National College of Ireland

Limiting Identity Data Theft obtained through malicious means in a Digital performing World.

MSc Internship
Cyber Security

Anish Kishan Kanhai
Student ID: 16135695

School of Computing
National College of Ireland

Supervisor: Ben Fletcher

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Anish Kishan Kanhai
Student ID:	16135695
Programme:	MSCTOP-UP CYB **Year:** 1
Module:	Internship
Supervisor:	Ben Fletcher
Submission Due Date:	17 August 2020
Project Title:	Limiting Identity Data Theft obtained through malicious means in a Digital performing World.
Word Count:	5025..... **Page Count**...15.....

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.
I agree to an electronic copy of my thesis being made publicly available on NORMA the National College of Ireland's Institutional Repository for consultation.

Signature: AKKANHAI

Date: 17 August 2020

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Limiting Identity Data Theft obtained through malicious means in a Digital performing World.

Anish Kishan Kanhai
16135695

Abstract

With the use of Digital connectivity many people survived recently, while plenty of businesses & offices temporarily closed their premises globally. It's the Internet or world wide web, that prevented those who work in the IT industry or in another industry, while performing their day to day tasks completely digital, from becoming redundant. The research presented in this document outlines a strategy to limit [with the intend to] even mitigate Data Theft. The ease and accessibility to data which includes company confidential or personal identifiable information has become much more vulnerable, as even employers don't know from either where or with whom their employees are working at times. In this paper we have scrutinized some major data breaches that occurred recently, and propose concepts and strategies to protect (identity) data from being exposed.

1 Introduction

The amount of internet users has rapidly increased over the decades, at present it includes 59.6% of the world population[1], equivalent to 4,648 million people browsing online details are given in Figure 1. Predictions are made that there will be 6 billion internet users by 2022, and more than 7.5 Billion Internet Users by 2030 [2] (approx. 90% of the World Population at that time).

Dec. 2016	3,696 millions	49.5 %
June, 2017	3,885 millions	51.7 %
Dec 2017	4,156 millions	54.4 %
Jun 2018	4,208 millions	55.1 %
Dec 2018	4,313 millions	55.6 %
Mar 2019	4,383 millions	56.8 %
Jun, 2019	4,536 millions	58.8 %
Jun, 2020	4,648 millions	59.6 %

Figure 1: Increase of internet users over the past 4.5 years.

[1] Internet World Stats, https://www.internetworldstats.com/emarketing.htm
[2] Cybersecurity Ventures Herjavec Official Annual Cybercrime Report (2019)

One of the biggest reasons are the IT Giants, that have made the internet more accessible by developing certain tools and interfaces where people can find everything they need. This enormous increase, brings also its side effects as we do not only deal with genuine web surfers, but also those that try to exploit it, gain access to unauthorized information or even steal an individual's personal (identifiable) information.

In this paper we are going to identify a strategy to limit Data Theft, as employees are working remotely, it's not feasible to limit the data they need to access and still expect a high productivity, majority of companies have or had teams that could only work on-site those who are responsible for a data center or site reliability teams. Despite that we connect to any network and with the use of a virtual private network we can get corporate access, it's of crucial importance that malicious networks are being detected, for which we have Intrusion Detection & Intrusion Prevention measures in place, but are those reliable and fit for a global Work from Home office environment. We have become dependent on the Internet, and are sometimes forced to connect to public network either in a library, cafeteria or at airports, with exposure to all risks that may occur. Our research also focusses on users whom participated in our survey, to investigate human behavior who are frequently associated with the weakest link.

Data that is been uploaded and shared contains many additional (including geolocation, date- & timestamp, IP location) information, that malicious users would like to capture for illicit purposes.

The structure of this report is as follows:
Section 1. Introduction the mindset behind a secure digital infrastructure, focusing on securing data, to mitigate data theft and leakage.
Section 2. Literature Review, an extensive research through previously published academic journals, papers and articles related to data theft & data leakage.
Section 3. Research Methodology, outlining our research strategy including study area, data sources and analysis.
Section 4. Design Specification, our proposed concept with justifications based on the knowledge concluded from the Literature Review & Research Methodology.
Section 5. Implementation, a justification of how and why decisions were made, including a partial technical implementation of the proposed solution.
Section 6. Evaluation, of our proposed/developed ICT solution, how it benefits.
Section 7. Conclusion & Further Work, discussion on & concluding our research including, with the possibilities for further works related to this research context.

