ABSTRACT
Lately there is interest in understanding Internet usage from developing regions such as Africa. Towards this goal we captured an anonymised trace from a cellular operator in Rwanda representing data traffic of 200,000 users for a week in February 2015. In this paper we highlight the key insights that we discovered focusing on device types, content being accessed and its geographical proximity to users.

1. INTRODUCTION
Africa is currently experiencing a high increase in mobile data usage. It is expected that mobile Internet traffic across Africa will increase 20-fold by the end of the decade [1]. This fast changing infrastructure is being led by different entities such as charities and government organisations as well as Internet-scale companies such as Google and Facebook that all see economic and social potential in connecting Africa [2, 3].

The Internet is predominantly accessed via cellular networks in this region due to the lack of fixed-broadband infrastructures [4]. In fact fixed-line broadband and cellular broadband subscribers account for 0.4% and 20% respectively of African population [2]. Given the low-bandwidth high-latency network characteristics, companies design their applications specifically for users in developing countries [5]. However it remains unclear what the African user experience looks like and whether there are specific patterns that can be leveraged to optimise Internet access over cellular networks. Addressing these issues would hopefully enable more people from Africa to access the Internet.

Towards this goal we captured a week worth of anonymised data traffic from a cellular operator in Rwanda representing 200,000 users. We also gather additional datasets (e.g. geographical location of destination IPs) that are important for the purpose of our analysis. We investigate in this paper user behaviour, network utilisation and content of interest to these users. We highlight interesting observations such as the type of devices being used and the corresponding effect on traffic behaviour. We also illustrate the geographical proximity of popular content (such as Google and Facebook) with respect to Rwanda.

We describe the various datasets that we captured for our analysis in Section 2. We present the device types and top brands being used by Rwandan users in Section 3. We then provide insights about the network usage based on these devices in Section 4. We also discuss insights about the popular content being accessed in Section 5. We finally present related work (Section 6) and conclude (Section 7).

2. METHODOLOGY AND DATASETS

2.1 Mobile Data Traffic Capture
There are three cellular operators in Rwanda (MTN, Tigo and Airtel). Our operator provides 2G and 3G data connectivity for its user base. Based on our discussion with the operator and other researchers in the region this user base consists mostly of people who live in cities. For international connectivity, the operator is using two ISPs with a combined bandwidth around 0.5 Gbps.

2.1.1 Passive Measurement from Gn Interface
A typical cellular network architecture consists of radio access and core network. The radio access network connects user equipment to the core network. The IP-based core network includes Serving GPRS Support Nodes (SGSN) and Gateway GPRS Support Nodes (GGSN) enabling data access. The actual traffic between user equipment and the destination host is tunnelled using GPRS tunnelling protocol (GTP) [6].

We choose to capture traffic flowing between SGSN and GGSN (Gn interface) as it provides the required level of visibility for our analysis. We use passive measurement techniques in order to better understand usage behaviour with minimal measurement bias. We leverage traffic mirroring capabilities in modern switches in order to capture a copy of the traffic flowing in the core network and dump it to a storage server as shown in Figure 1. We captured all traffic from Friday 6 February 2015 7pm to Thursday 12 February 2015 11am local time. Concurrently we had a data anonymisation process running on the captured trace to discard any potentially sensitive data and store the resulting dataset in encrypted form (as we will describe in Section 2.1.2). As we discussed before the maximum mobile data bandwidth provisioned by the operator is 0.5 Gbps. A single stor-
We captured TAC codes in our trace. We leveraged a map-

different device types used to access content. We observed

2.2 Device Type Database

Storage Server

Figure 1: Cellular network architecture and tap point

age server with a gigabit interface is capable of handling this
load. We use tcpdump to capture the mirrored traffic from the
switch to a RAID array. As all packets are captured by
one host, timestamp synchronisation is not a problem. We
also did not observe any significant increase to the switch
CPU load when traffic mirroring is enabled. Therefore we
believe that the process of traffic mirroring did not affect the
timing of the packets or the performance of the switch for
the purpose of our study.

Our captured traffic represents 200,000 unique users with
13 TB and 1.25 TB total download and upload volume re-
spectively. The trace contains over 30 Billions raw IP pack-
ets.

