Exploring Web Analytics to enhance Cyber Situational Awareness for the Protection of Online Web Services

Cyril Onwubiko
Cyber Security Intelligence, E-Security Group, Research Series, London, UK

Abstract— Web Analytics is a tool for monitoring online interactions to digital services, typically focused on entity profiling and analysis for market campaign, user behaviour, site performance and market intelligence. In this research, web analytics is applied for intelligence-centric data gathering and analysis to enhanced cyber situational awareness for monitoring critical online web services. A number of intelligence sources such as web logs, browser fingerprints, mobile and tablet fingerprints and endpoint fingerprint are gathered, fused, analysed in real time for enhanced situational awareness for the protection of online web services.

Keywords— CyberSA, Web Analytics, Digital Service, Entity Profiling, Social Intelligence, User Behaviour, Google Analytics

Introduction

Most businesses today have a website. A website can be used for vital personal, social or business purposes. Most websites are used for a number of use cases. For example, while Twitter is used for social interactions, most companies have found it a great source to advertise their products and services, and hence use it as a business tool. Likewise, the Twitter owners operate it as a business, and hence it serves a business function to them, while still serving a number of social needs as a social network. The same can be said of Facebook, YouTube etc. alike.

Whether a website, portal or social network is used for social, personal or business purposes, the organisations that own them are always interested on the return on investment they offer. For instance, a company may want to know the number of products purchased through their website in a certain period, or how a particular product lunch performed, or the number of users who visited their site following a campaign etc. To determine any of these pertinent business goals, the website, portal or social network must be monitored, and one tool for doing so perfectly well is by deploying a web analytics tool to monitor the web services.

The common use cases for web analytics are for gathering business intelligence, market analysis and research, business campaign, and monitoring site performance. However, in this paper we investigate how useful web analytics can be for cyber situational awareness purposes. In our work, we explore an intelligence-centric fusion of multiple sources of web analytics intelligence to improve monitoring and detection. Unlike in other application of web analytics [1, 2, 3], our application is for intelligence-centric purposes. The work presented in this paper implements a free version of Google Analytics [4] as a tooling to gathering intelligence around the Centre for Multidisciplinary Research, Innovation and Collaboration (C-MRiC.ORG [5]) conference website; and using the data obtained from Google Analytics to offer insight on a number of intelligence use cases such as operating system fingerprinting, browser fingerprinting, endpoint and geo-location fingerprinting etc. Collectively, these intelligence are fused and analysed, Results obtained from the analysis offer pertinent cues for enhanced situational awareness for better protection of the site.

The rest of this paper is organized as follows: Section I discusses about web analytics, Section II discusses how Cyber situational awareness makes use of cues obtained from web analytics in order to rank, rate and rationalize the importance of specific interactions. Section III explains our intelligence-led detection model, which shows how the different sources are fused, and analysed. In Section IV results are provided and explained, and finally, the work is concluded in Section V.

I. Web Analystics

Web analytics is the tool, technology and method for gathering, analyzing and interpreting or measuring web usage data [3]. There are a number of purposes or use cases for wanting to deploy or use a web analytic tool to monitor an online service. The benefits of web analytics can be seen in its use for monitoring digital services, such as online banking, online portal and websites for the following purposes:

1. Web analytics applications or tools are used to obtain information (profiling) of a digital visitor (customer, user or bot) to a website. This can be used to determine, for instance, the pages viewed, or documents downloaded, speed of page interaction etc. of the interaction or transaction.

2. It can be used to estimate traffic to a website immediately following the launch of a new product/service in order to monitor trends and gauge sales conversions.
3. It can be used to monitor and analyse online marketing campaigns and web contents, and to monitor customer online journey and interaction in order to gather customer intelligence.
4. Today, organisations use web analytics tools to research about their customers (customer intelligence), for instance, what they purchase, and often, knowing how long it took customers to make up their mind in buying a particular product or service.
5. Web analytics can be used to complement Web Fraud Detection tools in order to enforce financial compliance. For example, web analytics can be used in conjunction with web fraud detection tool for Anti Money Laundering (AML) detection, enforcement and compliance.
6. Finally, web analytics can be used for Cyber Security & Intelligence purpose, for example, for geo-location, user profiling, content and export restrictions.

