On the Structure and Characteristics of User Agent Strings

IMC 2017
November 2
London
United Kingdom

Jeff Kline & Aaron Cahn (*comScore*)
Paul Barford (*comScore, University of Wisconsin - Madison*)
Joel Sommers (*Colgate University*)

comScore

Introduction and Motivation

About comScore

- We measure and report on audiences for publishers, brands, app developers, etc.

- To measure this, we need the data. To get the data, we partner with brands, publishers, app developers, etc.

- The result is telemetry with worldwide reach. Our telemetry is deployed by major publishers, campaigns and apps.

- Volume on a typical day is ~50B records; each record represents an HTTP(S) request.

- We also maintain a large research panel, we measure TV traffic…

- comScore Labs is the research arm of comScore. It is based in Madison, Wisconsin. We have strong academic roots.

Study Objectives

Describe the User Agent (UA) space from the perspective of a large-scale real-world data corpus

- How large is the space?
- How does it evolve over time?
- How well does the UA fulfill its purpose?
- What about anomalies?

UA History

The UA is transmitted as part of the HTTP header

RFC 1945

10.15 User-Agent

The User-Agent request-header field contains information about the user agent originating the request. This is for **statistical purposes**, the tracing of protocol violations, and **automated recognition** of user agents for the sake of tailoring responses to avoid particular user agent limitations.

…

Example:

```
User-Agent: CERN-LineMode/2.15 libwww/2.17b3
```

Reference: https://tools.ietf.org/html/rfc1945#page-46

About the study's data

Archive spanning 2 year time window. Day 0 is January 1, 2015.

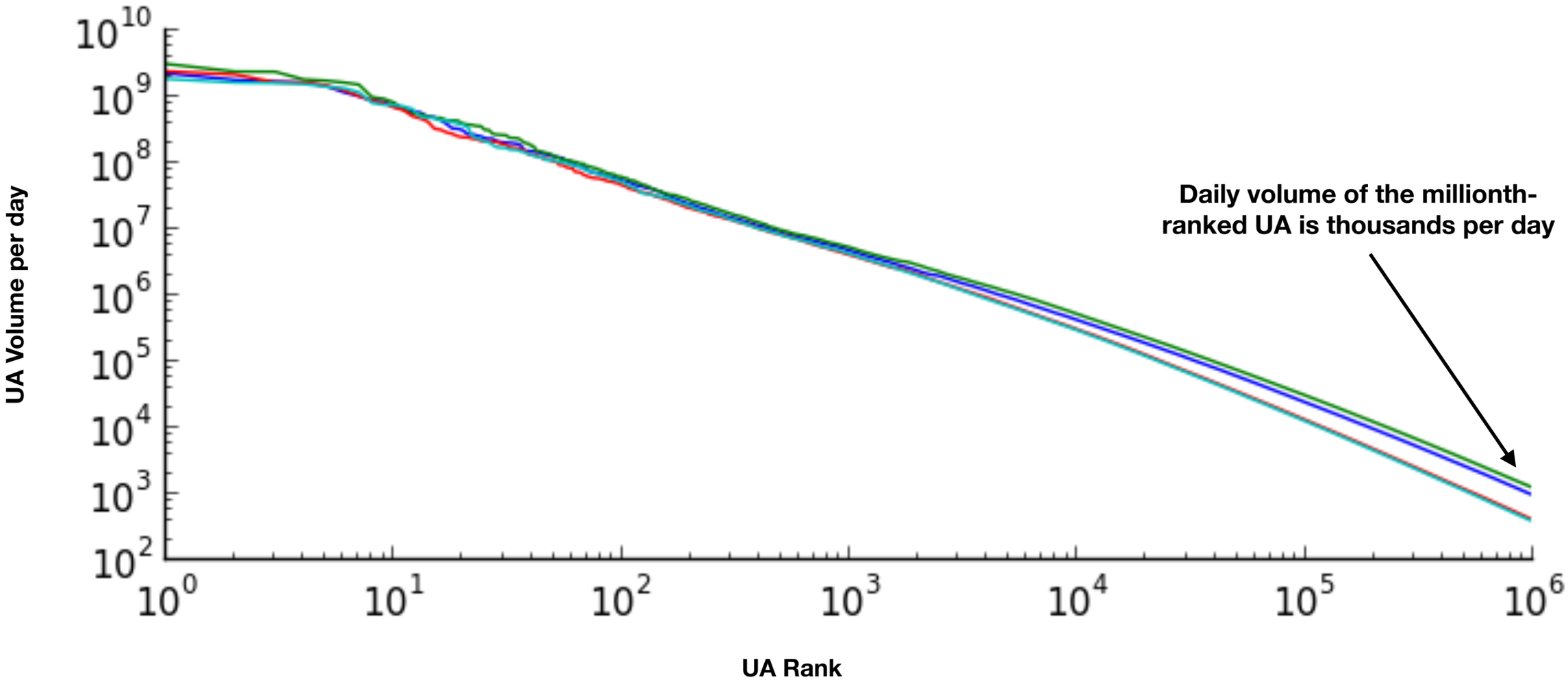
Each record in the archive records the `Volume` of requests that `UA` issued to our web servers on `Day`. The schema is:

```
Day    UA    Volume
```

