– Independent Work Report Fall, 2015 –

Understanding Pricing Effects on Mobile Data Usage

Ava Chen ’16

Advisor: Nick Feamster

One-Semester Project

Abstract

This paper performs an exploratory study on mobile usage patterns in the United States versus South Africa over a temporal period of three years (2013-2015). Using data collected from the MySpeedTest application, we analyze differences in usage behavior for the top 5 applications in terms of total data usage in each country, comparing usage on different connection types (Wi-Fi vs. cellular) as well as for devices on different data plans (unlimited vs. limited monthly data cap vs. prepaid). The study also attempts a deeper analysis into the behavioral effects of mobile pricing practices such as zero-rating, which is when a network carrier does not charge customers for cellular data used by certain services or applications. Our findings show that US users slightly prefer cellular connections to Wi-Fi connections for most of the US top 5 most used applications, while South African users generally prefer Wi-Fi connections (with the notable exception of Facebook). Further, US users on unlimited and limited plans display much higher average monthly usage than those on prepaid plans, while South African users on prepaid plans generally display much higher usage than those on unlimited and limited plans. Although we observed increased mobile data usage for certain applications during and after zero-rating periods, a lack of sufficient data rendered many of these results inconclusive. In addition, data insufficiencies prevented a conclusive study on the behavioral effects of changing data prices for users on prepaid plans. Thus, we urge the importance of recruiting users to download and use MySpeedTest, especially focusing on South African users on prepaid plans.

1. Introduction

In recent years, there has been a rather heated ongoing debate regarding whether offering different pricing plans, such as zero-rated services and applications, might slant user behavior toward certain content on the Internet [9]. Our study is arguably the first that offers real data to address this research question.

The motivation behind this research is to perform an exploratory analysis on the effects of various features on mobile data usage. Our goal is to study user behavior in the context of these features, specifically zooming in on behavioral discrepancies between users on different data plans and connection types, as well as taking into account pricing effects such as zero-rating and prepaid data pricing.

This is interesting because obtaining insights into the relationship between pricing effects and user behavior can help users better understand how to manage their mobile data under different plans. Such an understanding has become increasingly relevant in developing countries like Nigeria, South Africa, and Kenya, which have seen the rise of modern phenomena like mobile leapfrogging [13]. Mobile leapfrogging describes the fairly recent trend of consumers skipping use of fixed-line technologies like PCs and diving straight into the more affordable and convenient mobile option. The resulting rise in popularity of mobile phones in these developing regions brings with it a growing importance to understand and optimize mobile data usage.

Beyond understanding how users manage their data plans, we want to gauge whether and to what degree they may respond to differences in pricing plans. To this end, our study can help network carriers and regulators determine which data plans elicit different types of data usage, and whether varying the prices of prepaid plans affects user behavior. Along with network carriers and regulators, mobile applications and organizations like Facebook's Internet.org can benefit from analyzing possible behavioral effects of practices like zero-rating.

2. Related Work

There has been some related work in the areas of mobile data pricing, usage prediction, and behavioral trend analysis. We will focus on four prior studies: the first proposes a tool for time-dependent smart mobile data pricing, the second attempts to predict user behavior at a micro level, and the last two analyze behavioral trends on a macro scale. Our work is more aligned with the final approach, as our motivation is to obtain general insights into various pricing effects on mobile data usage. However, it is still useful to gain a broad understanding of the work that has been done in the realm of understanding mobile data usage patterns and smart data pricing.

In 2012, Ha et al. [7] proposed the architecture, implementation, and proof-of-concept for a tool called TUBE, which provides time-dependent mobile data pricing by creating a price-based feedback control loop between Internet service providers and their end users. On the ISP side, TUBE computes time-dependent prices to balance the cost of congestion during peak periods with that of offering lower prices during off-peak periods. On the end user side, the tool provides a graphical user interface for users to respond to offered prices either automatically or manually. After conducting a user trial with 50 iPhone or iPad 3G data users charged according to TUBE's algorithms, results showed that the tool helped flatten temporal demand fluctuations to reduce ISP costs (thus benefiting ISPs) while allowing for end users to save money by selecting the time and volume of their usage (thus benefiting customers). Although such an experimental study investigating a functional prototype for smart data pricing is the first of its kind, there is still more work to do in analyzing the empirical effects of different pricing practices already adopted by network carriers on end user behavior. Our study will attempt to discern these relationships.

In 2010, Choujaa et al. [6] presented an information-theoretic approach to predicting human behavior from selected mobile phone data points, using cellular data collected via the Reality Mining project. They had three goals: 1) To use specific time points in a day to predict a user's behavior at another time point, 2) To find the most useful time points in history to predict a user's future behavior, and 3) To determine the difficulty of predicting a user's behavior at a given time

from another user's behavior at another time. They were able to quantify predictability without using specific predictors by *a*) selecting time points to reduce uncertainty of a user's activity at a given time of day and up to three weeks into the future, and *b*) determining a user's activity at a certain time of day given another user's activity at another time. However, the purposes of our study are slightly different – instead of using time to predict user behavior on a select group of mobile users, our analysis attempts to explore the relationships between various pricing-related features and potential behavioral responses for a broad demographic of users.

In 2014, the Sandvine Global Internet Phenomena Report [2] published findings on fixed and mobile access networks in various continental regions around the world. Most relevant to our topic are mobile access trends prevalent in North America and Africa. The 2014 report claimed that peak period mobile traffic in North America is dominated by real-time entertainment and social networking applications. The former accounts for for 36.50% of aggregate traffic on the network, and the latter accounts for 26.36%. The high representation of social network traffic, coupled with the fact that social applications typically generate much less traffic than streaming applications, speaks to the popularity of social networking applications among users. In addition, the report noted that the upward trend in Facebook's traffic share could be attributed primarily to the introduction of a new video autoplay feature, which automatically streams videos on a user's Facebook newsfeed. With regard to traffic trends in Africa, findings report that peak period mobile traffic is dominated by web-browsing (34.85%) and communications (28.92%) applications. In addition, Africa is the only region in which Opera Mini, a web browser focused on data efficiency, is among the top 10 most popular applications, which may speak to a higher dedication to conserving data usage. The 2015 Sandvine Global Internet Phenomena Report [3] noted a slight increase in North American real-time entertainment traffic share and a slight decrease in social networking traffic share. However, these two categories remain by far the dominant traffic contributors in North America. In Africa, web browsing and communications still dominate traffic composition. Notably, WhatsApp network traffic increased by almost 50% to now contribute 10.86% of total network traffic. We will provide a possible explanation for this heightened popularity in Section 6.2.1, in which we discuss Cell C's

zero-rating of WhatsApp in 2014-2015. It is also interesting to note that real-time entertainment has become one of the top 5 contributing traffic categories, holding 6.44% of traffic share. According to the report, this may indicate the beginning of significant growth in this category as both networks and devices improve in Africa. In addition, the contribution of social networking applications increased from 6.06% in the 2014 report to 8.11%, perhaps also hinting at a rise in popularity of this category as well. Our results in Section 5.1.2 largely agree with the Sandvine findings reported in both years, although it is important to note that the Sandvine results are not broken down by country, while our results are specific to users in the United States and South Africa. In our study, four of the top 5 applications in the United States by number of bytes were social networking or real-time entertainment applications. On the other hand, two of the top 5 applications in South Africa were web-browsing and communications applications, while the presence of YouTube and Facebook in the country's top 5 may attest to the rise in popularity of real-time entertainment and social networking applications in the country.

In 2015, Mathur et al. [10] performed a multi-dimensional study on data usage practices in South Africa, a region where data costs are high and usage-based data plans are prevalent. They collected 339 survey responses on mobile data usage and cost management practices from June-July 2014, conducted in-depth interviews with 43 of the survey respondents from June-August 2014, and analyzed MySpeedTest data usage logs (discussed in greater detail in the next section) for 121 unique devices from November 2012-June 2015. The study concluded that mobile users in areas where data is limited and/or expensive are very cost-conscious; these users frequently adopt a variety of non-trivial strategies in an attempt to optimize their mobile data usage. For example, users often switched off cellular data connections or postponed mobile use until connected to Wi-Fi, in addition to avoiding data-intensive applications and changing settings to disable automatic software updates. Our findings in Section 5.1.3 indicate that this may be the case for apps like Google Play Store, which has an option to turn off automatic updates when not connected to Wi-Fi. We discovered that Wi-Fi usage of the app seemed on average much higher than cellular usage in South Africa, indicating that users had likely opted to disable automatic updates until connected to Wi-Fi.

3. Approach

Although the above studies have made significant progress in analyzing mobile data usage patterns and smart data pricing, they do not necessarily provide a holistic understanding of the empirical relationships between user behavior and different types of features, specifically ones involving mobile pricing practices like zero-rating. The key insight to our approach is that it analyzes longitudinal data on a diverse range of global users to explore relationships between such pricing effects and mobile usage. An empirical study comparing different types of mobile usage across different countries and carriers in the context of pricing effects like zero-rating and prepaid data pricing actually hasn't been analyzed at great length before.