Web analytics can be used to gather very useful and pertinent intelligence regarding a digital visitor, customer journey, interaction or a transaction, including but not limited to:

1. the browser type that was used (Browser fingerprint) for the transaction,
2. the operating system type that was used (Operating system fingerprint),
3. the country from where the attack was launched or originated (Country fingerprint / Geo-Location),
4. the language that was used, that is, the browser and keyboard language settings, which normally reveal the language locale (Language fingerprint),
5. the entity (bot or human) behavioral fingerprint of the entity behind the attack such as frequency, recurrence and recency. Recency being particularly useful in terms of whether the entity has been seen before, and if so, how recent. This also helps with historical trending and analysis.
6. whether it entity visiting the website is a robot (a.k.a. Bot) or a human agent that was used. Cues such as click speed, click velocity and transaction speed can be used to deduce or determine if it is a Bot or human. This is distinct to the use of CAPTCHA to minimize Bot interactions, finally,
7. the type of endpoint that was used, such as Mobile, Tablet, Desktop, Server etc. can be monitored, and intelligence of these gathered and analysed.

II. Using Web Analytics for Intelligence Gathering and Enhanced Cyber Situational Awareness

We propose and present our intelligence-centric fusion web analytics model (See Figure 1). The model is a representation of our fusion architecture for intelligence gathering using web analytics. As we mentioned above, we deployed a free-version of Google Analytics; a sample code of the deployment is shown in Table 1.

In this elementary experimental work, we gathered and analysed the following intelligence, web log, browser, geo-location, device information and endpoint fingerprints, in order to monitor the conference website, and consequently to detect threats targeting the website. We also gathered a variety of other intelligence which have not yet been comprehensively analysed for insights. Hence, our model is extensible to incorporate other data sources, and which can be used for a number of intelligence-led analysis and insights.

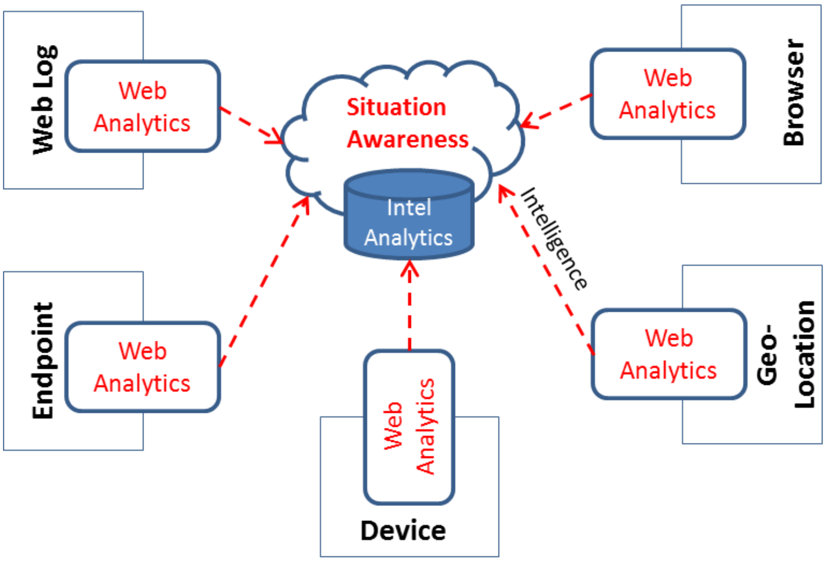


FIGURE 1: INTELLIGENCE-CENTRIC WEB ANALYTICS CYBER SITUATIONAL AWARENESS

In analyzing intelligence gathered, it is pertinent to note that relationships between intelligence can be analysed in order to understand the association or relationship amongst them, which could reveal very subtle but easily dismissed cues which can be fundamentally useful. Relationship can be analysed using a number of techniques such as graph theory, fuzzy logic and any of the machine learning algorithms. Our analysis focused primarily on the built-in analysis technique with Google Analytics technology, and exploring the technology for historical trending, realtime analytics, and further, comparing historic data against realtime data. At this very simplistic nature, we were still able to detect targeted attacks, and other useful information related to particular transactions.