The long tail

The number of distinct UA's encountered per day is O(millions).

A lot of the mass of the rank-order distribution lives in the tail.
The tail is long and there is no clear threshold.



UA Aggregation

Aggregating over UA is a basic task of web log analysis

- Want to get Android or iPhone traffic? This almost works…

```
.*Android.*
.*iPhone.*
```

- Chrome only traffic? Tablets? MS Edge? QQ? FB? OTT? Apps? IOT?

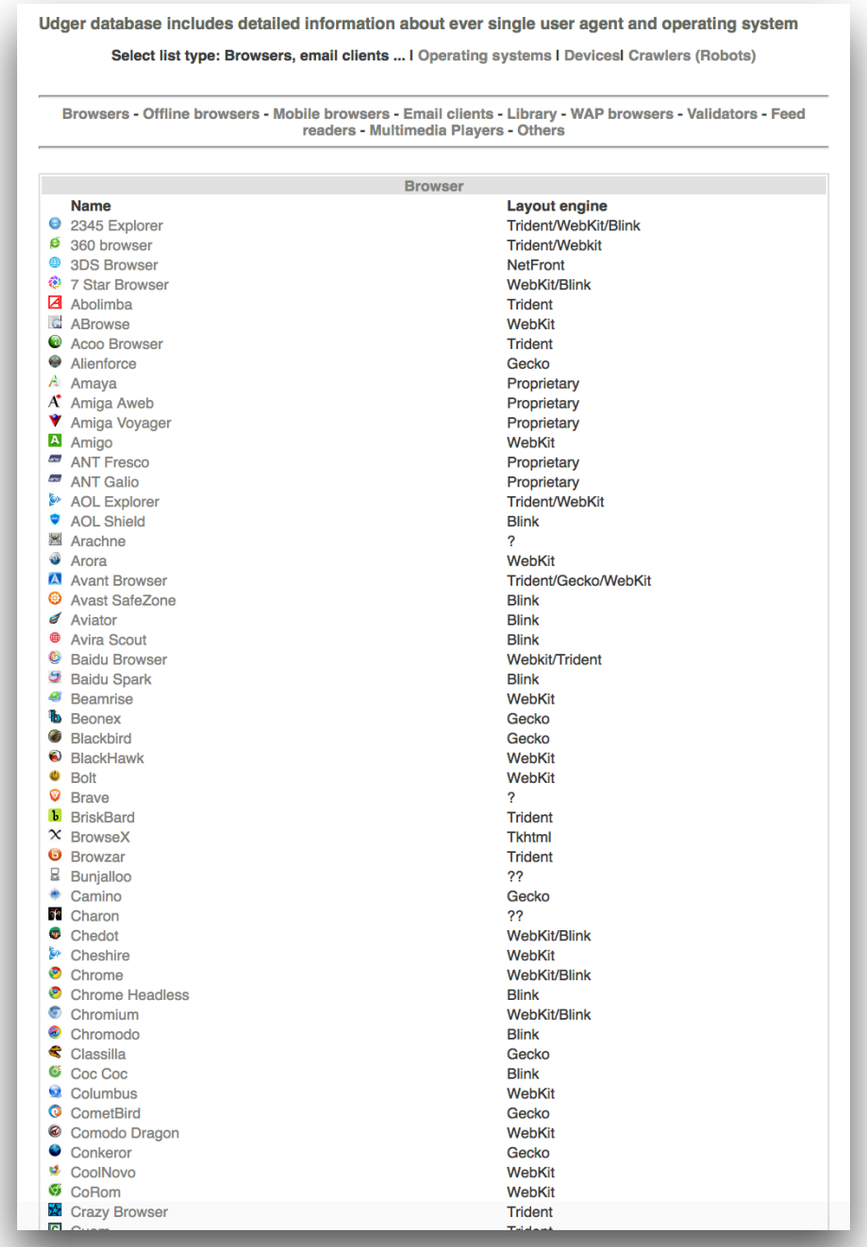
- Even for these simple questions, this task is complicated. Long-term maintenance is a challenge.