2 Related Work (Literature Review)

In order to protect data, we need to understand where it is being located, in the world of digitization we make us of Information Systems, which is designed to collect, process, store, and distribute data or information [1]. It's critical that these Information Systems are well-secured, to minimize data exposure, for this research we investigated recent academic papers, journals and publications to understand that despite technology is advancing and individuals are aware of the risks, still one attack after the other takes place.
In this overview of related documentation that we have gone through, our aim is to understand the current findings including solutions, we will determine if those are still relevant as why further research is mandatory.

Problem Statement
Data theft and leakage remains a big issue, despite some much awareness, and advanced measures in place. In our research we will investigate and outline the security measures, related to stored data and human interaction with technology. The following hypothesis has been established.
H^0 Online Data theft and leakage = Preventable by implementing security features
H^A Online Data theft and leakage ≠ Preventable by implementing security features
We will be focusing on multiple angles and perspectives, within the cyber security domain.

2.1 Recent Data Compromise Methods

According to the ISTR one out of 10 URL's are malicious [3]. The most prevalent types of cybercrimes being reported last year to the FBI's Internet Crime Complain Centre [4] were Phishing, Vishing, Smishing & Pharming, Non-Payment/Non-Delivery, Extortion, and Personal Data Breach. Among these the top three with the highest losses included Business Email Compromise, Confidence/Romance Fraud, and Spoofing (see Appendix 1. glossary). Human interaction exploitation is the major culprit for attackers to be successful, with the use of social engineering fraudulent attempts with the use of technical mediums are being initiated to steal data[5]. The curiosity and appearance from a trusted source lead people to click, download, install, open malicious content, with the consequence of losing data or money. Humans are unknowingly being engaged in malicious activities, from all attacks less than 1% makes use of System Vulnerabilities [6].

In a recent incident that occurred in April '20, The Anti-Phishing Working Groups eCrime eXchange team received reports of 1,054 attacks against Zoom, whereby phishers emailed out fake Zoom video-conferencing meeting notifications, directing victims to Webpages specially designed to steal their Zoom login credentials, these where used to log in to their corporate video conferencing accounts. The harvested passwords where tried on other sites and services where the victims may have registered themselves, another attack type offered Internet users the opportunity to download the Zoom client, but instead, delivered malware files unto their device(s) [2]. Almost one in ten targeted attack groups now use malware to destroy and disrupt business operations, a 25 percent increase from the previous year [7].

One of the main causes why some cyber-attacks are very successful and at times even easily performed by attackers is that there is too much focus on ICT controls and too little on office policies, procedures, regulations, monitoring and due diligence [3]. This has been confirmed in a research paper published by the Rotterdam University of applied sciences, which after having read forced us to consider to perform our research also from multiple angles and perspectives. To obtain privacy-protecting and secure Information Systems, the gap between non-technical (ethical, legal, social and economic) aspects and technological ones needs to be bridged. In this way we have Privacy to protect people and their (personal) data, and we have security to protect systems, equipment and assets [4]. The opportunity to commit fraud is frequently associated with weak governance, poor internal controls and accountability [5]. There is still a disparity about who is responsible for Data [6]. The already since two year effective General Data Protection Regulation, doesn't mention the word Privacy in its

[3] Internet Security Threat Report (ISTR) Volume 24, February 2019
[4] 2019 Internet Crime Report, Federal Bureau of Investigation, Internet Crime Complaint Center
[5] APWG Phishing activity trends report: 4rd quarter 2019 Anti-Phishing Working Group
[6] Proofpoint HUMAN FACTOR REPORT 2019
[7] ISTR 24: Symantec's Annual Threat Report Reveals More Ambitious and Destructive Attacks

legislation, even the term data privacy can't be found in it [8]. The GDPR rather enforces the Protection of Data and how it should be handled This framework is useful to limit data exposure, reduce scammers and provide more transparency if a data breach occurs by reporting this to the authorities within a specific timeframe.

Another synonym for Data Theft is Data leakage whereby confidential information (which consists of intellectual property, monetary data, PII data like contact information, credit-card details, & other confidential information [7] is being transferred to untrusted environments or compromised. Causes for this are as follows[8];

- Intentional data leakage by an adversary internal to the organization.

A major incident occurred in 2018 "the Apple data theft case", whereby filed a criminal complaint was filed charging a former Apple employee with the crime of stealing trade secrets regarding their autonomous vehicle project[9].

- Data leakage by a person external to the organization but with (temporary) access rights to the victim organization's resources

According to the data of the National Computer Information Security Evaluation centre of China, in the incidents that caused heavy losses due to internal important secrets leaked through the network, only 1% were stolen by hackers, while 99% were caused by intentional or unintentional leaks from internal employees [7].

- Unintentional leakage by Internal users or administrators.

In may 2019, due a faulty database script in Pardot (a B2B marketing automation tool) users where able to see all company data, regardless their access permissions [10].