The model shows the interaction, fusion and relationship among the various sources of web analytical collections employed in this study, as follows:

1. Web Log – Logs generated by the website, produced as a result of entities visiting the site which are recorded and stored. Logs contain useful information such as source IP address of the visitor (that is the origin of the transaction), its geography, location and other information obtained during the visit.
2. Browser – Browser information contains records of the type of browser used for the transaction, the browser type, version, and other fingerprints associated with that browser such as language setting, flash version, etc.
3. Geo-Location – Geo-location information contain records of the geography, city, country, continent and including the ISP that the transaction is originated from. With 3rd-party geo-location database augmented with our analytic tooling, we are also able to deduce the street name and map of the originating transaction. This makes pinpointing the attack or transaction at a fine-grain precision straightforward and achievable.
4. Device – Device information contains records of the type of device, such as Server, Desktop, Mobile or Tablet. Other information recorded as the make, model of the Mobile or Table device
5. Endpoint – Endpoint information contains records of the operating systems, browser, screen size, flash version, Java version, keyboard setting/locale and installed apps on the endpoint. These information are used to uniquely identify the endpoint (see Figure 2)

All these provide a rich source of intelligence about the digital visitor, and analysing these pieces of intelligence, together with other sources of data (big data analytics), can only but contribute toward enhanced Cyber situational awareness.

A. *Profiling*

Profiling in relation to this study is the ability to understand the transaction occurring at the online service (website). This is achieved by monitoring transactions occurring at the website using web analytic tooling. Profiling is useful for forensics in order to demonstrate that the entity that carried out the transaction can be determined and that most, if not all, features of the transaction are identified and recorded, and when required played back, or re-produced.

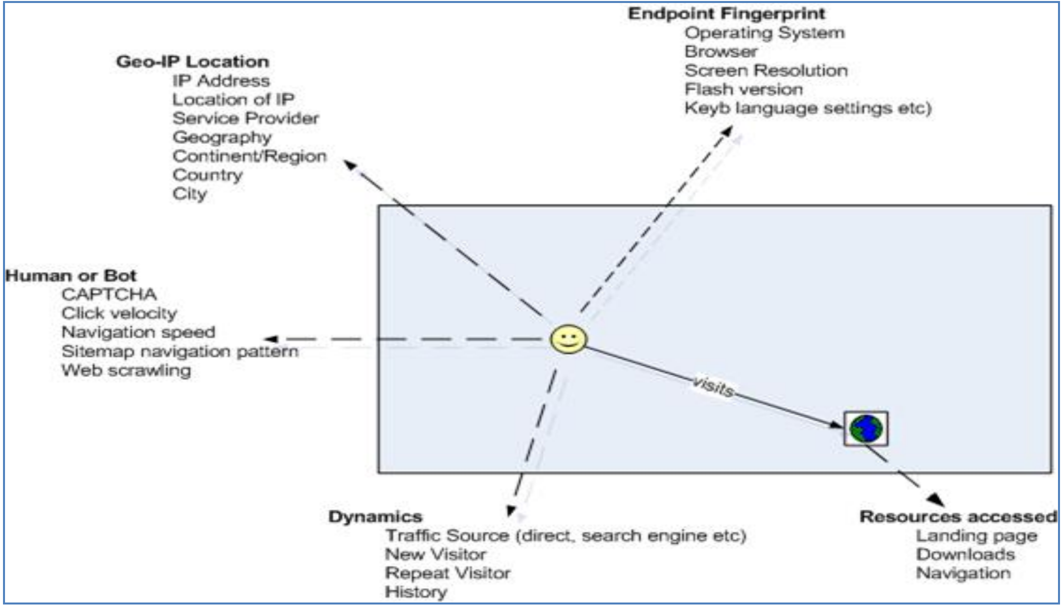


FIGURE 2: TRANSACTION PROFILING

B. *Key Features*

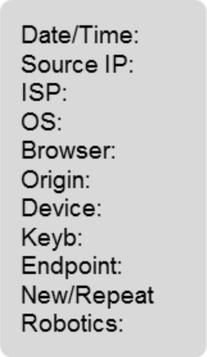
Three key features we investigated are as shown in Figure 2 are:

1. Endpoint fingerprint
   - Operating system fingerprint
   - Browser fingerprint
   - Screen resolution
   - Screen colours
   - Flash version
   - Source IP Address
   - Language setting / locale
   - Keyb setting / locale
   - Robotics (Human or Robot)
   - Device type (Mobile, Tablet, Desktop or Others)
   - Smartphone and Tablet OS,

2. Geo-location fingerprint
   - Service provider fingerprint
   - City fingerprint
   - Country fingerprint
   - Continent fingerprint

3. History fingerprint
   - New visitor
   - Repeat visitor
   - Occurrence
   - Recency

These three key features are carefully combined and collated into an 11-attribute parameter fingerprint referred to as the transaction fingerprint, as shown in Figure 3.