- Validation?

- comScore's internal categorization code-base is thousands of lines long

A publicly-visible and independent view that illustrates UA complexity

https://udger.com/resources/ua-list



Udger database includes detailed information about ever single user agent and operating system

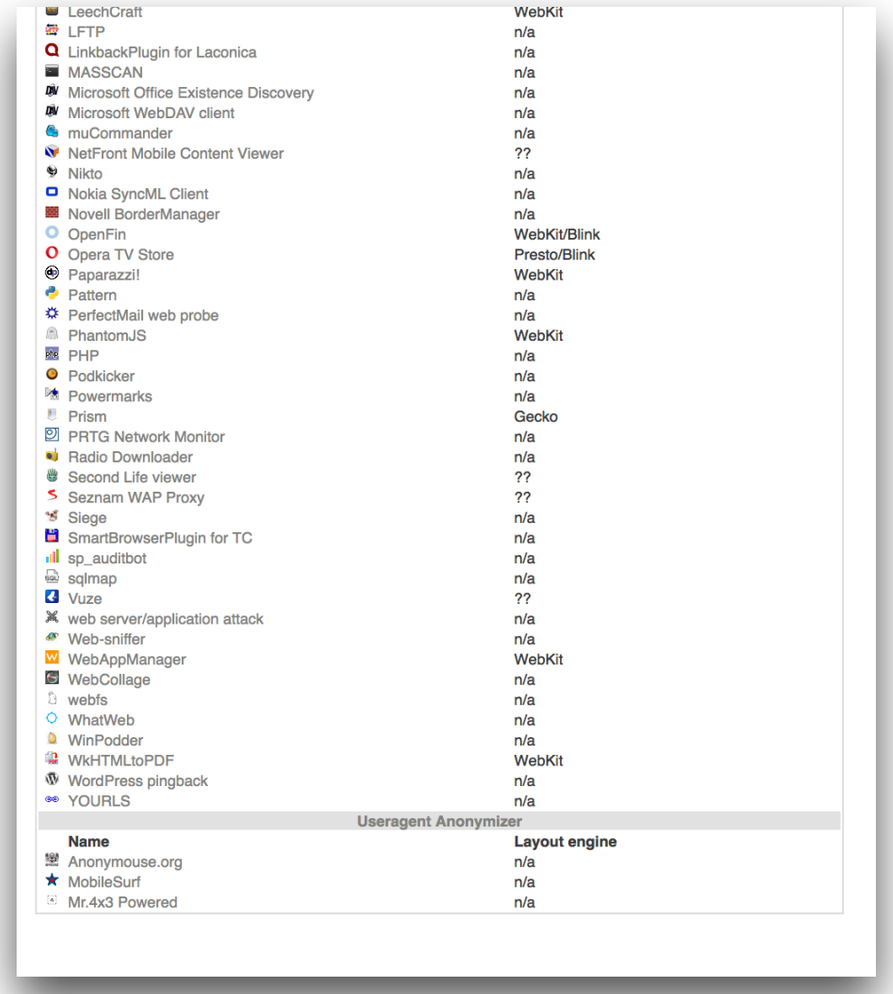
Select list type: Browsers, email clients ... I Operating systems I DevicesI Crawlers (Robots)

Browsers - Offline browsers - Mobile browsers - Email clients - Library - WAP browsers - Validators - Feed readers - Multimedia Players - Others