2.2 Technical Constraint, how Data become vulnerable

Data leaks are harmful to both enterprises, and people, in Aug. 2018, Catawba Valley Medical Centre informed 20,000 patients about their personal data being breached [10]. This is very dangerous especially if we consider (the amount of) IoT devices in the healthcare industry, as it transforms the way we live and work, a key enabler in the healthcare industry. With the use of IoT sensors connected to (wireless) networks and remotely controlled with smart devices in order to monitor patients in hospitals or at their homes [11] this can end up in a disaster. We haven't touched the cloud computing industry yet, it's the go-to platform for data storage, data processing, data analytics, and data backup and recovery, due to its immense benefits such as being highly available, massively scalable, hugely cost-saving, and quickly deployable. The cloud presence is becoming a norm in almost 70% of the enterprises across the globe, and have at least one application running on the cloud" [12].

The biggest problem in the cloud is the cloud itself, with very little effort, resources are being toggled on or off as needed. It's easy for businesses to scale up or down their computing power, but all this flexibility & ease with which operations can provision or reconfigure infrastructure, can have a security impact as one act of negligence can lead to a full database exposure [11]. According to Cisco's white paper, cloud data centre traffic will represent 95% of total data centre traffic by 2021. Eventually cloud computing will wipe out data centres altogether [12]. This will have a massive impact on the amount of Data that is being generated. Almost all devices, services and platforms provide apps, that can be installed for a better user

[8] https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:02016R0679-20160504&from=EN
[9] Former Apple Employee Charged with Data Theft
[10] Massive Salesforce Outage Resolved With Gradual Access Restoration
[11] Sophos 2020 Threat Report, https://www.sophos.com/en-us/medialibrary/PDFs/technical-papers/sophoslabs-uncut-2020-threat-report.pdf
[12] Cisco Annual Internet Report (2018–2023) White Paper

experience, be it a digital camera, a Smartwatch or internet banking. These mobile (software) applications either connected to a cloud environment or not use log files with log messages, that are printed during the execution of the program at its run time. These log files are often the only data source available for developers/engineers to diagnose program failures to gain better insights in determining the root cause of problems that occur, like security attacks, hardware failures, and configuration issues [13].

It has been very unfortunate that with the use of the cloud as a mobile app backend or mobile backend as a service (mBaas), we witnessed massive data leaks from the cloud. Reports state that insecure backend databases of mobile apps exposed approximate 280 million sensitive user records including personally identifiable information (PII) such as user-names, passwords, emails, phone numbers, and location details [14]. Appendix 2 provides a list of occurred data breaches.

A reason for Cloud vulnerabilities to occur may be due to the fact that cloud providers including Amazon, Backendless, Google, Kinvey, Microsoft, Oracle, etc. provide typically four components with mBaas: Database Management, User Management, Storage, and Notification Delivery with corresponding API's such as setValue, removeValue, createUserWithEmailAndPassword, sendPasswordResetEmail, putFile, listFilesAndDirectories, CreatePlatformEndpoint and publish [12]. More and more mobile apps are using these cloud APIs for various services like authentication, authorization, and storage, without directly setting up and managing those proper backend infrastructures. By inspecting the usage of these cloud APIs, it can be understood how each mobile app manages its customers data. Logs files may be analyses for this, there are different kinds of logging practices out of which verbose logging is recording more information than usual, especially for troubleshooting purposes. Despite all the warnings, poor logging is still a common practice in software development[15]. Which leads potentially to a data leak, as mentioned earlier this is more severe on mobile applications or IoT devices, since these carry sensitive identification information ranging from physical device identifiers (e.g., IMEI MAC address) to communications network identifiers (e.g., SIM, IP, Bluetooth ID), and application-specific identifiers related to the location and the users' accounts, which can easily be retrieved with connected API's or calling getDeviceID [16]. Appendix 3 provides full details of retrievable data which can be exposed.

3 Research Methodology

Our research is based on limiting data theft, to meet our research objectives, we have included a theoretical and practical part. In the first we analyse datasets to gain a better understanding on Data Theft vulnerabilities by making use of qualitative and quantitative methods, including primary and a obtained secondary data sources. In the practical part, we propose a solution in the hope to mitigate security risks. Our empirical research is an amalgamation of both.