Date/Time:
Source IP:
ISP:
OS:
Browser:
Origin:
Device:
Keyb:
Endpoint:
New/Repeat
Robotics:

FIGURE 3: TRANSACTION FINGERPRINT

The transaction fingerprint uniquely identifies each transaction, by date and time, source IP, service provider, operating system, browser type and version, city, country and continent of origin of attack, device type, make and model, keyb settings and language locale, endpoint type, whether the transaction is new or a repeat visitor, and whether the transaction is current (that is how recent), and finally whether the attributes deduce indicates the transaction to be carried out by human or Bot.

C. *Analysis*

Our basic analysis for this episode of the research is based on transaction fingerprinting comparison. So each transaction is gathered, recorded and compared with historical fingerprints. This is to ascertain whether the transaction entity is one who has previously seen, or whether some key attributes of that entity had changed, in which case, multiple fingerprints are stored against that entity. For example, an entity can have multiple devices – Mobile, Tablet and Desktop, and can complete a transaction from any number of different devices. Similarly, an entity can originate a transaction from multiple locations, provided they are not concurrent and from geo-velocity constrained locations. For instance, if a transaction with identically similar transaction fingerprints is originated from two geographically time-dispersed locations in time frames shorter than that achievable through any transportation means (flights included) then the transaction is marked as untrustworthy, and the riskiness of the transaction is heightened and consequently flagged as suspicion or attacker originated botnets.

D. *Web Analytics Tools*

Web analytics tools for customer intelligence gathering, monitoring online campaigns and monitoring and tracking sales conversions exist, such as: Google Analytics (GA), AWStats [6], Piwik Web Analytics [7], Open Web Analytics (OWA) [8]. Google Analytics offers both free and premium versions of their web analytics tool. AWStats is a free powerful and feature-rich web analytics tool for analysing web logs. Piwik is a free open source web analytics tool available in 48 languages, it provides users insight on their websites' visitors allowing them to run online marketing campaigns, and monitor digital visitors' interactions in order to optimise your campaign strategy and enhance overall customer experience. OWA is an open source web analytics tool that can be used to monitor, track (attribution) and analyse how digital entities use your website and application.

III. IMPLEMENTATION AND RESULTS

First, we generated the Google Analytics Tracking code, a unique 8-digit GA identity (GA-ID) assigned to the website by GA. Then, a Jscript code or PHP code associated to the GA-ID is generated, which must be copied and applied at the backend of the website to be monitored.

Here is a sinppet of the code we applied to C-MRiC.ORG for the implementation.

TABLE 1: SAMPLE GOOGLE ANALYTICS CODE

```
<script>
  (function(i,s,o,g,r,a,m){i['GoogleAnalyticsObject']=r;i[r]=i[r]||function(){
  (i[r].q=i[r].q||[]).push(arguments)},i[r].l=1*new Date();a=s.createElement(o),
  m=s.getElementsByTagName(o)[0];a.async=1;a.src=g;m.parentNode.insertBefore(a,m)
  })(window,document,'script','//www.google-analytics.com/analytics.js','ga');

  ga('create', 'UA-xxxxxxxx-1', 'auto');
  ga('send', 'pageview');
```

```
</script>
```

Where UA-xxxxxxxx-1 is the 8-digit GA code, in this case x represent a binary number

The javascript code is what gathers, records and reports the events as seen on the website. The analysis engine is provided through Google technology, which is presented through a dashboard. The GA portal then analysis, and reports the different attributes of the transaction that is monitored, such as Intelligence Events, Realtime Monitoring, Audience, Acquisition, Behaviour and Conversions [**Error! Bookmark not defined.**]. Reports and dashboards are customisable, allowing you the flexibility to customise your report to tailor your analytic purpose.

A. Results

The GA implementation has been deployed for over 12 months, and since then the site has been monitored, and intelligence gathered analysed. The data shown in these results are those extracted for the rolling 6 months.