Browser

Name	Layout engine
2345 Explorer	Trident/WebKit/Blink
360 browser	Trident/Webkit
3DS Browser	NetFront
7 Star Browser	WebKit/Blink
Abolimba	Trident
ABrowse	WebKit
Acoo Browser	Trident
Alienforce	Gecko
Amaya	Proprietary
Amiga Aweb	Proprietary
Amiga Voyager	Proprietary
Amigo	WebKit
ANT Fresco	Proprietary
ANT Galio	Proprietary
AOL Explorer	Trident/WebKit
AOL Shield	Blink
Arachne	?
Arora	WebKit
Avant Browser	Trident/Gecko/WebKit
Avast SafeZone	Blink
Aviator	Blink
Avira Scout	Blink
Baidu Browser	Webkit/Trident
Baidu Spark	Blink
Beamrise	WebKit
Beonex	Gecko
Blackbird	Gecko
BlackHawk	WebKit
Bolt	WebKit
Brave	?
BriskBard	Trident
BrowseX	Tkhtml
Browzar	Trident
Bunjalloo	??
Camino	Gecko
Charon	??
Chedot	WebKit/Blink
Cheshire	WebKit
Chrome	WebKit/Blink
Chrome Headless	Blink
Chromium	WebKit/Blink
Chromodo	Blink
Classilla	Gecko
Coc Coc	Blink
Columbus	WebKit
CometBird	Gecko
Comodo Dragon	WebKit
Conkeror	Gecko
CoolNovo	WebKit
CoRom	WebKit
Crazy Browser	Trident

8 pages later…



LeechCraft	WebKit
LFTP	n/a
LinkbackPlugin for Laconica	n/a
MASSCAN	n/a
Microsoft Office Existence Discovery	n/a
Microsoft WebDAV client	n/a
muCommander	n/a
NetFront Mobile Content Viewer	??
Nikto	n/a
Nokia SyncML Client	n/a
Novell BorderManager	n/a
OpenFin	WebKit/Blink
Opera TV Store	Presto/Blink
Paparazzi!	WebKit
Pattern	n/a
PerfectMail web probe	n/a
PhantomJS	WebKit
PHP	n/a
Podkicker	n/a
Powermarks	n/a
Prism	Gecko
PRTG Network Monitor	n/a
Radio Downloader	n/a
Second Life viewer	??
Seznam WAP Proxy	??
Siege	n/a
SmartBrowserPlugin for TC	n/a
sp_auditbot	n/a
sqlmap	n/a
Vuze	??
web server/application attack	n/a
Web-sniffer	n/a
WebAppManager	WebKit
WebCollage	n/a
webfs	n/a
WhatWeb	n/a
WinPodder	n/a
WkHTMLtoPDF	WebKit
WordPress pingback	n/a
YOURLS	n/a

Useragent Anonymizer

Name	Layout engine
Anonymouse.org	n/a
MobileSurf	n/a
Mr.4x3 Powered	n/a

Selected UAs

```
Mozilla/5.0 (Windows NT 10.0) AppleWebKit/537.36 (KHTML, like Gecko)  Chrome/42.0.2311.135
Safari/537.36 Edge/12.10136


Dalvik/2.1.0 (Linux; U; Android 5.1; F100A Build/LMY47D)


Mozilla/5.0 (Windows NT 6.1; WOW64; Trident/7.0; rv:11.0) like Gecko


Instagram 17.0.0.15.91 Android (22/5.1.1; 240dpi; 480x782; LGE/lge; LGL52VL; m1; m1; en_US)


UCWEB/2.0 (MIDP-2.0; U; Adr 4.2.2; en-US; Micromax_A76) U2/1.0.0 UCBrowser/10.7.9.856 U2/1.0.0
Mobile


Mozilla/5.0 (Windows NT 6.1; WOW64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/29.0.1547.59
Mozilla/5.0 (Windows NT 6.1; WOW64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/29.0.1547.59
Safari/537.36 QQLive/9212159/50170335 Safari/537.36


Mozilla/5.0 (iPhone; CPU iPhone OS 10_3_1 like Mac OS X) AppleWebKit/603.1.30 (KHTML, like
Gecko) Mobile/14E304 [FBAN/FBIOS;FBAV/91.0.0.41.73;\ FBBV/57050710;FBDV/iPhone8,1;FBMD/iPhone;
FBSN/iOS;FBSV/10.3.1;FBSS/2;FBCR/Verizon; FBID/phone;FBLC/en_US;FBOP/5;FBRV/0]
```

These are not really “long tail”. Each has millions of records per day.