Our research makes use of data collected by MySpeedTest, a mobile phone application that collects information on a device's network performance (e.g., throughput and latency), user behavior (e.g., data consumed by and usage frequency of different applications), and metadata (e.g., physical location, data cap plan, battery life). Active measurements such as speed tests initiated by users collect network performance data, while passive measurements collect mobile usage data when the device is on and connected to the Internet. We will be focusing on the relationships between different pieces of metadata associated with each device, and the mobile usage behaviors on that device.

4. Data

For our study, we analyzed usage data measured by MySpeedTest from January 2013 to November 2015. Table 1 shows the relevant tables in the database and the columns we extracted from them for our study, with the columns in italics used to join them together.

The process of collecting, cleaning, and preprocessing the data was an iterative one, and continued to be modified as we ran into new challenges and made new discoveries about the data. We include in Appendix A implementation details for this portion of the project, including the main challenges we faced and solutions we adopted to arrive at a working usage table for our analysis. Table 2 shows this final usage table created from the relevant tables and columns in the original database.

Table 1: Relevant tables and columns extracted from database

Table	Column	Description
application	name	Application name
	package	Package name corresponding to application
application_use	*measurementid*	ID of measurement
	package	Package name corresponding to application used in measurement
	total_sent	Counter for total number of bytes sent by end of measurement
	total_recv	Counter for total number of bytes received by end of measurement
network	*measurementid*	ID of measurement
	connectiontype	Connection type used in measurement (e.g., Wi-Fi, 1G, 2G, etc.)
measurement	*measurementid*	ID of measurement
	deviceid	ID of device reporting usage in measurement
	time	Time of measurement (granularity = 15 minute intervals)
device	*serialnumber*	Serial number of device
	deviceid	ID of device
	networkname	Network carrier to which device belongs
	networkcountry	Country to which device's network carrier belongs
	datacap	Monthly data cap of device (e.g., unlimited, X megabytes cap, prepaid)
sim	*serialnumber*	Serial number of device
	operatorname	Network carrier to which device belongs
	networkcountry	Country to which device's network carrier belongs

For the `connectiontype` column, we categorized all connection types labeled "Mobile" or "Mobile:$\{1,2,3,4\}$G" as "Cellular" connection types. For the `datacaptype` column, we categorized all data plans with monthly limits between 250 MB and 2+ GB as plans with "Limited" data caps.

5. Exploratory Analysis

With the data in this malleable form, we were able to perform some exploratory analysis, all of which was done in various iPython notebooks using the Pandas [11] and MatPlotLib [8] libraries

Table 2: Usage table used for analysis

date	date of measurement, derived from measurement.time (format: YYYY-MM-DD)
month	month and year of measurement, derived from measurement.time (format: YYYYMM)
name	application name associated with measurement, from application.name
deviceid	device associated with measurement, from measurement.deviceid
total_usage	total usage associated with measurement, summed across upstream and downstream bytes, derived from application_use.total_sent and application_use.total_recv
connection	original connection type associated with measurement, from network.connectiontype
connectiontype	connection type categorized into "Wi-Fi", "Cellular", and "Unknown", derived from network.connectiontype
datacap	original data cap associated with device, from device.data cap
datacaptype	data cap type categorized into "Unlimited", "Limited", "Prepaid", and "Unknown", derived from device.datacap
networkcountry	network country associated with device, from device.networkcountry or sim.networkcountry
networkname	network name associated with device, from device.networkname or sim.operatorname

for analysis and visualization, respectively. This step consisted of comparing usage across different countries for different connection types and data cap types during a three-year period, from January 2013 through November 2015. For the purposes of our study, we focused on data collected from devices in the United States and in South Africa, the former due to the context we have regarding usage patterns in our own country, and the latter as a precursor to our pricing analysis.

The main goal in this portion of our analysis is to compare usage behaviors between the United States and South Africa for different connection types and data cap types across a longitudinal period. The research question we attempt to address is twofold: 1) whether cellular vs. Wi-Fi connections affect data usage in each country, and 2) whether unlimited vs. limited vs. prepaid data plans affect data usage in each country.

5.1. Results

We report a total of 12,277 unique users collectively using 67,821 applications throughout the entire period of study from January 2013 through November 2015. Across all users, the median number of mobile applications per device is 39. However, there appears to be a right skew in this distribution, as the average is much higher at 57 applications per device. This is verified by the fact that three quarters of the users have 75 or less applications installed on their devices, while there are a few devices with much higher numbers of installed applications (in the hundreds). Across all users, the median of the average number of applications used per day is 24.21, indicating that users generally use over half of their installed applications every day, on average. However, many of these applications are likely background processes.

To calculate an appropriate metric for usage of a given application by a particular category of users over a certain period of time, we first found the average monthly usage of the application across all users in that category. We then took the median across all months in the period of interest. This made our metric more robust to potential outliers in the dataset while reflecting general usage patterns across our user base.

5.1.1. Data Composition

First, we want to explore the overall composition of our data, which can be gleaned by looking at a breakdown of users and measurements for the period of study.

Table 3 shows a quarterly breakdown of users by data cap type for the duration of our study. As mentioned, we have grouped together all data plans with monthly limits between 250 MB and 2+ GB into the "Limited" category. Note that Q4 2015 excludes December, as the last measurement taken prior to writing up these results was on November 30, 2015.

From these results, it appears that 2013 showed a much higher count of active users, about an order of magnitude more than in 2015. 2014 had around 1,000 active users per quarter, with a blip in Q2 due to some problems with the MySpeedTest application. 2015 had much lower numbers overall, with a couple hundred active users per quarter.

In 2013, we see a comparable amount of users on unlimited plans as users on plans with a monthly

Table 3: Quarterly user breakdown by data cap type

Quarter	# Users	Unlimited	Limited	Prepaid	Unknown
2013 Q1	4,797	2,150	2,068	0	579
2013 Q2	3,784	1,694	1,664	0	426
2013 Q3	2,610	1,244	1,074	0	291
2013 Q4	2,408	1,217	941	0	250
2014 Q1	1,566	812	592	1	161
2014 Q2	595	362	187	0	46
2014 Q3	1,192	756	355	17	64
2014 Q4	919	593	252	24	50
2015 Q1	522	341	150	6	25
2015 Q2	438	224	147	11	56
2015 Q3	338	175	114	9	40
2015 Q4	262	132	92	7	31

limit, although the former category has slightly more users than the latter in each quarter. This difference in users between the two categories is magnified as the number of overall users drops through the next two years. By 2015, the majority of users are on unlimited plans, although users on limited plans are still well-represented in the data.

There are at most a handful of users on prepaid plans in each quarter (throughout 2013 there were no users at all who reported being on prepaid plans). This could pose some trouble for our data pricing analysis, for which only devices on prepaid plans are relevant. We will discuss these problems in Section 6.3. An important issue of note here is that the data cap associated with each device is reported by the respective user. It was brought to our attention that this self-reporting may contain factual errors, as users may be unsure of their data plans and can easily misreport this information. Thus, we must keep in mind that our breakdown by data cap type may not be 100% trustworthy due to possible misreporting. In fact, South African data plans are predominantly of the prepaid variety, but the MySpeedTest dataset contains users on unlimited plans (which do not exist in South Africa). This discrepancy is certainly concerning, and leads us to believe that our

breakdown by data cap type may not be entirely trustworthy, especially in South Africa, based on possible misreporting of the original data.

For the purposes of our comparison analyses between connection types and data cap types, we decided to exclude those devices whose data cap types were unknown.

Table 4 shows a quarterly breakdown of measurements by connection type for the duration of our study. We aggregated the measurements to a daily granularity, and as mentioned, we have grouped together all types of cellular connections (e.g., 1G, 2G, 3G, 4G) into the "Cellular" category.

Table 4: Quarterly measurement breakdown by connection type

Quarter	# Measurements	Wi-Fi	Cellular	Unknown
2013 Q1	3,546,252	1,348,135	1,944,605	253,512
2013 Q2	4,647,397	1,735,783	2,448,425	463,189
2013 Q3	3,873,565	1,556,440	1,960,871	356,254
2013 Q4	3,046,596	1,290,031	1,492,056	264,509
2014 Q1	1,358,882	598,714	641,662	118,506
2014 Q2	358,670	161,089	159,139	38,442
2014 Q3	1,533,182	649,263	728,297	155,622
2014 Q4	1,320,134	538,473	656,357	125,304
2015 Q1	520,320	215,099	261,076	44,145
2015 Q2	544,442	210,364	284,078	50,000
2015 Q3	557,535	206,781	303,257	47,497
2015 Q4	392,635	154,419	204,050	34,166

These results reflect the decline in active users throughout the period of study. The average quarterly number of measurements decreased from 2013 to 2014 by about 60%, and from 2014 to 2015 by well over 50%.

