook, Gmail, and WhatsApp are the three most used app. Here we define most frequent app as the app being used for the most number of days for all users. However, if we change our metric to the most usage counts, WhatsApp, Facebook, and Chrome became the top three apps with the most number of user sessions in Frappe.

Secondly, we wanted to determine the most used app grouped by countries. This is because simply looking at the total lengths or counts of sessions does not reflect the geographical disparities in app usages and they are often heavily influenced by outliers. We found out that while WhatsApp remains the most popular communication app in most European countries, LINE is predominant in Asian regions like (Chinese) Taiwan, Japan, and Indonesia. In United States, Gmail is the most used app 1,594 associating records. On the browser front, while Chrome is the most used app in Canada and Australia, Firefox is most popular in Lithuania. We did not include the statistics of all the countries since there were not sufficient records to statistically declare the popularity of an app in certain countries.

Moreover, we computed the number of usages in each category and their average ratings. Communication, with 25,149 overall counts in the dataset, has the most number of usages. Social Media, Tools, News & Magazines, and Productivity are the second to fifth most used categories on Google Play respectively. Regarding the ratings, 4.21 is the benchmark, that is the average rating of all apps. If we look into the average ratings by categories, Brain Puzzle has the highest rating of 4.60, due to the relatively smaller number of apps belonging to this category. Comics, Cards & Casino, Arcade & Action, Books & Reference are the second to fifth highest rated categories. Among the top 10 highest rated categories, 4 of them generally categorized to self-improvement, and 3 of them belongs to entertainment.

The App Category Analysis provided insights further than the names of the trending apps. By joining the two tables, we were able to uncover users' preferences across different categories, geographic locations, languages, ratings, and prices of the apps.

D. Usage Sequence Analysis

In order to understand the user behaviours, trends and the contexts under which they switch between different apps, we need to derive information from the sequence of apps that user use, from the order of apps in the data. Using this sequence, we can produce insights on understanding the switching between different apps for the users. Two of the data sets allow us to do that as follows:

1) LSApp: From MapReduce in data munging stage, we now have a database table containing each sequences of previous two, current, and next two apps for each users. Using this for all users, we use Hive Queries to observe which of these sequences are the most popular across users. Table VI shows the output of the 10 most common sequences of app usages for our sample.

The results appear to be to confirm the intuitive hypothesis that a user is more likely to switch amongst the same platform of apps. The switching of apps between Google and Facebook's app products are the most recurring across users. Considering the integration of services and how different apps from these companies complement each other, this isn't surprising, and rather, emphasizes on the network effects that these products create. Moreover, we see that besides the same company products, users are more likely to switch between apps from the same categories, and the most dominant of these are the communication apps. This probably highlights that user behaviours are directed by the particular motive or objective they have for mobile usage in a particular session. If not for an objective, it might perhaps be a user tendency to switch between similar applications. For either of the reasons, user are more likely to move between apps that are similar and belong to the same categories.

2) iSTAS: This dataset provides a similar conclusion from LSAPP. From the result gathered in iSTAS/App usage sequence (Fig.7), we can see that apps in the same category are often shown in a sequence. For example, message_service, messages, and messenger (Communication); as well as calculator, calendar, and calendar_storage (Utilities).

VI. Conclusion

We found in Frappe and LSApp that Social Media, Emailing, and Browser apps are the most popular apps for Android users arosss the globe. While Games, Utilities, and Web browser are the top 3 categories in iSTAS. In Frappe, we found out that Tools has the most competitive category while maintained a relatively high rating. It may be beneficial for third party app developers to work on apps that belong to those categories.

Furthermore, in iSTAS, we have learned that the user tend to use Apps in the same category in a sequence. Thus when implementing optimization algorithms for a mobile operating system, it may be more efficient for the system to prepare the resource for the app the user is currently using. However, further study is needed to verify this theory.

Additionally, we found out that there is a discrepancy of the prevailing apps in different geographic locations as the some mobile apps dominate market shares in certain countries.