Selected UAs

```
Mozilla/5.0 (Windows NT 10.0) AppleWebKit/537.36 (KHTML, like Gecko)  Chrome/42.0.2311.135
Safari/537.36 Edge/12.10136
```

(Not Chrome)

```
Dalvik/2.1.0 (Linux; U; Android 5.1; F100A Build/LMY47D)
```

(Not the Yamaha outboard motor)

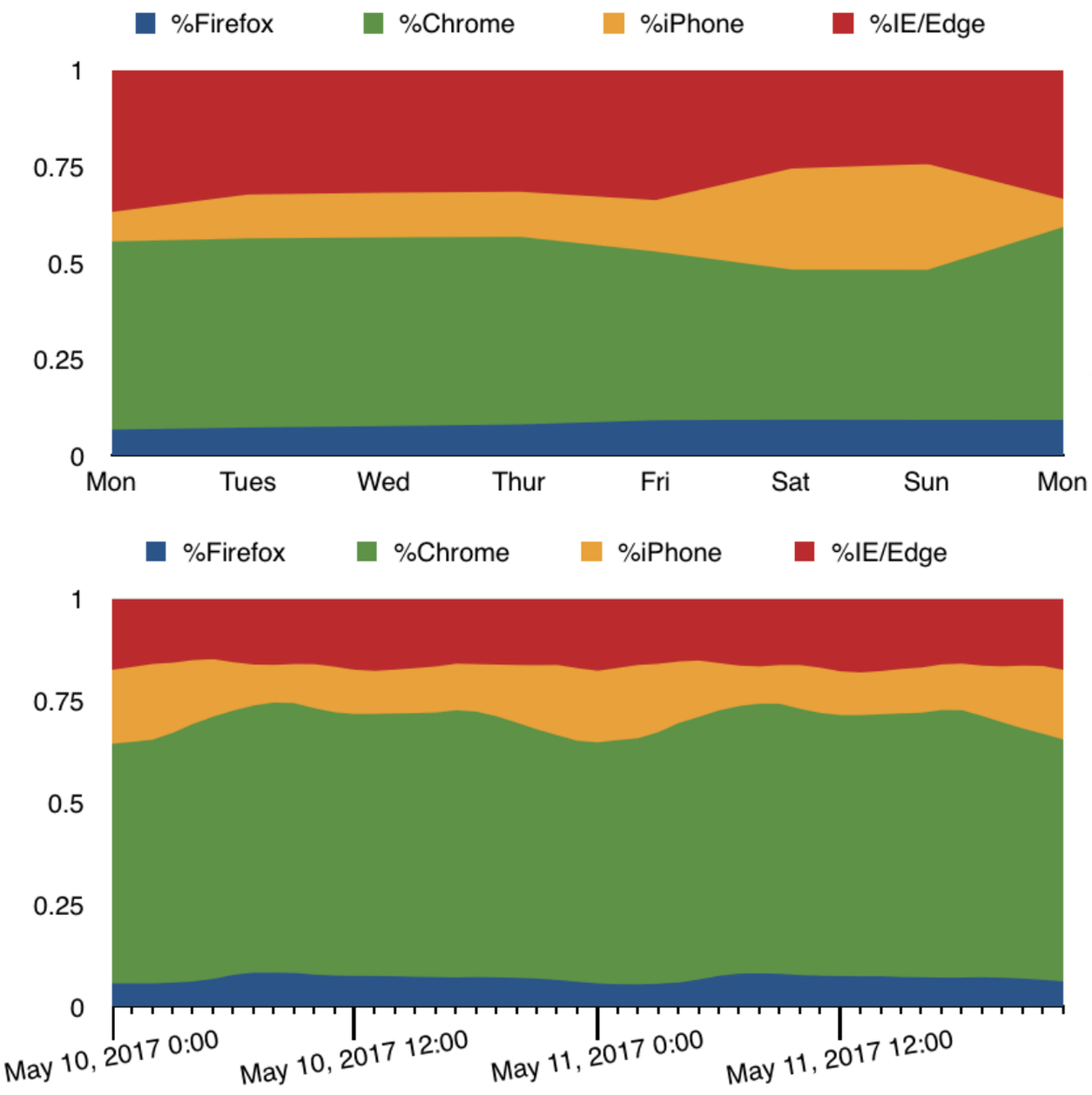
```
Mozilla/5.0 (Windows NT 6.1; WOW64; Trident/7.0; rv:11.0) like Gecko

Instagram 17.0.0.15.91 Android (22/5.1.1; 240dpi; 480x782; LGE/lge; LGL52VL; m1; m1; en_US)

UCWEB/2.0 (MIDP-2.0; U; Adr 4.2.2; en-US; Micromax_A76) U2/1.0.0 UCBrowser/10.7.9.856 U2/1.0.0
Mobile

Mozilla/5.0 (Windows NT 6.1; WOW64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/29.0.1547.59
Mozilla/5.0 (Windows NT 6.1; WOW64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/29.0.1547.59
Safari/537.36 QQLive/9212159/50170335 Safari/537.36
```