As we can see from the number of quarterly Wi-Fi vs. cellular measurements in Table 4, users seem to use cellular connections slightly more than Wi-Fi connections. A possible explanation for this is that generally when people are on their mobile devices, they may not be in an area with Wifi (hence the term "mobile" device). Thus, they are more likely to use up cellular data on various

apps they may need or want to access at any given moment. However, the prevalence of Wifi usage indicates that users may still opt to conserve cellular data usage when they are in Wifi areas, connecting to the Internet on Wifi when possible.

Finally, for the purposes of our comparison between connection types and data cap types, we decided to exclude those measurements whose connection types were unknown.

5.1.2. Top 5 Used Apps in US vs. ZA

Throughout the period of study, we report a total of 1,034 US users, with 403 on unlimited plans, 627 on limited plans, and 4 on prepaid plans. We report a total of 249 South African users, with 63 on unlimited plans, 176 on limited plans, and 10 on prepaid plans.

For our comparisons, we will consider the top 5 mobile applications in each country, ranked by total data usage in bytes throughout the period of study from January 2013 through November 2015. We excluded the MySpeedTest application from this ranking. The top 5 used applications from January 2013 through November 2015 are shown in Table 5.

Table 5: Top 5 apps by total data usage for US vs. ZA

US	ZA
1. Netflix	1. YouTube
2. YouTube	2. Facebook
3. Facebook	3. Chrome
4. Google+	4. Correo
5. Browser	5. Google Play Store

These results agree with the Sandvine hypothesis [2, 3] that North American data usage is dominated by social networking and real-time entertainment applications, as evidenced by the top 4 applications in the United States: Netflix, YouTube, Facebook, and Google+. It makes sense that Netflix and YouTube would be the top contributors to mobile usage in America, because streaming applications typically generate a lot of traffic. Facebook and Google+ take the next two spots, attesting to the popularity of social media applications in users' mobile experiences. The Sandvine reports also declare web-browsing and communications applications to be the top contributors of

African mobile traffic, which is reflected in the presence of Chrome and Correo (a mail app) in our list of South Africa's top 5 applications. Sandvine's 2015 report shows a significant rise in popularity of real-time entertainment, as well as a small increase in traffic contribution of social networking applications, which can explain the presence of YouTube and Facebook in the list.

5.1.3. Effects of Connection Type on Mobile Data Usage in US vs. ZA

Figures 1 and 2 show the usage breakdown by connection type for the top 5 apps in each country throughout the entire period of study, from January 1, 2013 through November 31, 2015.

Figure 1: Usage breakdown by connection type for top 5 apps in the US

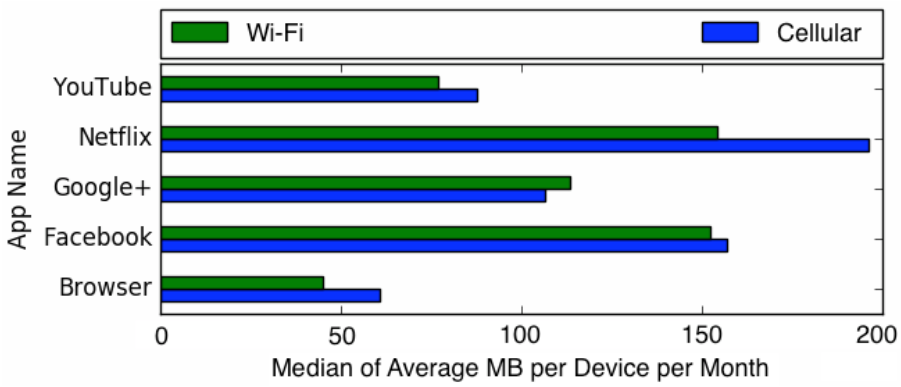
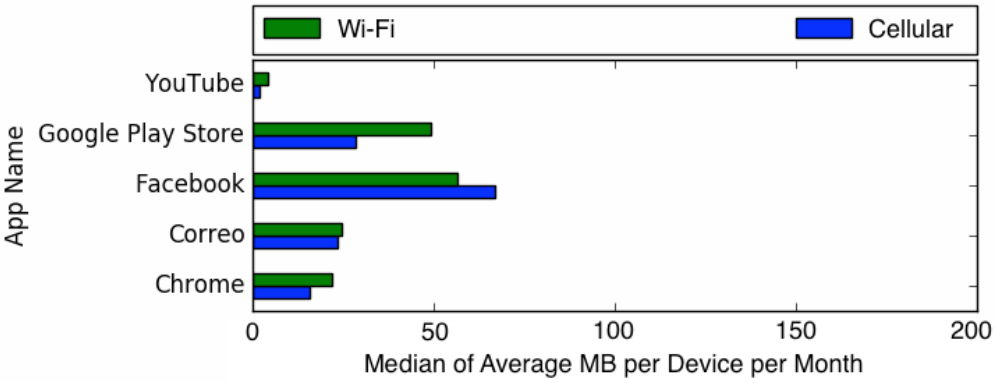


Figure 2: Usage breakdown by connection type for top 5 apps in South Africa



We can clearly see that US mobile usage dwarfs South African mobile usage for the top 5 apps in each country. In addition, US users generally access the relevant applications on cellular connections more frequently than on Wi-Fi connections, even for data-intensive streaming applications like Netflix and YouTube. In South Africa, on the other hand, users seem more wary of cellular data

usage, preferring Wi-Fi connections for almost all of the listed applications. This difference may imply that US users are not as cognizant of data usage as South African users. It also points to the possibility that mobile speeds in the US may be much better than those in South Africa, or that perhaps US mobile carriers offer better data plan options that allow for more cellular data usage. It is interesting to note that Facebook has the second highest median of average monthly Wi-Fi and cellular usage in the United States, behind only Netflix. In fact, Facebook and Google+ report higher usage than YouTube, even though as a streaming application YouTube generally uses much more data. This points to the huge popularity of social networking applications among users in the United States.

In South Africa, Facebook has the highest median of average monthly Wi-Fi and cellular usage, and is the only application showing higher cellular usage than Wi-Fi usage. The popularity of the application indicates the growing prevalence of social media in developing countries like South Africa. Perhaps the higher cellular usage on the app can be attributed to the introduction of Facebook's new video autoplay feature automatically streaming videos on users' newsfeeds, as mentioned in the Sandvine reports [2]. Users may not be aware of the amount of cellular data used simply by scrolling through their newsfeeds. YouTube usage is relatively low compared to the communications and social networking applications, perhaps implying that South African users may still be more wary of using data-intensive real-time streaming applications than communications and social networking applications on their mobile devices. However, this is still surprising given that YouTube was in fact the most used application by number of bytes in South Africa. The discrepancy may be due to extremely high usage by certain outliers that ranked the application so high in terms of total data usage, whereas our metric comparing median of average monthly usage per device mitigates the influence of these outliers and in fact reflects much lower YouTube usage on a typical device. However, it would be interesting to look into the distribution of YouTube usage in South Africa to discern possible reasons for this disparity. Finally, the prevalence of Google Play Store could perhaps be attributed to automatic updates that the application often performs in the background, which users may or may not be aware of. There is an option on the app to

disable automatic updates when not connected to Wi-Fi. The fact that Wi-Fi usage of the app nearly doubles cellular usage indicates that users may have taken advantage of this option. This agrees with the findings made by Mathur et al., which stipulate that users adopt various strategies to optimize mobile data usage, including changing settings to disable automatic software updates and postponing use until connected to Wi-Fi. Our observations here all indicate a higher dedication among South African users to conserving data usage when on a cellular connection.

5.1.4. Effects of Data Cap Type on Mobile Data Usage in US vs. ZA

Figures 3 and 4 show the usage breakdown by data cap type for the top 5 apps in each of the two countries throughout the entire period of study. Again, we categorized all data plans with monthly limits between 250 MB and 2+ GB as plans with "Limited" data caps.