Figure 4 shows the number of sessions recorded on a daily basis from October 2015 to March 2016. The high ramp up in sessions was due to the call for papers sent during February and March, and also, March was the busiest being the deadline for paper and abstract submissions, which accounts for the ramp up in online visits to the site.

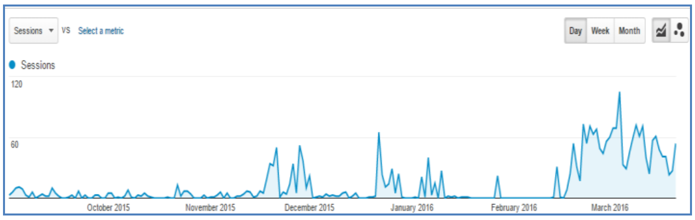


FIGURE 4: SESSION FROM OCTOBER 2015 TO MARCH 2016

Figures 5 to10 show respectively, browser fingerprint, device fingerprint, country fingerprint, city fingerprint, percentage of transaction per country, source of re-direction or referrals channels to the site, service provider fingerprint, and country of origin of transaction or attack.

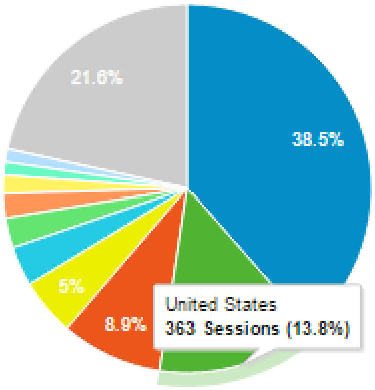
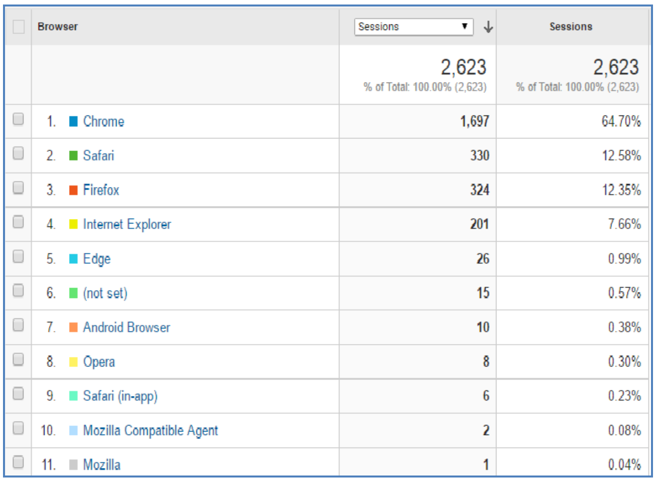
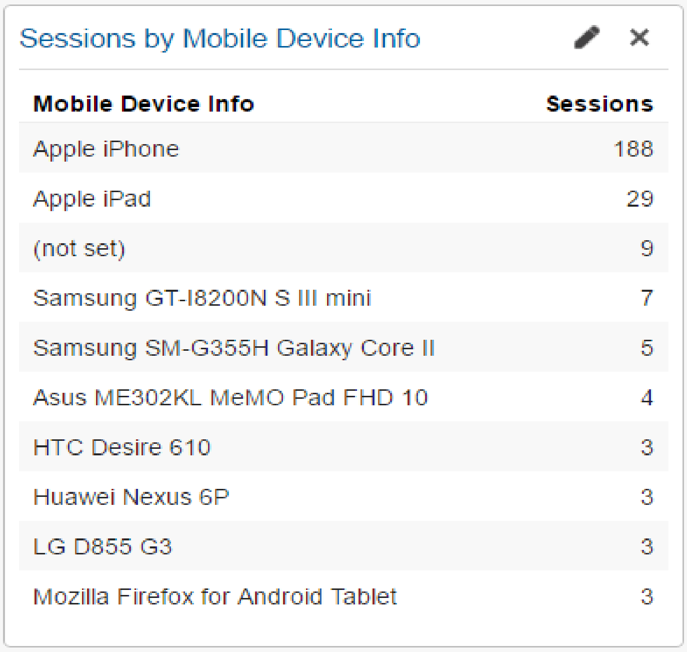