(Old Chrome? Old Chrome?)

```
Mozilla/5.0 (iPhone; CPU iPhone OS 10_3_1 like Mac OS X) AppleWebKit/603.1.30 (KHTML, like
Gecko) Mobile/14E304 [FBAN/FBIOS;FBAV/91.0.0.41.73;\ FBBV/57050710;FBDV/iPhone8,1;FBMD/iPhone;
FBSN/iOS;FBSV/10.3.1;FBSS/2;FBCR/Verizon; FBID/phone;FBLC/en_US;FBOP/5;FBRV/0]
```

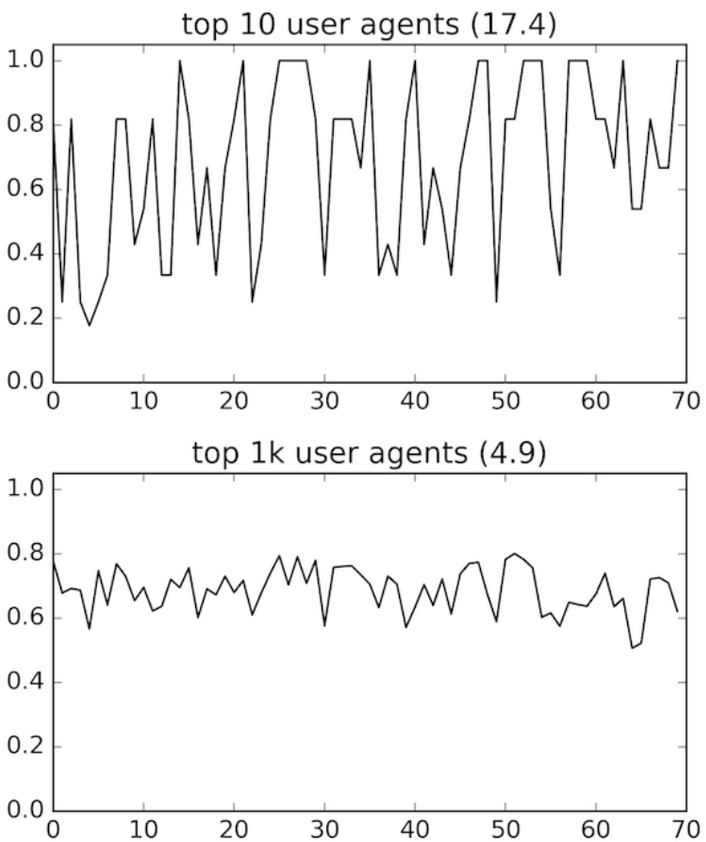
These are not really “long tail”. Each has millions of records per day.

Time-dependent features of the UA distribution



Hour-of-day and day-of-week matter in the UA distribution. This matters for results that relate PII to the UA.