Figure 3: Usage breakdown by data cap type for top 5 apps in the United States

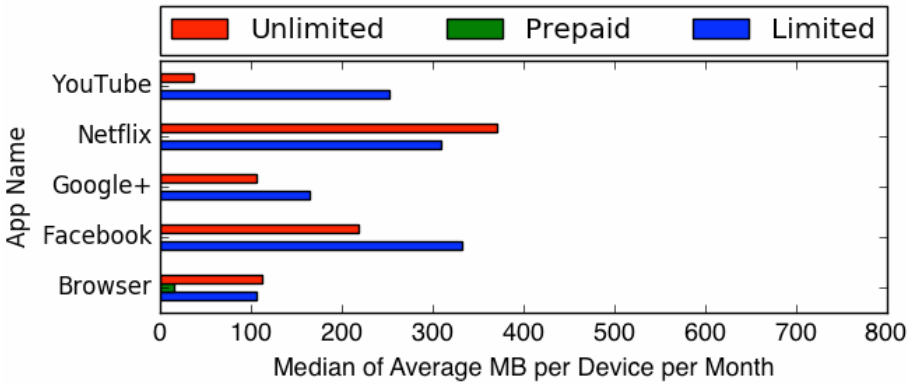
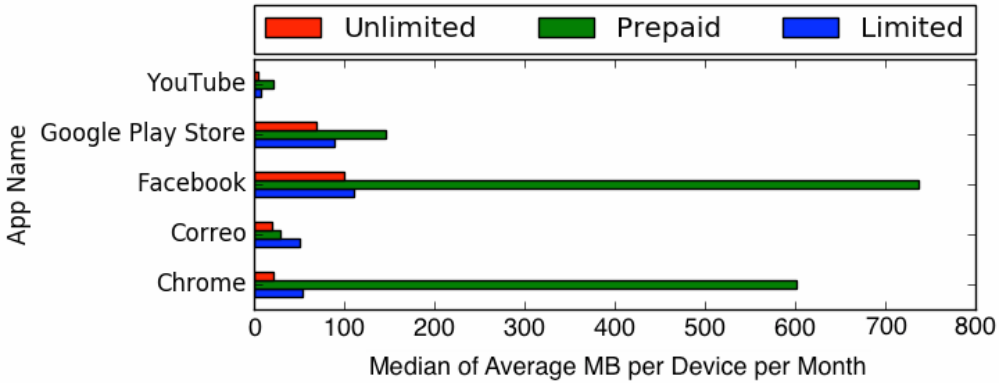


Figure 4: Usage breakdown by data cap type for top 5 apps in South Africa



Both unlimited and limited data plans are widely represented in US usage, whereas we see virtually no appreciable usage by the few users on prepaid plans. For three of the five applications,

usage by devices with a limited data cap is significantly higher than usage by devices on unlimited plans. However, unlimited users in the US do contribute more usage than limited users for extremely data-intensive, long-term streaming applications like Netflix. It is interesting that, unlike Netflix, YouTube usage is heavily dominated by users on limited plans. Overall, the prevalence of limited data plans in US usage again points to the possibility that carriers in the United States may offer better mobile data plan options, perhaps with higher monthly data caps. It would be interesting to obtain context into the variety of data plan options offered by the major network carriers in the United States, perhaps by studying how often US users on limited plans actually hit their monthly data caps, and how flexible carriers are in allowing users to pay for increased caps should they be close to their limits in any given month.

In South Africa, we see a much different breakdown in which users on prepaid plans generally show much higher usage than those on other plans for nearly all of the applications. Most notably, these users dwarf users on unlimited and limited plans with regard to Facebook and Chrome usage. Perhaps this implies that South African users on prepaid plans reserve use of their mobile devices for social networking and web browsing activities. These results are generated by a total of only 10 users on prepaid plans, so our findings may exhibit some skew due to the low sample size. Since South African data plans are predominantly of the prepaid variety, it would be interesting to reevaluate these results once the dataset contains more prepaid users. Recall that our data cap type breakdown may not be completely accurate due to possible errors in self-reporting, so our findings in this section remain largely inconclusive.

5.1.5. Differentiated Effects of Data Cap & Connection Type on Data Usage in US vs. ZA

In Section 5.1.3, we compared mobile data usage on Wi-Fi vs. cellular connections in both countries (regardless of data cap type), while in Section 5.1.4, we compared data usage by users on unlimited vs. limited vs. prepaid plans in both countries (regardless of connection type). We now dig deeper into the individual effects of each feature on usage behavior, holding the other feature constant. We are interested in breaking down mobile data usage by both features simultaneously to avoid conflating their respective effects on usage. For this portion of the analysis, we studied the

two applications that were in the top 5 for both the United States and South Africa: YouTube and Facebook. Table 6 shows the results of the usage breakdown by both features for each application in the two countries, again using median of average monthly usage as our metric of comparison.

Table 6: US vs. ZA usage breakdowns by data cap and connection type for YouTube and Facebook

App Name	Data Cap	US		ZA	
		Wi-Fi(MB)	Cellular(MB)	Wi-Fi(MB)	Cellular(MB)
YouTube	Unlimited	24.587	10.893	2.444	1.904
	Limited	88.588	133.549	4.331	1.508
	Prepaid	0.049	0.021	8.721	0.034
Facebook	Unlimited	94.597	116.971	49.427	25.399
	Limited	144.927	290.357	42.036	68.579
	Prepaid	0.000	0.000	139.814	443.703

First we'll look at usage behaviors in the US, recalling that breakdown by only connection type revealed higher cellular usage than Wi-Fi usage. By holding data cap type constant and comparing usage on different connection types, we find that users on limited plans contributed most to this difference for these two applications, using both significantly more on cellular than on Wi-Fi. This again points to the possibility that US limited data plans have higher monthly data caps that may in fact place little constraint over cellular usage. Interestingly enough, users on unlimited plans reported much lower Wi-Fi and cellular usage than those on limited plans. Users on prepaid plans contributed almost no appreciable usage, but the small amount of usage on YouTube was expectedly higher on Wi-Fi connections than on cellular connections, as users would likely not want to use up their prepaid cellular data on a data-intensive streaming application.

With regard to South African usage behaviors, we recall that breakdown by only connection type revealed higher Wi-Fi than cellular usage on YouTube but higher cellular than Wi-Fi usage on Facebook. Breakdown by only data cap type revealed users on prepaid plans dominating average usage, especially for Facebook. Again holding data cap type constant and comparing usage on different connection types, we see that users on prepaid plans indeed contribute the most overall

usage on average. Oddly enough, cellular usage of the app triples Wi-Fi usage among this category of users, implying that prepaid users make up the category that contributes most to the overall higher cellular than Wi-Fi usage on Facebook. Examining YouTube usage for each data cap type, we observe that the preference for Wi-Fi over cellular usage for the app is clearly reflected in the table for users on all data cap types. The most pronounced difference can be seen among prepaid users, followed by users on limited plans and finally by users on unlimited plans.

6. Pricing Analysis

In addition to the above exploratory comparisons between usage behaviors in the United States and South Africa, we also want to study the effects of zero-rating and prepaid data pricing on mobile data usage, specifically in South Africa. We focus on South Africa for our pricing analysis because prepaid plans are most prevalent there, and mobile data in the region is relatively limited and expensive. Thus, we believe the effects of zero-rating and data pricing would be most relevant and pronounced in an area where users are perhaps more sensitive to changes in mobile pricing.

6.1. Pricing Behavior Context in South Africa

In the context of zero-rating, our study specifically compared data usage on WhatsApp between the 4 predominant carriers in South Africa: Vodacom, Cell C, MTN, and Telkom. This is because Cell C zero-rated WhatsApp for mobile users on its network from November 19, 2014 to August 31, 2015, before switching to a bundle offer under which users could use the app for 30 days up to a fair usage cap of 1 GB, excluding voice calls, for ZAR 5. With this context, it would be useful to analyze the Wi-Fi vs. cellular usage behaviors of Cell C users compared to users on other networks on WhatsApp, during the zero-rating period as well as the bundle offer period. In addition, we know that Cell C, in collaboration with Facebook, launched Free Basics (a.k.a. Internet.org) on its network on July 1, 2015. This zero-rated the app on Cell C, so we wanted to do a similar comparative study on Internet.org data usage for Cell C vs. the other network carriers in South Africa during this period. Finally, we looked into Twitter usage trends, knowing that MTN had zero-rated the application from May 1 to July 31, 2014, and again during the ICC Cricket World Cup from February 14 to

March 29, 2015. Although the Cricket World Cup ended in March, the offer was maintained and is still active today. With this knowledge, we thought it would be interesting to see whether temporal usage patterns reflected this time-dependent zero-rating.

With regard to data pricing, we wanted to make use of Research ICT Africa's information containing quarterly costs of 1 GB data bundles for prepaid plans on all 4 major South African carriers, from Q1 2014 through Q4 2015 [1]. We noted that for some carriers, these prices remained relatively consistent, while others increased or decreased their prices throughout the time period. With this context, we wanted to analyze potential behavioral differences among users on prepaid plans that might reflect data pricing changes made by different carriers.

6.2. Effects of Zero-Rating on Mobile Data Usage in ZA

South Africa's mobile market is dominated by four main carriers. Table 7 shows the number of subscribers and percent of market share held by each of these four carriers, as reported in 2014 [5]. The largest mobile operator by number of subscribers is Vodacom, although its market share decreased by over 10% from 2012 to 2014. Vodacom's closest competitor is MTN, whose market share has remained relatively stable. Historically, Cell C and Telkom have had a much smaller subscriber base, but Cell C has seen massive amounts of growth in subscriber numbers in recent years [5].