2.1.2 Dealing with Sensitive Data

We take user privacy very seriously. In this project we
comply with best practice enforced by the university ethics
committee [7]. We do not capture any application level pay-
load (user content); we only store TCP/UDP packet headers
including GTP user headers. The only exception is DNS to
enable us understand the type of content being accessed. We
also capture GTP control packets. However, any sensitive
information that can infer user identity such as International
Mobile Subscriber Identity (IMSI), Mobile Subscriber Num-
ber (MSISDN) and International Mobile Station Equipment
Identity (IMEI) is hashed using SHA256 with salt. We only
keep Type Allocation Code (TAC); the initial eight-digit por-
tion of IMEI [8] to be able to identify the type of devices
used. Moreover we do not know the geographical location
of users of this study.

This data anonymisation process was done in Rwanda to
comply with rules and regulations. The resulting dataset is
stored encrypted using AES256 and the raw trace was de-
stroyed.

2.2 Device Type Database

As we discussed in the previous section, we captured TAC
codes from GTP control packets to be able to identify the
different device types used to access content. We observed
6,000 unique TAC codes in our trace. We leveraged a map-
ping database from imeidata.net to convert these codes
to models.

We gathered the different features of these models from
additional online databases. Based on these curated fea-
tures we classify these devices into six categories: (i) fea-
ture phone, (ii) smart phone (with touchscreen), (iii) usb,
(iv) tablet, (v) router (including MiFi) and (vi) modem (Ta-
ble 3).

2.3 IP to ISP/Content Database

For the purpose of our analysis, we map a given destina-
tion IP in a TCP/UDP packet to additional metadata about
the ISP, domain name and content being served by this des-
tination. In order to construct this mapping, we rely on three
data sources.

Leveraging MaxMind GeoLite Autonomous System Num-
ber (ASN) [9] database, we map a given IP to the register-
ing ISP or organisation. We also use information from DNS
queries that we captured to tag an destination IP with the
user-requested domain name. Thirdly, we performed reverse
DNS queries to Google public DNS for all (16 Millions) desti-
nation IPs that we have found in the trace.

For example, we map 173.194.112.20 to Google Inc.,
google.com and fra07s27-in-f20.1e100.net. This enables us
to have insight about the content being served from these IPs
in addition to other useful metadata as we will discuss next.

2.4 IP to Geographical Location Database

Being able to accurately pinpoint geographical locations of
servers is imperative to this study. For example we want
to understand whether content being served is located in the
African continent. Relying solely on standard GeoIP databases
e.g. MaxMind GeoLite2 [10]) would lead to wrong results
because: (i) these databases use registration information for
location mapping (i.e. IPs registered by Google Inc. would
be incorrectly located in the U.S.) and (ii) anycast routing
traffic is sent to the nearest destination (i.e. DNS queries to
8.8.8.8 are routed to the closet server, which is dependent on
the location of clients).

To circumvent this challenge we use international-link re-
sponse times to select IPs that are incorrectly located by
MaxMind GeoLite2. Similar to related work [11], we work
out international-link (SGSN to destination) latency from
TCP SYN-SYN/ACK-ACK messages (Figure 1).

For example we inspect IPs that are registered in the U.S.
but with response times below 200 ms or in Europe but with
response times below 150 ms. These values are chosen based
on typical round trip latencies from Rwanda to the U.S. and
Europe.

For this set, we performed traceroute and ping to
each IP from a node in Kenya. Because of Rwanda prox-
imity to Kenya, we believe that these results are similar for
users from these two locations. We also compared response
times for these active measurements with the response times
that we observed in the original traffic trace to confirm that
they are similar. From traceroute and ping results, we can infer the location of a given IP from the hop count, location of last hop and end-to-end latency. To reduce uncertainties we choose to map location at continent granularity. However we drill down in Africa; we get to country granularity by doing ping to a given IP from different locations in Africa (using ISP looking glass) until we get to the smallest latency for which the corresponding location is tagged to the IP. We check the correctness of our results by doing random checks to compare the assigned location with encoded location in DNS and ISP names that we have (as discussed in Section 2.3).

We acknowledge that this best-effort methodology comes with some uncertainties. We performed the active measurements after we captured traffic from the operator. With the dynamic nature of the Internet, things could look differently. For example CloudFlare has announced a point of presence in Kenya that is in operation since August 2015 [12]. In February 2015, we think that CloudFlare traffic was served from either South Africa or the U.A.E. (based on response times). However we have not noticed any other discrepancies.