3.1 Study Area

None of the scientific and informative research papers that we analysed during our literature review could provide a specific solution. The Security Effectiveness Report 2020, concluded in their findings: "*Security controls are not performing as expected"* and mentioned it's alarming that **alerts are only generated for 9% of attacks [13]**. Luo et. all, proposed a

[13] Deep Dive Into Cyber Reality, Security Effectiveness Report 2020, Mandiant

framework called PrivacyProtector, focussing on protected data collection with the ideas of secret sharing and share repairing (in case of data loss or compromise) for patients' data privacy, which they are till date still unable to implement successfully [11]. Ghouse discussed the following technique for data leakage prevention (DLP), Hash-based fingerprinting, due to which sensitive data won't be identified, they refuted their own findings as with Hashing techniques (MD5 or SHA) any minor modification within the data will automatically change the complete hash value, as hash values of the smaller datafiles are also susceptible to change even a minor change within the data, also considering the processing required for this it is inappropriate to use [8]. LeakScope's solution for data vulnerabilities in the cloud, is clearly not perfect and has many limitations, as it gives false negatives. A reason given is that detection of vulnerabilities is based on APIs, if there are other APIs that also include developer credentials in their parameters, LeakScope will miss the identification of these strings [12]. At present we are in a stage to examine several possibilities to prevent unauthorized data loss, our humble work in this contribution will be one of those attempts. Nothing is guaranteed except change[17], with the scientific approach that we take, we hope that our change will be one for good.

3.2 Data Sources

Our primary qualitative dataset that we generated through a questionnaire being sent in survey format. We aimed for 1000 Participant globally we got a response of 850 or 85%. The data has been collected between May and June 2020, aimed at active online digital platforms users located worldwide. In our analysis we included the whole population, only to validate the value of the standard deviation we had to use random sampling, with 40 respondents from each month to compare. In our questionnaire we formulated the questions is such a way that a better insight will be given how human beings interact with technology including their personal data. We have to consider that due to geographic differences, some discrepancies may appear in the data, for example citizens in the west have to comply with different laws & regulations, while in other continent users may be either more active on social media or have no access to public Wi-Fi networks, only by authenticating themselves with a OTP (by phone or email). This questionnaire is being conducted, in order for us to determine the digital behaviour better.

Our secondary publicly available (obtained) quantitative dataset, The UNSW-NB15 Source Files [18][19]. This dataset includes nine of the following attack types, Fuzzers, Analysis, Backdoors, DoS, Exploits, Generic, Reconnaissance, Shellcode and Worms. Our research is focussed on browser based attack types, and we segmented the data on the following protocols with their respective ports, SMTP (25), HTTP (80), FTP (21), DNS (53) & UDP/TCP (1043). We investigated the frequency of attacks, the interval ratio or duration and success rate. The outcome of both analysis in comparison to the earlier performed literature review will altogether assist us to conceptualize with the intend to develop our ICT Solution to limit data theft.

3.3 Research Analysis

First and Foremost, we oriented the gathered and obtained datasets [named Survey & Attacks] after converting it to .csv files, we segmented only the columns relevant to our research and analysed it in R and Tableau, Figure 4 provides the Data Analysis overview as performed by us.

```
> str(Attack)
89581 obs. of  41 variables>
```

```
str(Survey)
850 obs. of  10 variables
```

```
> nrow(Attack)
[1] 89581
> ncol(Attack)
[1] 41
```

```
> nrow(Survey)
[1] 850
> ncol(Survey)
[1] 10
```

The Survey dataset consisted the following questions to provide us a better insight on human behaviour with digital mediums, devices and platforms.

1. Gender
2. Age Group
3. Social Media Platforms used
4. Device / Interface mostly used
5. Auto-Saved passwords
6. Public Wi-Fi usage & accessibility
7. Use of a Password Manager
8. Has respondent's account ever been compromised
9. Using one's Real Identity online
10. Most occurring Cybercrime

```
> names(response)
 [1] "Gender"          "AgeGroup"        "MediaPlatform"   "Device"
 [5] "SavePassword"    "PublicWifi"      "PasswordManager" "Compromised"
 [9] "RealID"          "MFC"
```

The Survey Dataset consisted mostly of nominal data, due to which we couldn't perform many statistical test's on it, after converting some variables into numerical or ordinal data, we determined the following as show in Figure 2.

```
> var(Survey)
```

	Gender	Device	SavePassword	PublicWifi	PwManager	RealID
Gender	0.58970554	-0.06937574	0.06610130	0.189069494	-0.02548881	0.029729093
Device	-0.06937574	1.01118963	0.01201413	0.072084806	-0.12414605	-0.196230860
SavePassword	0.06610130	0.01201413	0.23067138	0.118539458	0.01121319	0.181012956
PublicWifi	0.18906949	0.07208481	0.11853946	1.167773852	-0.06487633	-0.003203769
PwManager	-0.02548881	-0.12414605	0.01121319	-0.064876325	0.35401649	0.102520612
RealID	0.02972909	-0.19623086	0.18101296	-0.003203769	0.10252061	2.884994111

[Figure 2. Output of the Variance function on the Survey Dataset in R]

The Attack Data set contained the following information this has been filtered on the rows and columns that are relevant to this research. Appendix 4 includes a list of the Table headers with the full description of the data as it uses abbreviations.