Visual Representations and Insights

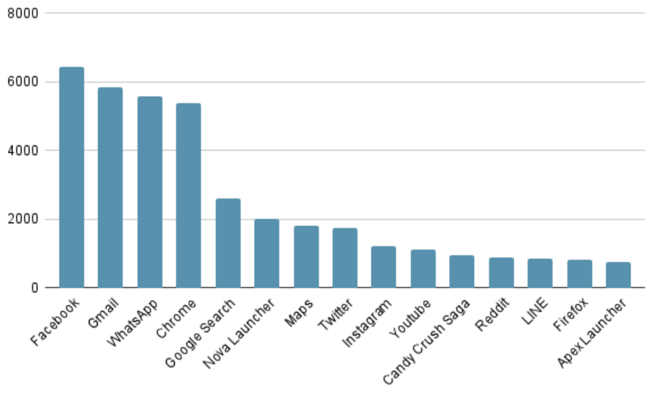


Fig. 1. Most Used Apps in Frappe

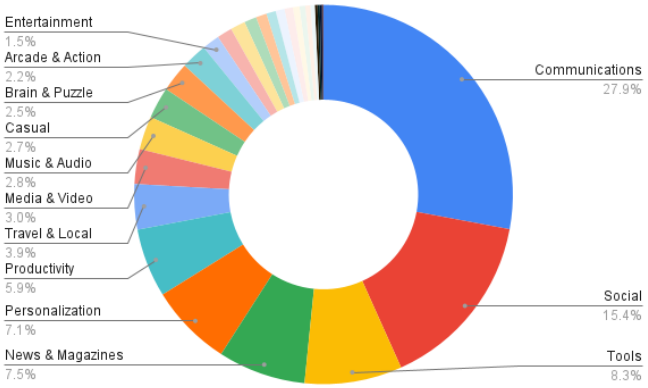


Fig. 2. Percentage Distribution of Apps on Google Play

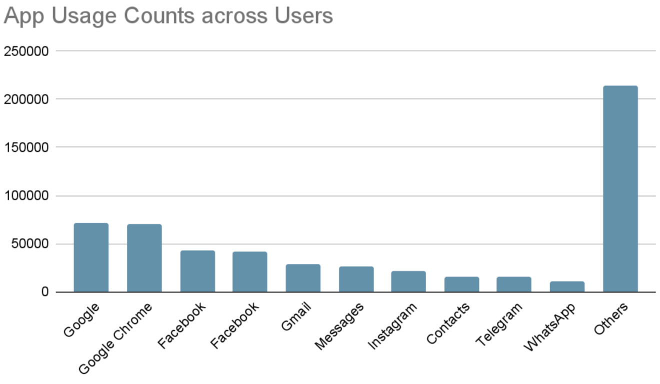


Fig. 3. App usage counts in LSApp

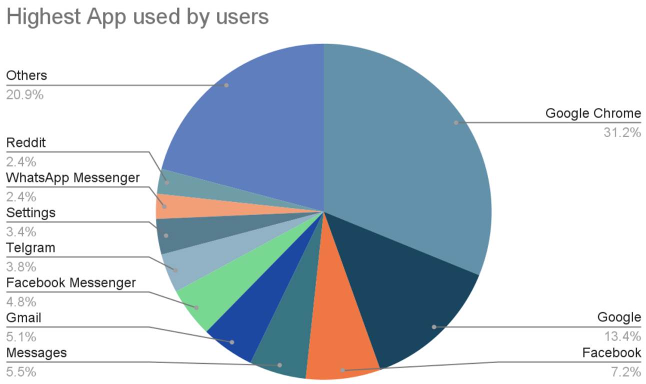


Fig. 4. Highest Used App for Each User

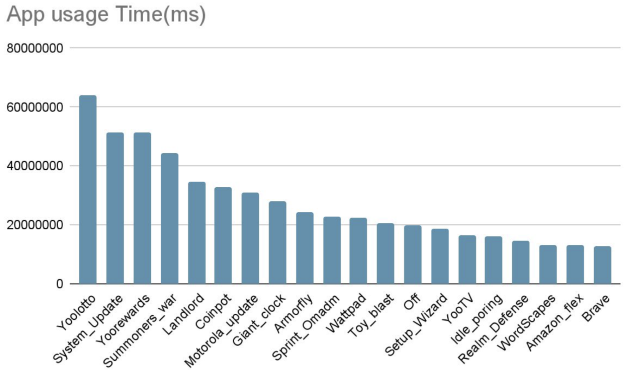
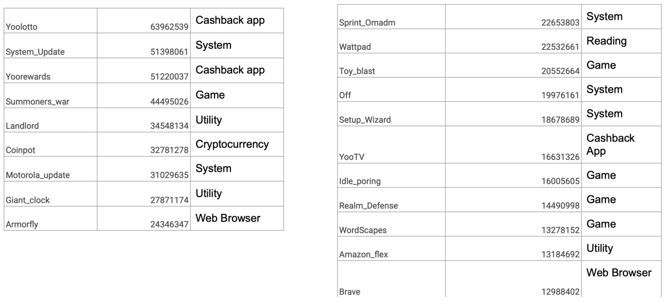