2.5 Data Processing

To handle this data scale, we provisioned a 7-node cluster running Apache Hadoop [13] and Apache Spark [14] for data processing and analysis. We use Apache Zeppelin [15] for interactive data exploration and visualisation. We also developed our system in way that makes modifying parameters and reproducing results automatically executed.

GTP is a tunnelling protocol that encapsulates the actual IP traffic. We extract attributes from context creation and modification messages and link them to the actual user traffic. In this way we can classify TCP/UDP traffic according to the characteristics of the corresponding user. For example we can tag an IP packet with the connection speed (2G/3G) and the device type. We also have a notion of a session defined from tunnel context setup to tear down.

In order to follow guidelines set by our ethics committee dealing with sensitive data, we do not capture any application level details (with the exception of DNS). Nevertheless we can still infer some high-level information about the type of traffic content being accessed using the following heuristics.

We rely mainly on DNS to tag IPs with the user-requested domain names. We augment this set with results from our reverse DNS lookups and ASN mappings as we discussed in Section 2.3. Additionally we collected public information about IPs and ports used in popular applications. For example we classify traffic to Amazon Inc. registered IPs on ports 5242–5245,4244 and 9785 as Viber traffic [16]. When we observe that a specific destination IP belongs to an end-user on a tier3 ISP, we classify traffic as peer-to-peer. For CDN (e.g. Akamai and CloudFlare) traffic, we infer the type of content from the initial user-requested domain names by-passing any intermediate mappings.

3. DEVICES

<table>
<thead>
<tr>
<th>Type</th>
<th>Device count %</th>
<th>Downstream volume %</th>
</tr>
</thead>
<tbody>
<tr>
<td>feature phone (e.g. ITEL IT5130): phones without touchscreen</td>
<td>18.8%</td>
<td>3.4%</td>
</tr>
<tr>
<td>smart phone (e.g. TECNO P5): phones with touchscreen</td>
<td>15.6%</td>
<td>29.2%</td>
</tr>
<tr>
<td>USB (e.g. ZTE MF90): devices accessed via USB</td>
<td>3.7%</td>
<td>33.1%</td>
</tr>
<tr>
<td>router (e.g. HUAWEI B681): devices accessed via WiFi</td>
<td>0.5%</td>
<td>19.3%</td>
</tr>
<tr>
<td>Tablet (e.g. HUAWEI S7): devices with large touchscreen</td>
<td>0.5%</td>
<td>3.2%</td>
</tr>
<tr>
<td>Modem (e.g. HUAWEI E881): PCMCIA devices and GPRS modules</td>
<td>0.6%</td>
<td>11.8%</td>
</tr>
</tbody>
</table>

Table 1: Device types

We observed in our traces around 200,000 uniques users that were accessing content using over 6,000 unique devices. From the TAC codes database that we assembled (Section 2.2), we classify these devices intro six types based on their features (Table 3) and aggregate the corresponding device count and download volume.

Around 80% of users are feature phones. Feature phones are first and second generation phones from 2000–2005 that are mainly designed for voice features. They offer constrained means to access the Internet though pre-packaged applications and they are not typically 3G enabled.

The top two feature phones models are ITEL IT5130 and ITEL T340, which constitute 30% of feature phone users. These two models cost 27$ and 30$ respectively making them widely affordable in a developing country like Rwanda. However feature phones consume minimal traffic (4% of downstream volume) mainly due to their limitations.

On the other hand, routers exhibit a converse behaviour to feature phones. At just 0.5% of the number of devices, they consume around 20% of downstream traffic. We expect the Internet to be accessed from a PC or laptop via these routers, which enables a much wider range of Internet usage compared to feature phones.

The top two routers in terms of downstream volume are HUAWEI B681 and ZTE MF29A. They are typically installed at home or small office environment in these regions. Routers represent a different segment of users anyway.

Generally usb, smart phones and router devices account for 78% of the downstream traffic volume while constituting 18% of the observed user base. This long tail behaviour of downstream traffic per device type is an indication of the limitations posed by feature phones on their users despite the low cost.

3.1 Brands and Models

Taking a closer look at top models (Table 2) and brands (Table 3) we observe that ZTE and HUAWEI account for...
4. NETWORK BANDWIDTH USAGE

In this section we explore upstream and downstream network usage per hour. We explore patterns for different device types. We also highlight the main geographical regions from which traffic originates and terminates.