Table 7: Subscriber and market share distributions in 2014 for top 4 ZA carriers [5]

Carrier	# Subscribers	% Market Share
Vodacom	31.4 million	40.0%
MTN	28.0 million	35.0%
Cell C	18.1 million	25.0%
Telkom	1.8 million	2.3%

With this context, we begin the analysis of our results regarding zero-rating among different carriers in South Africa. For reference throughout the discussion of our findings, table 8 shows each zero-rating offer that our study analyzes.

Table 8: Zero-rating practices among South African carriers

Carrier	Application	Offer	Duration
Cell C	WhatsApp	No offer	Before 11/19/14
		Zero-rated	11/19/14–08/31/15
		Bundle offer	09/01/15–present
Cell C	Free Basics	No offer	Before 07/01/15
		Zero-rated	07/01/15–present
MTN	Twitter	No offer	Before 05/01/14
		Zero-rated	05/01/14–07/31/14
		No offer	08/01/14–02/13/15
		Zero-rated	02/14/15–present

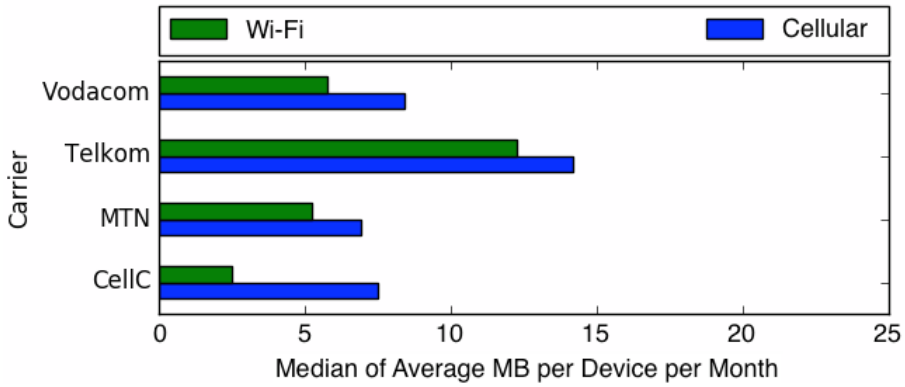
6.2.1. Effects of Zero-Rating WhatsApp on Cell C's Network

As seen in Table 8, WhatsApp's zero-rating offer on Cell C's network began on November 19, 2014, before which Cell C had no promotional offer for WhatsApp users on its network. After the offer expired on August 31, 2015, Cell C adopted a bundle offer in which, for a fee of ZAR 5 (approx USD 0.327), users could use up to 1 GB on WhatsApp for 30 days, excluding voice calls.

Our goal was to analyze whether mobile usage behavior reacted to these different types of promotions, transitioning from no offer to completely zero-rating the app to offering a generous amount of monthly usage for a small fee. Following are the results of our study comparing WhatsApp mobile usage on Wi-Fi vs. cellular connections during these three periods for Cell C vs. the other 3 main carriers in South Africa, which to our knowledge offered no WhatsApp promotions throughout the last three years.

Figure 5 shows WhatsApp's usage breakdown by connection type for the 4 main South African carriers during the period with no zero-rating offer. It appears that cellular usage is consistently higher than Wi-Fi usage for users on all 4 carriers. Telkom users generally seem to use WhatsApp the most, while MTN and Cell C users appear to use the app the least. Cell C has the lowest Wi-Fi usage on WhatsApp, chalking in at a median of around 2.5 MB of average monthly usage per device. Cellular usage on the app is about three times as high, at around 7.5 MB. This difference between

Figure 5: WhatsApp usage breakdown for 4 main ZA carriers: no offer



Wi-Fi and cellular usage among Cell C users is interesting, since mobile use of the app had not yet been zero-rated during this period. There may be some underlying factors that could contribute to the lower Wi-Fi usage on Cell C, such as lack of connectivity on home Wi-Fi connections, but this is difficult to determine definitively. Overall, though, WhatsApp usage on Cell C during this period is roughly similar to usage on the other carriers in terms of having comparable overall median of average monthly usage as well as displaying higher cellular than Wi-Fi usage.

Figure 6: WhatsApp usage breakdown for 4 main ZA carriers: Cell C zero-rating offer

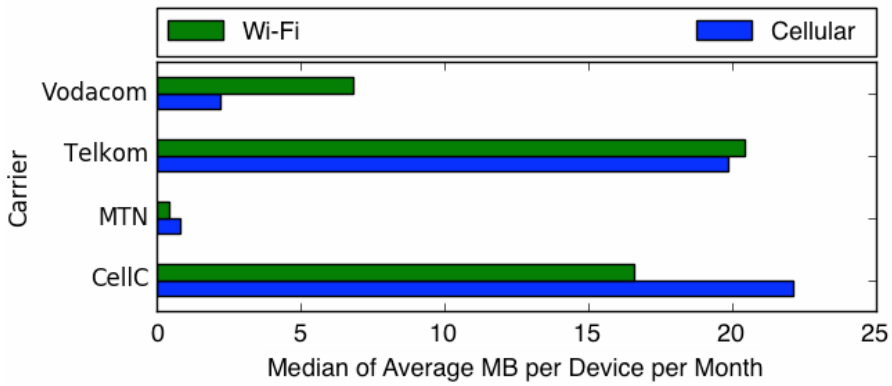


Figure 6 shows WhatsApp's usage breakdown by connection type for the 4 main South African carriers during Cell C's zero-rating of the application. Cell C usage has indeed shot up, with Wi-Fi usage increasing almost sevenfold to a median of around 17 MB of average monthly usage per device, and cellular usage increasing almost threefold to around 22 MB. This indicates that zero-rating WhatsApp may have had remarkable effects on its usage, greatly increasing the application's use not only on cellular connections, but actually even more so on Wi-Fi connections. The fact that

overall usage on Cell C seemed to increase so dramatically during this zero-rating period implies that perhaps WhatsApp became much more popular overall, regardless of connection type, because users felt they could use the app freely on any connection at no cost.

Oddly enough, we also see behavioral changes in WhatsApp usage for the other three carriers during this period. Cellular usage on Vodacom decreased, as did overall usage on MTN, while overall usage on Telkom increased. Wi-Fi usage on both Vodacom and Telkom became higher than cellular usage during this period as well. However, we found that the number of unique users on each carrier during this period may not be high enough to draw conclusive results. We report only 10 WhatsApp users on Vodacom, 1 WhatsApp user on Telkom, 4 on MTN, and 4 on Cell C. See Table 9 for a breakdown of the number of users in each carrier and time interval.

Figure 7: WhatsApp usage breakdown for 4 main ZA carriers: Cell C bundle offer

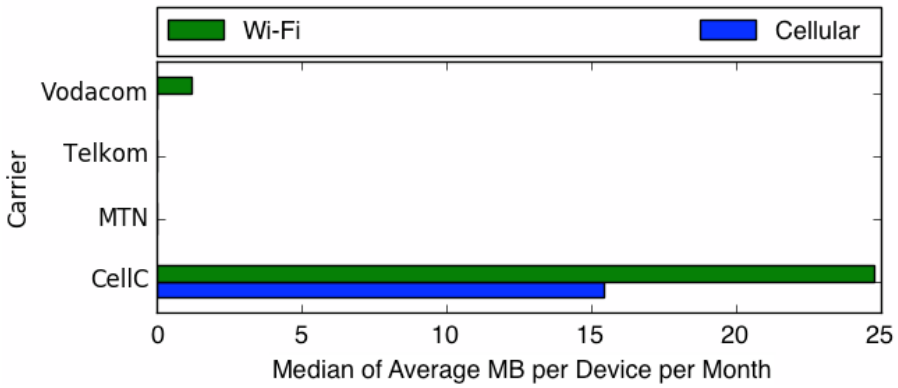


Figure 7 shows WhatsApp's usage breakdown by connection type for the 4 main South African carriers during its bundle offer period. We discovered in our data that in this period there was only 1 Vodacom user and 1 Cell C user using WhatsApp, with no Telkom or MTN users on WhatsApp at all (See Table 9). Due to this lack of appreciable data, we are unable to perform a thorough comparison between WhatsApp usage on different carriers during the bundle offer period. However, we note that the single user on Cell C does display a high monthly median usage on WhatsApp, comparable to usage seen during the zero-rating period, and certainly much higher than usage before any of Cell C's promotional offers. To reiterate, though, our sample size is too small for these results to be seen as entirely conclusive.