FIGURE 5:ACCESS PER COUNTRY



	Browser	Sessions ▾ ↓	Sessions
		2,623 % of Total: 100.00% (2,623)	2,623 % of Total: 100.00% (2,623)
1.	Chrome	1,697	64.70%
2.	Safari	330	12.58%
3.	Firefox	324	12.35%
4.	Internet Explorer	201	7.66%
5.	Edge	26	0.99%
6.	(not set)	15	0.57%
7.	Android Browser	10	0.38%
8.	Opera	8	0.30%
9.	Safari (in-app)	6	0.23%
10.	Mozilla Compatible Agent	2	0.08%
11.	Mozilla	1	0.04%

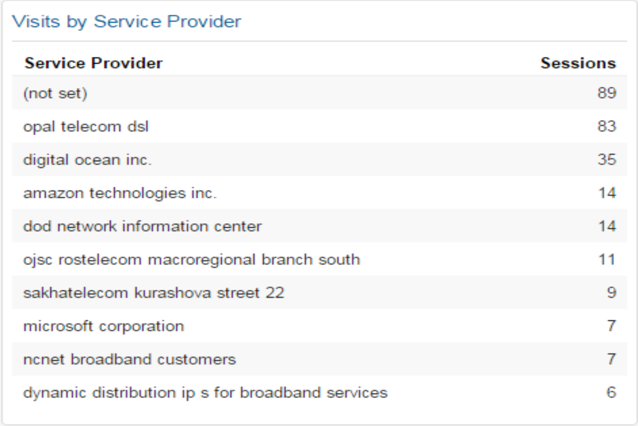
FIGURE 6: BROWSER FINGERPRINT



Sessions by Mobile Device Info

Mobile Device Info	Sessions
Apple iPhone	188
Apple iPad	29
(not set)	9
Samsung GT-I8200N S III mini	7
Samsung SM-G355H Galaxy Core II	5
Asus ME302KL MeMO Pad FHD 10	4
HTC Desire 610	3
Huawei Nexus 6P	3
LG D855 G3	3
Mozilla Firefox for Android Tablet	3

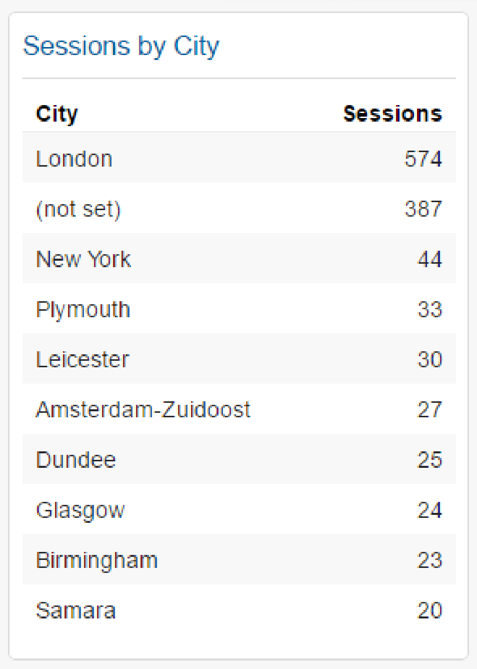
FIGURE 7: MOBILE FINGERPRINT



Visits by Service Provider

Service Provider	Sessions
(not set)	89
opal telecom dsl	83
digital ocean inc.	35
amazon technologies inc.	14
dod network information center	14
ojsc rostelecom macroregional branch south	11
sakhatelecom kurashova street 22	9
microsoft corporation	7
ncnet broadband customers	7
dynamic distribution ip s for broadband services	6