4.1 Users

Overall traffic peaks at 350 Mbps and 35 Mbps for downstream and upstream respectively, which represents 10:1 ratio. Figure 2 illustrates repeated daily patterns. As our traces start from Friday 6 February 2015 7pm Rwanda local time, these figures mainly show traffic patterns for a whole weekend and three full weekdays.

Figure 2(a) and Figure 2(d) break down traffic by device type. As we discussed before smart phone, usb and router devices have the largest contribution.

For smart phone users, we observe generally more traffic during the weekend and weekdays’ evenings. This is probably reflecting outside working hours pattern. On the other hand, router users consume more traffic during weekdays’ up to 5pm then tailing off. We believe that some of these routers are installed in office environments. We also note that tablet users have a small contribution. This might be related to financial barriers.

4.2 Geographical Locations

Based on the geographical mapping of IPs that we discussed in Section 2.4, we classify by continent upstream and downstream traffic in Figure 2(b) and Figure 2(e) respectively.

For downstream, approximately 35% of traffic is served from Africa while 40% of traffic comes from Europe. The remaining 25% is split across the other continents (12% in North America, 8% in Asia and 5% for the rest). We highlight that contrary to our expectation, a large portion of downstream traffic is located close to users. This is due to the presence of a few content distribution networks (e.g. Akamai and CloudFlare) as well as major providers (e.g. Google) in the region.

Table 2: Top models by downstream volume

<table>
<thead>
<tr>
<th>Model</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>HUAWEI B881 (router)</td>
<td>12.1%</td>
</tr>
<tr>
<td>ZTE MF190 (usb)</td>
<td>11.7%</td>
</tr>
<tr>
<td>ZTE MF190U (usb)</td>
<td>10.0%</td>
</tr>
<tr>
<td>HUAWEI E881 (modem)</td>
<td>7.8%</td>
</tr>
<tr>
<td>ZTE MF29A (router)</td>
<td>4.3%</td>
</tr>
<tr>
<td>HUAWEI E303H (usb)</td>
<td>4.0%</td>
</tr>
<tr>
<td>TECNO P5 (smart phone)</td>
<td>1.9%</td>
</tr>
<tr>
<td>ZTE MF667 (usb)</td>
<td>1.8%</td>
</tr>
</tbody>
</table>

Table 3: Top brands by downstream volume

<table>
<thead>
<tr>
<th>Brand</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>ZTE</td>
<td>30.0%</td>
</tr>
<tr>
<td>HUAWEI</td>
<td>29.1%</td>
</tr>
<tr>
<td>SAMSUNG</td>
<td>13.7%</td>
</tr>
<tr>
<td>TECNO</td>
<td>7.6%</td>
</tr>
<tr>
<td>APPLE</td>
<td>4.0%</td>
</tr>
<tr>
<td>ITEL</td>
<td>2.2%</td>
</tr>
<tr>
<td>NOKIA</td>
<td>1.8%</td>
</tr>
<tr>
<td>HTC</td>
<td>1.4%</td>
</tr>
</tbody>
</table>

On the other hand, upstream traffic exhibit a different behaviour. Although Europe contribution remains at 40%, upload traffic to North America is significant.

We are interested in understanding which geographical locations in Africa serve content for Rwandan users. Figure 2(c) and Figure 2(f) illustrate the top three African countries in term of upstream and downstream traffic respectively.

For downstream, Kenya comes first with 55% of the African-based traffic, followed by Rwanda and South Africa at 32% and 11% respectively. Lately Kenya has been playing a key role as a technology hub for East Africa, which attracts providers there. On the other hand, upstream traffic to African counties albeit minimal is more concentrated at Rwanda and South Africa with minimal traffic going to Kenya. In the next section we investigate whether this is related to the type of content being accessed.

5. CONTENT

In this section we analyse the top content being accessed by users based on device type, traffic direction (i.e. upstream and downstream), and geographical location. As discussed in Section 2.3 we use heuristics that rely on DNS, IP registration and port number to tag TCP/UDP streams with information about the corresponding content.

We group services belonging to the same provider together. For example we refer to windows update and bing traffic as microsoft. We also group content of a specific type together (e.g. news and antivirus). For Google related content, we classify it as googlevideo (e.g. youtube) or google (e.g. Google search).