Table 9: Number of WhatsApp users for each carrier and time interval

Carrier	Before 11/19/14	11/19/14–08/31/15	08/31/15–present
Vodacom	89	10	1
Cell C	46	4	1
MTN	60	4	0
Telkom	14	1	0

After performing this analysis on Cell C's zero-rating effects on WhatsApp, we can now understand some possible context for why WhatsApp's network traffic increased by almost 50% in 2015 to contribute 10.86% of total network traffic in Africa, as discussed in Sandvine's 2015 report [3]. Perhaps zero-rating WhatsApp on one of the largest mobile network carriers in South Africa contributed to this significant increase in popularity of the application this past year. In addition, as mentioned in the beginning of Section 6.2, Cell C has enjoyed huge amounts of growth in its subscriber base in recent years [5]. Perhaps promotional offers such as this will serve well to increase the popularity of not only the zero-rated application, but also the zero-rating network. Unfortunately, we are unable to draw strong conclusions from our own results due to the limited number of users in our dataset.

6.2.2. Effects of Zero-Rating Free Basics on Cell C's Network

As mentioned in Section 6.1, Free Basics (a.k.a. Internet.org) was launched on Cell C's network in collaboration with Facebook on July 1, 2015. Thus, similar to our WhatsApp analysis, we wanted to compare usage breakdowns by connection type for the app before and after the launch date of this promotion. However, upon searching for the app among our entire user base, we found that only a handful of users had the app installed, and they were all on an India-based network called Airtel. Thus, due to the current lack of available data, we are at the moment unable to properly study the effects of launching the zero-rated Free Basics application in South Africa. However, we would like to highlight the extensibility of the code we have written to perform such an analysis. Once sufficient data is available, we can simply run our code on the relevant data and obtain the necessary results for analysis.

After doing some research into the connection between Airtel and Internet.org, we found that Airtel had launched Internet.org in Ghana in January 2015 [12] and in Malawi in May 2015 [4]. So, we believe that there may simply be a slight lag time for the application's launch in a country to spread across the launching network's user base. However, we also stipulate that perhaps zero-rating via Free Basics may not have as pronounced an effect as zero-rating WhatsApp, because Free Basics allows access to apps like Facebook, which is already so heavily ingrained in most users' virtual experiences and social media lives. Since Facebook is already one of our top 5 most used apps in South Africa, users are already committed to using the application as is, regardless of costs. This might contribute to the slow reaction of Cell C users downloading the Free Basics app and switching over to the free version of Facebook that it provides.

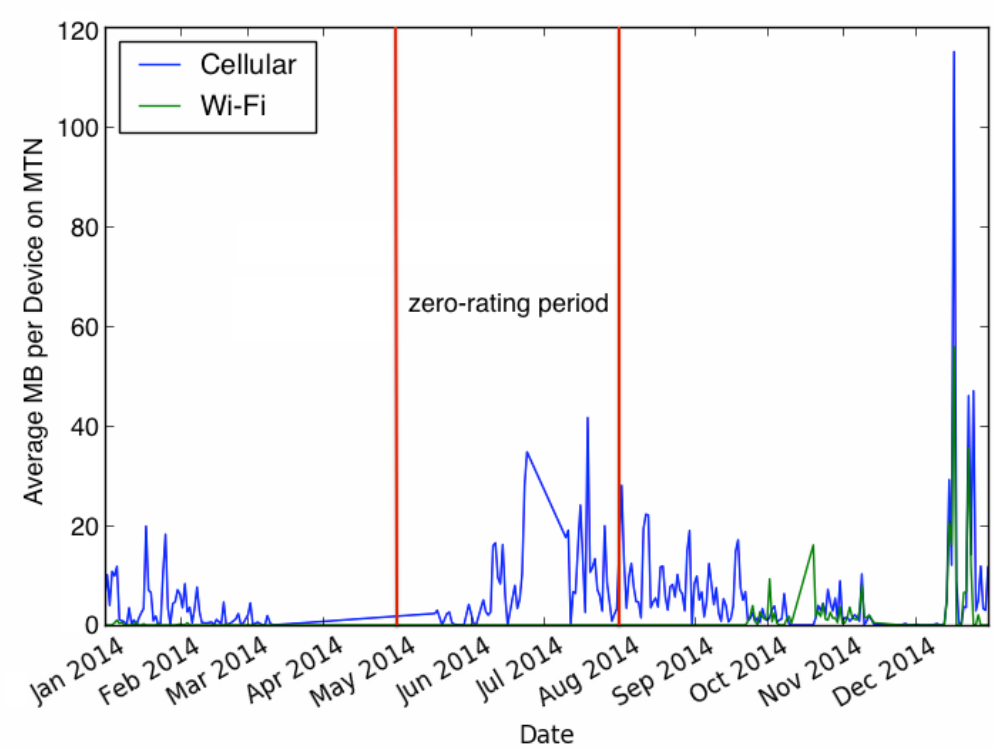
6.2.3. Effects of Zero-Rating Twitter on MTN's Network

We discussed in Section 6.1 MTN's promotional events zero-rating Twitter from May 1 to July 31, 2014, and from the beginning of the ICC Cricket World Cup in February 2015 onward. Thus, we wanted to study whether we could glean any temporal usage patterns that reflected MTN's time-dependent zero-rating of Twitter for certain events like the Cricket World Cup. However, we found that the dataset contained 0 MTN Twitter users after January 2015, so we were unable to analyze the zero-rating effects specifically during the Cricket World Cup. However, we did have data throughout 2014 – Figure 8 shows the corresponding time-series plot of average daily Twitter usage per device on Wi-Fi vs. cellular connections.

Our results indicate much higher cellular usage in general than Wi-Fi usage throughout the year, aside from a small bump in Wi-Fi usage over cellular usage in October. The period from the end of March through the beginning of May reflects the lack of data in the beginning of the second quarter due to a drop in usage resulting from problems with the MySpeedTest application.

The largest spike in both cellular and Wi-Fi data usage seems to correspond to the holiday season in December. This makes sense because social media platforms like Twitter are often used during holiday seasons for advertising and marketing, searching for and sharing gift ideas, posting wish lists, and spreading general holiday cheer. A 2013 Twitter blog post mentioned that holiday

Figure 8: Time-series of average twitter usage per device on MTN



shopping conversations in 2012 increased by 30% over the previous year, and Twitter generally sees significant peaks around key shopping days during the holiday season, such as Christmas Eve [14].

Interestingly enough, the period from June to July 2014 shows the second most significant peak in cellular data usage. This perhaps indicates that MTN's promotional event zero-rating Twitter during this time period may have had some effect on user behavior. Cellular usage seemed to begin picking up around early June, reflecting a slight delay in user response to the announcement of the promotional offer. Usage peaked at a little over 40 MB a day in mid-July, near the end of the event, before taping back down throughout August and September. The fact that usage remained relatively high throughout the months immediately following the event implies that the promotion did seem to maintain some heightened level of interest and use in the application for a while, but by the end of the year cellular usage of the app had gone back down to pre-promotion levels seen at the beginning of the year. Thus, it seems that promotional events like zero-rating Twitter for a certain period of time may not have lasting effects on the app's popularity, though it may increase short-term usage during and immediately following the promotion.

6.3. Effects of Varying Prepaid Data Prices on Mobile Data Usage in ZA

The metric we chose to examine the price of a 1 GB basket is the USD Purchasing Power Parity (PPP) price. The concept of Purchasing Power Parity takes into account an adjustment to the exchange rate between two currencies such that the exchange is at par with each currency's purchasing power in its respective country. Thus, using this metric better reflects differences in purchasing power between South Africa and the United States, instead of simply using the unadjusted exchange rate between ZAR (South Africa Rand) and USD (United States Dollar).

6.3.1. Quarterly Data Prices

Table 10 shows the USD PPP prices per 1 GB basket set by the 4 main carriers in South Africa. As we can see, Cell C, Vodacom, and MTN held the same price of $24.771 throughout 2014, while Telkom transitioned from a price ($29.925) initially much higher than the standard to a price ($19.784) much lower, between Q3 and Q4 2014. All 4 carriers collectively reduced prices between year end 2014 and the new year of 2015, settling at a new standard of $24.349, with Telkom at a comparably lower price of $19.359. Interestingly enough, between Q1 and Q2 2015, MTN deviated upward from the standard by raising its price from $24.349 to $26.029, while Telkom deviated even further downward by lowering its price from $19.359 to $16.105. Cell C and Vodacom maintained consistent prices throughout 2015.

Table 10: Quarterly USD PPP prices per 1 GB basket

Carrier	2014 Q1	2014 Q2	2014 Q3	2014 Q4	2015 Q1	2015 Q2	2015 Q3
Cell C	24.771	24.771	24.771	24.771	24.349	24.349	24.349
Vodacom	24.771	24.771	24.771	24.771	24.349	24.349	24.349
MTN	24.771	24.771	24.771	24.771	24.349	26.029	26.029
Telkom	29.925	29.925	29.925	19.784	19.359	16.105	16.105

We can understand these pricing choices in the context of the economic theory of perfect competition. Since none of the carriers holds a monopoly over South Africa's mobile network, we generally encounter a standard price shared by most carriers. At the outset, Telkom was the only

deviating carrier, with a much higher price than the rest. Perhaps because it realized it was losing market share with its comparably high price, it drastically dropped its price by over $10.00 in an attempt to compensate for this loss and draw customers back in order to increase its market share. In addition, the collective price drop between 2014 and 2015 shows that any change in price to the standard is quickly adopted by most of the carriers, because each carrier wants to maintain a competitive price in the market so as not to lose market share.