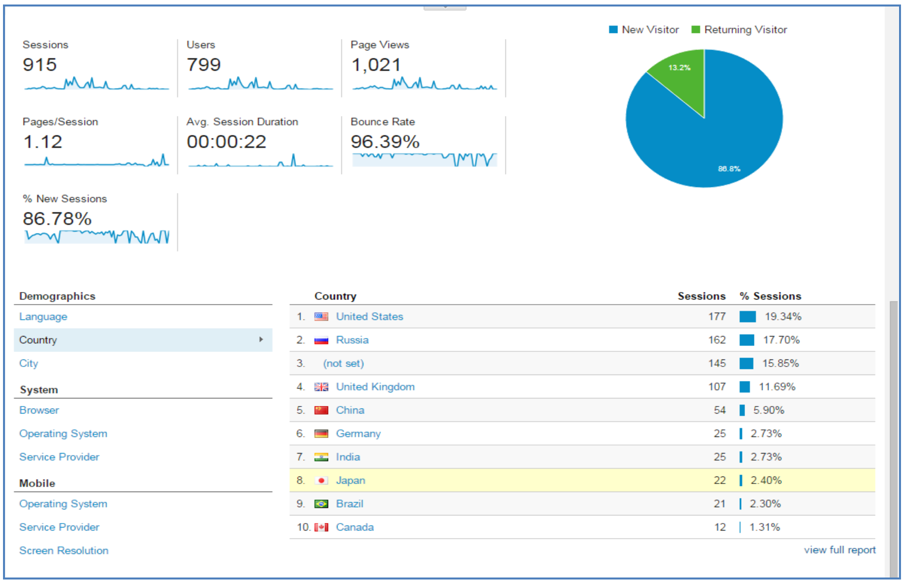
FIGURE 8: SERVICE PROVIDER FINGERPRINT



Sessions by City

City	Sessions
London	574
(not set)	387
New York	44
Plymouth	33
Leicester	30
Amsterdam-Zuidoost	27
Dundee	25
Glasgow	24
Birmingham	23
Samara	20

FIGURE 9: CITY-LEVEL GEO-LOCATION DATA PER SESSION



Sessions 915

Users 799

Page Views 1,021

Pages/Session 1.12

Avg. Session Duration 00:00:22

Bounce Rate 96.39%

% New Sessions 86.78%

	Country	Sessions	% Sessions
1.	United States	177	19.34%
2.	Russia	162	17.70%
3.	(not set)	145	15.85%
4.	United Kingdom	107	11.69%
5.	China	54	5.90%
6.	Germany	25	2.73%
7.	India	25	2.73%
8.	Japan	22	2.40%
9.	Brazil	21	2.30%
10.	Canada	12	1.31%