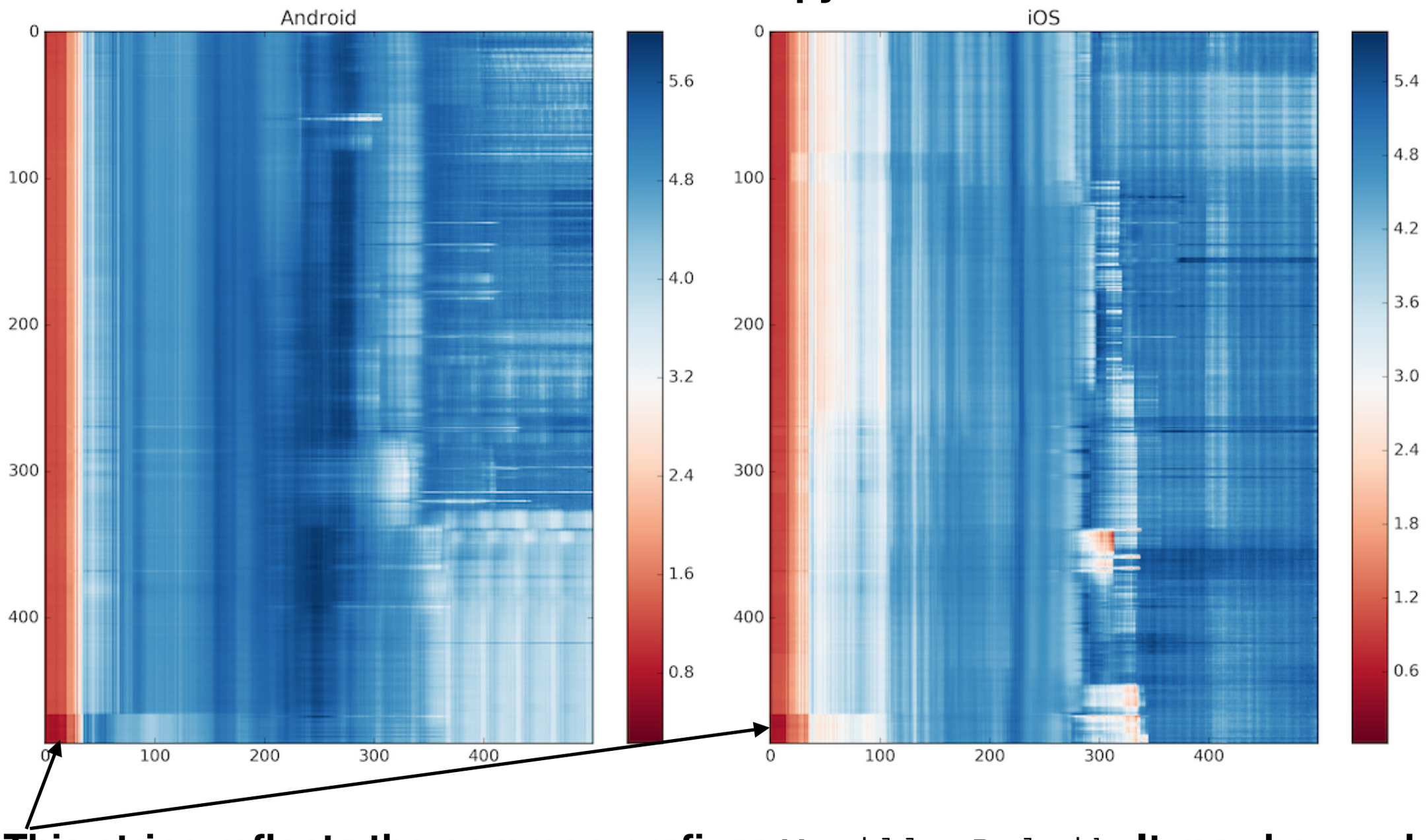
Time-dependent features of the UA distribution



The top 1k UA's churn in a stable manner.

The top 1k week-over-week sets have Jaccard similarity of ~0.7.

The UA space over time



This stripe reflects the common prefixes `Mozilla`, `Dalvik`. It may be used in conjunction with the legend to help interpret the representation.

Lessons

- UA categorization and parsing is (still) a challenge. This task is basic to web log analysis.

- The UA space is diverse and dynamic.

    - The week-over-week Jaccard similarity of the top 1k is relatively stable at about 0.7.

    - UA distribution depends on time-of-day and day-of-week (among other things)

- Introduce the character entropy matrix. It is simple to construct, interpret and it has been used to expose unexpected features within the UA-space.

If the community expresses interest, we will try to make a portion of our UA set available for academic research.

Thank you.

jkline@comscore.com
Jeffery Kline