With this context behind the changing prices offered by different carriers, it would be interesting to see whether Telkom prepaid users displayed any behavioral changes between Q3 and Q4 2014, when prices dropped by more than 30%. Similarly, it would be useful to look into changes in behavior of Telkom vs. MTN prepaid users between Q1 and Q2 2015, when MTN increased prices and Telkom further decreased them. For these comparisons, we could use Cell C and Vodacom users as a control, since these two carriers maintained their prices at the competitive standard.

6.3.2. Insufficiencies in Relevant Data

We attempted to study the behavioral changes between the above quarters of interest, but we found instead that the MySpeedTest dataset contained no appreciable data on which to perform any useful analysis. Table 11 shows the number of unique users on each of the 4 carriers' prepaid plans for the relevant quarters. Throughout 2014 and 2015, we had no Telkom users on prepaid plans at all, which precluded us from looking into the behavioral effects of Telkom's price drops. We had no MTN users on prepaid plans after Q1 2015, which prohibited us from observing any changes in behavior resulting from MTN's price increase in 2015. Without relevant data, we could not properly evaluate any possible effects of these two carriers changing their data prices for prepaid plans.

Table 11: Number of unique users on prepaid plans per quarter

Carrier	2014 Q1	2014 Q2	2014 Q3	2014 Q4	2015 Q1	2015 Q2	2015 Q3
Cell C	0	0	1	0	0	0	0
Vodacom	0	0	4	7	2	1	0
MTN	0	0	0	2	1	0	0
Telkom	0	0	0	0	0	0	0

7. Future Work

7.1. Recruitment

Recruitment is the most important aspect of future work in the enhancement of this study. As of our last measurement date on November 30, 2015, there were only 66 unique devices reporting active data usage. The lack of sufficient data in various categories, such as for certain carriers or certain data plans, reduces the robustness of our results and renders many of our findings inconclusive. In addition, although we attempted to explore yearly trends in mobile usage behaviors, we were unable to draw intuitive meaning from these analyses due to the lack of data in each individual year of study. This limited us to examining overall usage throughout all three years without differentiating between possible behavioral nuances in each year.

The area where we believe recruitment is highly crucial is South Africa, which is where our pricing analysis is most relevant. More specifically, we urge the recruitment of Cell C users for the purposes of enhancing our WhatsApp analysis. We also believe it is especially important to encourage users on the network to download Free Basics, since we currently have no data on the effects of Cell C zero-rating the app. In addition, recruiting MTN users will be necessary to properly analyze the network's current zero-rating policy with Twitter. Further, in order to reach any conclusive results on the effects of varying data prices on mobile usage, it is vital to recruit users on prepaid plans for all 4 major South Africa carriers. Perhaps most relevant for this purpose right now is recruitment centered around Telkom's and MTN's networks, as the former has displayed a trend in decreasing data prices this past year while the latter has been shown to increase them. The current lack of appreciable data for these two carriers has precluded a proper analysis on the behavioral effects of these changes in data pricing for prepaid plans. An equally important area of future work in this vein pertains to the improvement of the MySpeedTest application to prevent data cap type misreporting, perhaps by making data cap categories more explicit to users.

We have written up the code needed to perform the above pricing analysis studies, should significant data in these categories be collected and made available to us.

7.2. Prediction Using ML Techniques

It may also be interesting to narrow the scope of this research from a broad exploratory analysis to a study on the predictive powers of features (pricing and non-pricing related) on mobile data usage. In order to perform such an analysis using machine learning techniques, we would first need to convert the categorical variables that we are currently dealing with into numerical values that can be interpreted by ML estimators. To this end, we have written the code to perform one-hot encoding of these categorical features into category indices.

The next step could be to first run a basic linear regression on the one-hot encoded features, which would give a naive, rudimentary estimate of the predictive power of each feature. It would also be useful to look into running a Support Vector Machine on the data. Since the one-hot encoding results in a very sparse and high-dimensional feature space, using a coordinate descent method may be beneficial here. In addition, l_1 regularization can help take into account the correlation between features, allowing us to determine which features are predictive of each other and of mobile usage.

7.3. Public vs. Private Wi-Fi

Lastly, although we currently compare mobile usage between Wi-Fi and cellular connections, we can further break down the Wi-Fi connection type into public vs. private Wi-Fi. For the purposes of this study, we assume from our context that Wi-Fi connections can be accessed basically for "free," since users on Wi-Fi connections do not use up cellular data. However, this is only the case for public Wi-Fi connections; users generally pay for their own private Wi-Fi connections. As mentioned, the distinction between public and private Wi-Fi is not made in this study, but in the future delineating between the two could provide more enlightening insights into the relationship between connection costs and usage. To distinguish between public and private Wi-Fi connections, we can examine the traceroutes from each device for a given measurement.

8. Conclusion

This paper contributes two main findings: the first stems from our exploratory analysis comparing usage patterns between the United States and South Africa, and the second stems from our zero-rating analysis of WhatsApp and Twitter.

1) Although we faced limitations due to the lack of sufficient data, exploring behavioral patterns over a longitudinal period of three years (2013-2015) helped mitigate this issue. Thus, we are fairly confident in most of the results gleaned from the exploratory portion of our work. Our findings regarding the top 5 most used applications in each country are largely in line with the types of applications declared most popular in their respective continents by the Sandvine reports [2, 3]. Notably, we also contrasted the slight preference among US users for cellular connections with the general preference among South African users for Wi-Fi connections, perhaps indicating a higher dedication to conserving cellular data usage in the latter country. Facebook presents an exception to this observation. Arguably the most prevalent social media application in the United States as well as in South Africa, it displays higher average cellular than Wi-Fi data usage for both countries. Due to the possibility of self-reporting errors with regard to data cap types, it was difficult to draw conclusive results from our usage comparisons between users with different data plans.

2) From our analysis of WhatsApp zero-rating on Cell C's network in South Africa, we discovered that mobile data usage on both cellular and Wi-Fi connections greatly increased during the zero-rating period and remained high during a bundle offer period immediately following it, indicating an overall heightened popularity of the application. However, our dataset included only a handful of WhatsApp users on Cell C, so these conclusions are not as robust as we would like. A study on 2014 Twitter zero-rating on MTN's network in South Africa revealed a slight increase in cellular data usage on a typical device, while Wi-Fi usage remained relatively flat. We also noted a slight delay as users responded first to the announcement of the promotional event and then to the end of it. Since there were no Twitter users on MTN's network in 2015, we could not complete our Twitter zero-rating analysis for that year. In addition, because we discovered no Free Basics users in South

Africa at all, we were unable to study the effects of zero-rating the application on Cell C's network in 2015. Thus, although we do present some interesting findings with regard to possible behavioral effects of zero-rating, these results certainly require more study due to the lack of sufficient data.

To this end, a main purpose of our study is to motivate active recruitment of users as a necessary next step. As our conclusions show, this is particularly important in South Africa for the purpose of increasing the number and diversity of users in the country. We especially want to target recruitment toward users on prepaid plans, since these constitute the most prevalent type of data plan in South Africa.

9. Acknowledgments

I would like to thank Nick Feamster for introducing me to this very interesting topic and providing guidance, advice, and feedback throughout the research process.

I greatly appreciate Sarthak Grover's help with troubleshooting and solving the data collection, cleaning, and preprocessing challenges I encountered, as well as for being a reliable source of contact for pertinent issues or questions during the course of my research.

In addition, I would like to extend my gratitude to Enrico Calandro for providing useful context with regard to various zero-rating practices in South Africa as well as relevant quarterly data pricing information in the region. Thanks also to Guilherme Martins for his insights into navigating the MySpeedTest database and cross-checking my results.

Finally, I am grateful as always to my family and James Bartusek for their unwavering support and to Terrace F. Club for the neverending food=love.

10. Honor Code

This paper represents my own work in accordance with University regulations.

Appendices

A. Data Collection, Cleaning, and Preprocessing

Below are the main challenges we faced and solutions we adopted to arrive at the usage table used for analysis.

Challenge 1

We encountered our first challenge during the initial stages of exploring the data directly through psql queries to the database. Because the database is located at Georgia Tech, there was a pronounced latency effect that made it inefficient to interact fluidly with the database on a scalable level.

Solution 1

To remediate this challenge, we decided use an API that performed a single query to the data in the above tables for a certain date range and dump the result into a Pandas dataframe. We then used Python's Pandas library [11] to explore relevant parts of the data in an iPython notebook. This way, the latency effect existed as a one-time occurrence at the initial collection phase, and we were able to bypass any network latencies in our exploration of the data in the notebook.