FIGURE 10: NEW VS. RECURRING VISITORS/USERS

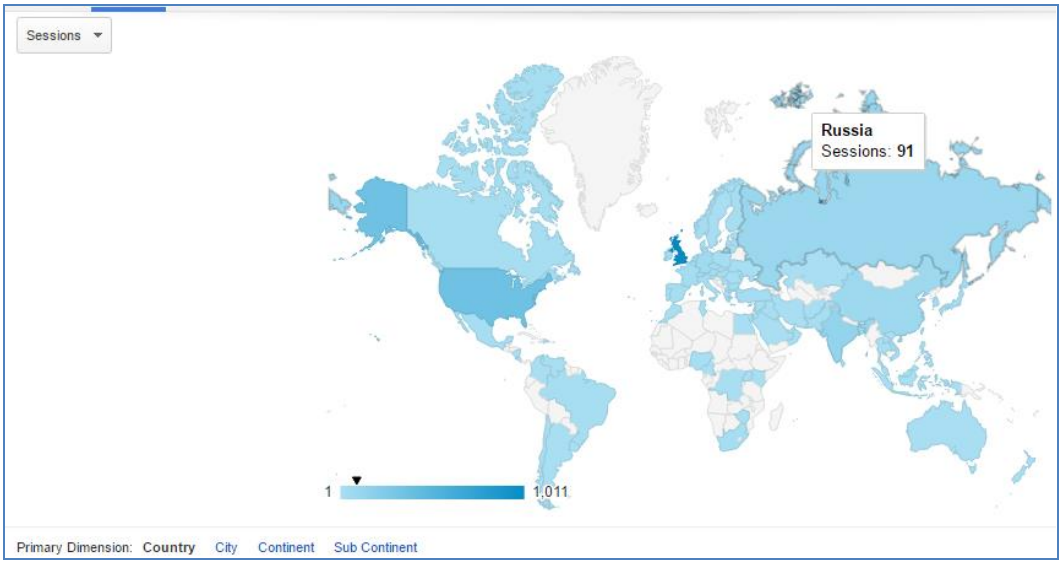


FIGURE 11: GEO-LOCATION BY COUNTRY

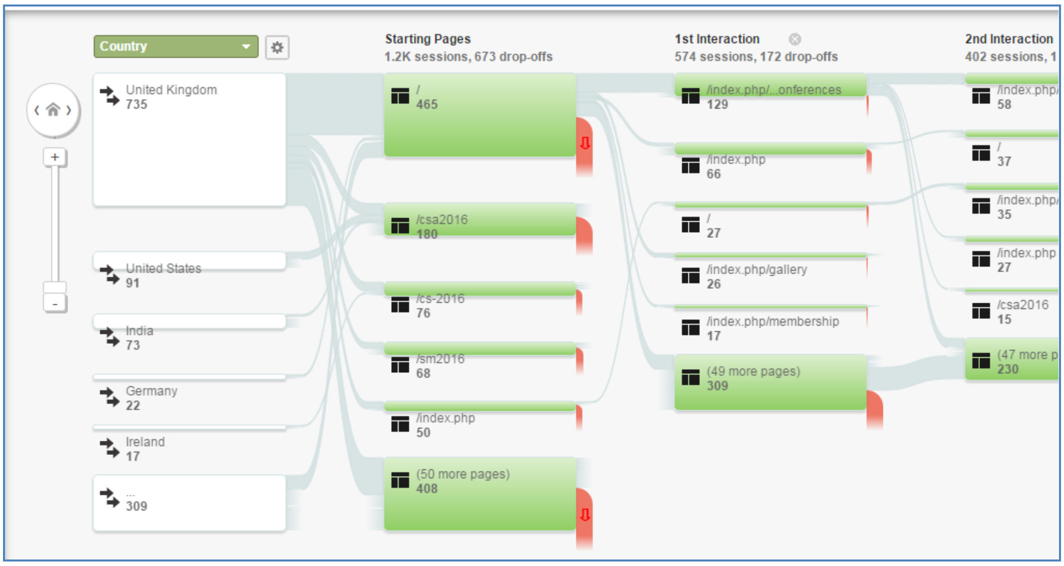


FIGURE 12: FLOW

You would have notified that in Figure 6 through to Figure 9 there is a class of transaction marked "(not set)". These are users who either have deliberately set their device not to reveal their location, browser type, OS and geo-location information. This is parameter is easy to set on mobile and tablet devices, by setting you Wi-Fi or mobile data network not to reveal or use location information. For these class of transaction, generic entry level web analytics may not deduce these set of information, this is because web analytics monitoring is based on passive monitoring. However, to gather fingerprints of devices whose location or network information are not reveal, active monitoring is required.

IV. CONCLUSION

- Web analytics is essential for monitoring digital online services for entity profiling, site performance, marketing, social intelligence and online campaign.
- Web analytics enables better and enhance Cyber situational awareness – this can be through website continuous monitoring and tracking of digital entities, fingerprints of objects of particular interests, which aids identification and detection of anomalous online behaviour, and detection and protection of online public services.
- Interesting & key features of Web analytics relevant to Cyber situational awareness are discussed such as

Endpoint fingerprinting, Location fingerprinting, and, Entity Profiling.

- Web analytics provides intelligence and indicators of compromise for enhanced web fraud detection.
- Finally, a Network-centric intelligence fusion model is discussed, which combines Wetware and Social networks intel in a big data repository to provide richer and enhanced situational awareness, useful when comparing trends, patterns & historical and retrospective analysis.

REFERENCES

[1] Zheng, G. & Peltsverger S. (2015) Web Analytics Overview, In book: Encyclopedia of Information Science and Technology, Third Edition, Publisher: IGI Global, Editors: Mehdi Khosrow-Pour

[2] Increasing Accuracy for Online Business Growth - a web analytics accuracy whitepaper – Accessed (March 2016), from https://brianclifton.com/blog/2008/02/16/accuracy-whitepaper/

[3] J. Burby and A. Brown (2007), Web Analytics Association – Web Analytics Definitions, Version 4.0 (Accessed March 2016) from http://www.digitalanalyticsassociation.org/Files/PDF_standards/WebAnalyticsDefinitionsVol1.pdf

[4] Google Analytics – https://analytics.google.com

[5] Centre for Multidisciplinary Research, Innovation and Collaboration (C-MRiC.ORG) – http://www.c-mric.org

[6] AWStats - http://www.awstats.org/

[7] Piwik Web Analytics, http://piwik.org

[8] Open Web Analytics – Open Source Web Analytics, http://www.openwebanalytics.com/