Fig. 5. App usage time in iSTAS



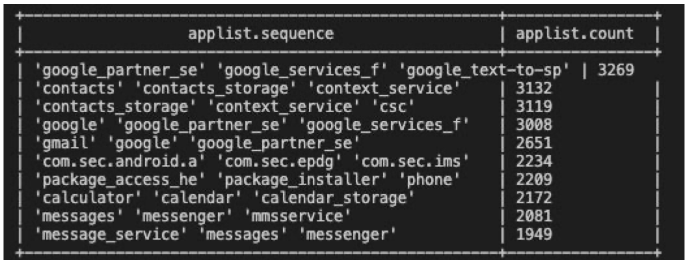
Yoolotto	63962539	Cashback app
System_Update	51398061	System
Yoorewards	51220037	Cashback app
Summoners_war	44495026	Game
Landlord	34548134	Utility
Coinpot	32781278	Cryptocurrency
Motorola_update	31029635	System
Giant_clock	27871174	Utility
Armorfly	24346347	Web Browser

Sprint_Omadm	22653803	System
Wattpad	22532661	Reading
Toy_blast	20552664	Game
Off	19976161	System
Setup_Wizard	18678689	System
YooTV	16631326	Cashback App
Idle_poring	16005605	Game
Realm_Defense	14490998	Game
WordScapes	13278152	Game
Amazon_flex	13184692	Utility
Brave	12988402	Web Browser

Fig. 6. Top 20 app category information in iSTAS

appSequence	sequenceCount
Google, Google Chrome, Google, Google Chrome, Google	17171
Google Chrome, Google, Google Chrome, Google, Google Chrome	16613
Facebook, Facebook Mr, Facebook, Facebook Mr, Facebook	7524
Facebook Mr, Facebook, Facebook Mr, Facebook, Facebook Msngr	7186
Google Chrome, Gmail, Google Chrome, Gmail, Google Chrome	4658
Gmail, Google Chrome, Gmail, Google Chrome, Gmail	4650
Google, Instagram, Google, Instagram, Google	4577
Google, Facebook Mr, Google, Facebook Mr, Google	4480
Facebook Msngr, Google, Facebook Mr, Google, Facebook Mr	4181
Instagram, Google, Instagram, Google, Instagram	4161

TABLE VI
App Usage Sequences for LSApp



```
+------------------------------------------------------------+-----------------+
|                     applist.sequence                       | applist.count   |
+------------------------------------------------------------+-----------------+
| 'google_partner_se' 'google_services_f' 'google_text-to-sp' | 3269            
| 'contacts' 'contacts_storage' 'context_service'            | 3132            |
| 'contacts_storage' 'context_service' 'csc'                 | 3119            |
| 'google' 'google_partner_se' 'google_services_f'           | 3008            |
| 'gmail' 'google' 'google_partner_se'                       | 2651            |
| 'com.sec.android.a' 'com.sec.epdg' 'com.sec.ims'           | 2234            |
| 'package_access_he' 'package_installer' 'phone'            | 2209            |
| 'calculator' 'calendar' 'calendar_storage'                 | 2172            |
| 'messages' 'messenger' 'mmsservice'                        | 2081            |
| 'message_service' 'messages' 'messenger'                   | 1949            |
+------------------------------------------------------------+-----------------+
```

Fig. 7. Top 10 App usage sequence slice by count in iSTAS

Acknowledgments

Acknowledgements to Professor Ann Malavet for NYU Peel HPC Team. Thanks to Baltrunas et al. and Aliannejadi et al. for providing the three datasets.

Appendix
Availability of Code & Data

All the code for processing and analyzing the datasets are available at https://github.com/cswpy/app_usage_analytics.

Frappe is available at https://www.baltrunas.info/context-aware.

LSApp is available at https://github.com/aliannejadi/LSApp.

iSTAS is available at https://github.com/aliannejadi/istas.

References

[1] Baltrunas, Linas, et al. ”Frappe: Understanding the usage and perception of mobile app recommendations in-the-wild.” arXiv preprint arXiv:1505.03014 (2015).

[2] Yu, Donghan, et al. ”Smartphone app usage prediction using points of interest.” Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies 1.4 (2018): 1-21.

[3] Huang, Jiaxin, et al. ”On the understanding of interdependency of mobile app usage.” 2017 IEEE 14th International Conference on Mobile Ad Hoc and Sensor Systems (MASS). IEEE, 2017.

[4] Aliannejadi, Mohammad, et al. ”In situ and context-aware target apps selection for unified mobile search.” Proceedings of the 27th ACM International Conference on Information and Knowledge Management. 2018.

[5] Aliannejadi, Mohammad, et al. ”Context-aware Target Apps Selection and Recommendation for Enhancing Personal Mobile Assistants.” ACM Transactions on Information Systems (TOIS) 39.3 (2021): 1-30.