Challenge 2

When we began to shift from exploring the data to performing some initial analysis, we found that a more pertinent challenge lay in the size of the database. Since the MySpeedTest application collects measurements at 15-minute intervals for every active device/application pair, regardless of whether positive usage was recorded in a given interval, we were faced with potentially billions of rows of data per month. This rendered any sort of longitudinal data manipulation and analysis very inefficient.

Solution 2

Our solution to this problem was to aggregate measurements for each device/application pair by date. To do this in Pandas, we selected the `name`, `deviceid`, `total_sent`, `total_recv`, `connectiontype`, and `time` columns from the dataframe to create a measurements table, and selected the `deviceid`, `networkcountry`, `networkname`, and `datacap` columns to create

a table containing the metadata for each device.

We extracted the date from each measurement's timestamp, and grouped the measurement table by `date`, `name`, `deviceid`, and `connectiontype`, summing `total_sent` and `total_recv` bytes for each group. The resulting dataframe contained the aggregated measurements for each device/application for each connection type by date.

Since our metadata table contained a row for each measurement, it was likely to contain many duplicates and possibly some empty values. So, we first converted all empty strings (which is how empty values are stored in the raw database) into NumPy *nan* values. We then dropped all duplicate rows as well as any rows containing *nan* values, since we only needed one valid row of metadata per device. After performing an inner join between the de-duplicated metadata table and the aggregated measurements table, we checked to ensure that the combined table had no *nan* values.

At this point, we were left with a much more manageable usage table aggregated by date, on which we could perform more efficient analysis.

Challenge 3

However, we realized that a bottleneck still existed in the data collection phase. If any part of a query needed to be changed for any reason, the whole query would need to be rerun on the raw data for the length of the desired time period as described in Solution 1, and the entire cleaning and preprocessing phase outlined in Solution 2 would have to be repeated on the resulting dataframe. This method of implementation was thus still not sufficiently scalable for our longitudinal analysis.

Solution 3

At this point, we realized that it would be best to perform the necessary cleaning, preprocessing, and aggregating directly via psql queries to select, join, and groupby the relevant columns in the above tables. First, we created a new table directly in the database that consisted of measurements for each device/package pair aggregated by date. To do this, we joined the relevant columns in the `application_use`, `measurement`, and `network` tables on `measurementid`, and summed up usage measurements (`total_sent` and `total_recv`) that were taken from the same device/package pair for each connection type on each given day.

This table was much more manageable in size to directly query. After dumping the results of a query on any desired time period into a Pandas dataframe, it was easy in Pandas to add the application `package`→`name` mapping by joining with the `application` table, as well as the metadata associated with each device by joining with the `device` table.

Challenge 4

Once we arrived at this step, we ran into another issue. While the tables resulting from Solutions 1 and 2 had no empty values, those generated with Solution 3 had hundreds of thousands of rows with empty `networkcountry` and `networkname` fields. We discovered that the `device` table had 2136 devices with no `networkcountry` field and 2357 devices with no `networkname` field. Although we realized the API we used in Solution 1 had likely removed all of these devices, we still found the revelation of such a significant amount of missing data to be a problem.

Solution 4

We could only partially improve upon the missing data situation. To do so, we tried to find other tables in the database that also contained metadata associated with each device. We found that the `sim` table contained the `networkcountry` and `operatorname` (equivalent to `networkname`) for each device, although it too had thousands of devices with missing fields in these two columns. However, we hoped to merge together the non-empty fields contributed by these two tables for a more complete picture of each device's metadata. So, we joined the `device` and `sim` tables and dumped the resulting table into a Pandas dataframe. The `device` table and the `sim` table each contributed one country and one carrier (i.e., network/operator name) to this table (the column names were changed to distinguish between the two versions of each field).

To minimize the number of missing values, for each row in the resulting table we selected whichever version of network country had a nonempty field, if any, and used this version as the corresponding device's country. We did the same for each device's carrier name. Although this marginally reduced the number of rows with empty values in the `networkcountry` amd `networkname` fields, we still ended up with a fair amount of devices with empty values in either or both of these two columns. For purposes of this study, we included the measurements

associated with these devices except when drilling down by country or carrier, in which case they got automatically filtered out.

Challenge 5

Once we were finally able to run some initial analysis on the data, we found that our daily-aggregated data usage numbers looked alarmingly high, often differing from the expected range by three or four orders of magnitude. This rendered any observations and results from our analysis inconclusive, since our numbers were nowhere near believable.

Solution 5

After much troubleshooting to better understand where the problem was originating, we traced the discrepancy all the way back to the `total_sent` and `total_recv` columns in the raw database. For a seemingly egregious device that appeared to use 140 GB of Google Play Store on a particular day, we found that consecutive measurements in the raw database displayed the same high `total_sent` number repeatedly for dozens of 15-minute intervals, before increasing to an even higher number and repeating that value for some number of measurements. The `total_recv` column displayed similarly odd behavior.

We then inferred that instead of the `total_sent` and `total_recv` columns representing the number of bytes sent or received during a given time interval, as we had been led to believe, these columns actually represented the byte count recorded at that time interval. This byte count only increased when the device/application pair reported a positive usage level in a given time period. Thus, adding up all the `total_sent` and `total_recv` measurements for a device/application pair on a given day grossly inflated our results.

Instead, aggregation by date for a device/application pair entailed simply subtracting the lowest byte count, presumably associated with the pair's first measurement of the day, from the highest byte count, presumably associated with the pair's last measurement of the day, to get the actual number of bytes the device used on the application on that day. After we changed the logic to create the table described in Solution 3, we reran all of our analysis code and achieved much more reasonable results.

References

[1] "Prepaid 1gb basket." Research ICT Africa. [Online]. Available: http://www.researchictafrica.net/

[2] "Global internet phenomena: 2h 2014." Sandvine Intelligent Broadband Networks, 21 November 2014. [Online]. Available: https://www.sandvine.com/downloads/general/global-internet-phenomena/2014/2h-2014-global-internet-phenomena-report.pdf

[3] "Global internet phenomena report: Africa, middle east & north america." Sandvine Intelligent Broadband Networks, 04 December 2015. [Online]. Available: https://www.sandvine.com/downloads/general/global-internet-phenomena/2015/global-internet-phenomena-africa-middle-east-and-north-america.pdf

[4] E. Alvarez, "Facebook's free internet service expands to malawi." EnGadget.com, 14 May 2015. Available: http://www.engadget.com/2015/05/14/facebook-internet-org-malawi/

[5] Q. Bronkhorst, "Sa mobile subscribers: Vodacom vs mtn vs cell c vs telkom." BusinessTech, 23 April 2015. Available: http://businesstech.co.za/news/mobile/85752/sa-mobile-subscribers-vodacom-vs-mtn-vs-cell-c-vs-telkom/

[6] D. Choujaa and N. Dulay, "Predicting human behaviour from selected mobile phone data points," in *Proceedings of the 12th ACM International Conference on Ubiquitous Computing*, ser. UbiComp '10. New York, NY, USA: ACM, 2010, pp. 105–108. Available: http://doi.acm.org/10.1145/1864349.1864368

[7] S. Ha *et al.*, "Tube: Time-dependent pricing for mobile data," *SIGCOMM Comput. Commun. Rev.*, vol. 42, no. 4, pp. 247–258, August 2012. Available: http://doi.acm.org/10.1145/2377677.2377723

[8] J. D. Hunter, "Matplotlib: A 2d graphics environment," *Computing In Science & Engineering*, vol. 9, no. 3, pp. 90–95, 2007.

[9] S. Leidel, "Zero rating: Why are people using a half-baked internet." DW Akademie, 12 February 2015. Available: http://www.dw.com/en/zero-rating-why-are-people-using-a-half-baked-internet/a-18887956

[10] A. Mathur, B. Schlotfeldt, and M. Chetty, in *Proceedings of the 17th ACM International Conference on Ubiquitous Computing*, ser. UbiComp '15. ACM, 2015, pp. 1209–1220. Available: http://dblp.uni-trier.de/db/conf/huc/ubicomp2015.html#MathurSC15

[11] W. McKinney, "Data structures for statistical computing in python," in *Proceedings of the 9th Python in Science Conference*, S. van der Walt and J. Millman, Eds., 2010, pp. 51 – 56.

[12] L. Mutegi, "Ghana: Facebook, airtel partner to bring internet.org app to ghana." AllAfrica (Nairobi), 26 January 2015. Available: http://allafrica.com/stories/201501270091.html

[13] D. Smith, "Internet use on mobile phones in africa predicted to increase 20-fold." The Guardian, 05 June 2014. Available: http://www.theguardian.com/world/2014/jun/05/internet-use-mobile-phones-africa-predicted-increase-20-fold

[14] J. Stringfield, "Study: Twitter's influence on holiday shopping." Twitter, 07 October 2013. Available: https://blog.twitter.com/2013/study-twitters-influence-on-holiday-shopping