There are two special categories that we define: (i) browser traffic that is served by companies such as Opera and Mozilla and (ii) p2p traffic on random ports and from tier3 ISP-based hosts.

Regardless of the device type being used, we generally notice a large portion of downstream traffic belonging to googlevideo and google. We also observe a significant p2p and facebook traffic with more weight on upstream traffic.

Browser dominates downstream traffic for feature phone users (Figure 3(a)) while instagram is being accessed by smart phone devices (Figure 3(b)). We observe whatsapp traffic for both type of handheld devices.

On the other hand router (Figure 3(c)) and usb (Figure 3(d)) devices consume slightly different content. Due the fact that typically a PC or laptop is being used to access the Internet via these devices, we observe microsoft (e.g. Windows update) and antivirus related traffic. It is interesting to find that users are accessing these services over 3G (as opposed to fixed-line broadband), which illustrates that it is their only means of connectivity.

We investigate top content based on the geographical location of destination servers. We observe that almost all traffic from Kenya (Figure 3(f)) belongs to Google service. In fact googlevideo from Kenya is the top content in terms of downstream volume across all locations. As we discussed before,
Figure 2: Bandwidth usage (average per hour and Hour 0 is Friday 7pm Rwanda local time)

upstream traffic to Kenya is minimal.
Content hosted in Rwanda (Figure 3(e)) has different characteristics. It serves traffic for Microsoft, Apple and Facebook services. We also note that the top content in terms of upstream volume is for Rwanda e-government (gov).
Europe (Figure 3(h)) plays a similar role as Kenya serving a large volume of Google traffic. Additionally Facebook traffic is largely from Europe. We observe also large p2p traffic from Europe. On the other hand, North America (Figure 3(h)) serves traffic that is not typically hosted anywhere else such as whatsapp and dropbox with more weight on upload traffic as well.

We explore further the contribution per geographic location for a given content, which illustrates how close it is for Rwandan users. We focus on Facebook and Google services as well as popular communication applications such as Viber, Whatsapp and Skype.

Figure 5 shows the break down for downstream traffic. We observe that Viber traffic is the closest to users with around 90% being served from within Africa and mostly from Rwanda. On the contrary, Whatsapp traffic is the furthest coming exclusively from North America. Google traffic is evenly balanced between Europe and Kenya while Facebook and Skype services are being served mostly from Europe.

Upstream traffic going to these services travels relatively further as shown in Figure 4(a). All Viber and Whatsapp traffic go to North America while Skype is evenly balanced between North America and Europe. Upload traffic to Facebook and Google is relatively closer terminating in Europe. Only googlevideo has users uploading content to African based servers (around 20% to Kenya).

6. RELATED WORK
There have been a few studies lately about the African web ecosystem. Zaki et al. [17] analysed web performance in Ghana illustrating challenged in slow DNS resolution and a lack of local caching. Fanou et al. [18] investigated Google Content Cache for African users showing that a large portion of traffic is served from North America and Europe. Schmitt et al. [19] analysed quality of services for Facebook and Google being access from a village in Zambia [19]. These active measurement projects are complemented by our study for a better understanding of user behaviour.

There are also many studies about mobile data usage using similar techniques as the ones we employed [20, 21, 11, 22, 23, 24]. We plan as future work to compare these results with the insights that we got for Rwandan users.

7. CONCLUSION AND FUTURE WORK
In this paper we studied country-scale mobile data usage from a cellular operator in Rwanda. We captured anonymised user traffic for a week in February 2015 and collected additional datasets (e.g. accurate GeoIP information) to aid our
analysis. We discussed key insights about user behaviour and content being accessed.

We showed the widespread dominance of Chinese brands in this market. We observed that many users are still using feature phones for Internet access. However we also illustrated that most of the traffic comes from usb and router devices, which is a strong indication that cellular data access is the main mean of connectivity in this region.

Google and Facebook services are popular to Rwandan users. A large portion of Google traffic is being served from the African continent (mainly from Kenya). However Europe is still the main geographical location from which content is served. For communication applications such as Viber and Whatsapp, user upload traffic travels a long way to North America.

For future work, we want to investigate more details about the performance (i.e. latency) of cellular networks in the region. This will enable us propose network optimisation techniques for specific usage.

8. REFERENCES

http://www.informationweek.com/software/social/facebook-overhauls-android-app-for-emerging-countries/d/d-
id/1278740, 2014.