We created several charts and graphs to identify outliers, compared to the full data as it may be an incidental occurrence or due to exceptional circumstances some attacks lasted longer with a higher success rate We have included the most significant ones in this papers, each under its relevant section.

Our obtained Attack dataset could be plotted nicely in R, figure 5 gives its output based on which we could identity if variables are correlating, and on which our research should be focussed on to gain a better insight on our proposed hypothesis.

4 Design Specification

After having completed the analysis, following outcomes became clear amongst all responders most users access internet applications through their browsers, 527 people are using a web browser to access a website or application [Figure 6], on a desktop/laptop it's more common to use either Chrome, Safari, Firefox and/or Edge (IE), while on mobile devices users usually install apps that are available, but due to storage limitations or annoying notifications individuals chose to use a web browser.

Web browsers allow multiple Internet Protocols and ports to be open, for a good and proper connection but causes also risks to be targeted in cyber-attacks.

According to the Attack Dataset, the most vulnerabilities took, while victims were connected to the internet which is almost impossible without a Domain Name System (DNS) which

makes browsing human readable, when we enter website URL's it's being automatically redirected to the proper IP Address of the webserver. Figure 7 shows the protocols users where connected to when being attacked and exploited.

If we take a look at the ports Figure 8 indicates that all port ranges have targeted with successful outcomes for the attacker, the most frequently used are between port 0 to port 1000.

Port 0 is reserved & blocked by ISP's [14] while the highest or biggest Port 65535 is an Ephemeral Port [15] which from 49152 all the way through 65535, all these are reachable through a network connected to the Internet. At times people access an unsecure website, the website page won't be loaded and a message appears as follows, **this webpage is not secure,** despite this warning users still have the possibility to click on continue to access the content. For this a General Online Presence Regulation must be in place, that each online service or hosted website should adhere to. Our proposed ICT Security Solution based on the performed investigations and analysis, is a browser extension which can be installed by users that will be directed to either secure websites and/or no (personal) data will be either captured or stored, even not during an Incognito or private Window browsing session. Our developed browser based (in this scenario) Chrome Extension has been named: Surf-Sue, which stands for Surf **Sec**u**re,** with which users can browse more safely and protect their identity online by not exposing sensitive information. Figure 8 displays the business process notation model, full details have been provided/included in our Configuration Manual.

5 Implementation

The concept of Surf-Sue is very well appreciated, at present we are in stage of launching a beta version, with all those that participated in the questionnaire. As it provides trust and alertness, users are being alerted whenever their data is being requested to be filled in or processed. Figure 9 shows the process including the attacks from which one is being protected.

We have come to this decision by comparing the outcomes of the following questionnaires:

[14] https://www.speedguide.net/port.php?port=0
[15] https://www.ncftp.com/ncftpd/doc/misc/ephemeral_ports.html

```
> my_table_2 <- table(MSc_Survey_Responses_$`8. Has your account
 ever been compromised (if unknown check https://haveibeenpwned.
com/)`, MSc_Survey_Responses_$`5. Do you auto-save passwords`)
> print.table(my_table_2)

                          No Yes
  1 - 2 times             17  85
  1 time                  68 204
  2 - 3 times             34  85
  More than 3 times       17  51
  No                     102 102
  Prefer not to Disclose  68  17
```

```
> my_table_2 <- table(MSc_Survey_Responses_$`5. Do you auto-save passwords`, MSc_Su
rvey_Responses_$`8. Has your account ever been compromised (if unknown check http
s://haveibeenpwned.com/)`)
> print.table(my_table_2)

    1 - 2 times 1 time 2 - 3 times More than 3 times  No Prefer not to Disclose
No           17     68          34                17 102                     68
Yes          85    204          85                51 102                     17
```

[Figure 10. Comparison of user accounts being compromised based on if they auto-save their passwords]

```
> my_table_1 <- table(MSc_Survey_Responses_$`6. Do you connect to "Public"
 Wifi Networks ?`, MSc_Survey_Responses_$`5. Do you auto-save passwords`)
> print.table(my_table_1)

                                                    No Yes
  No, never                                        102  68
  Yes, always (to save mobile data)                 51 170
  Yes, Only to trusted networks                     34  68
  Yes, when I have no other Internet connectivity option 119 238
```

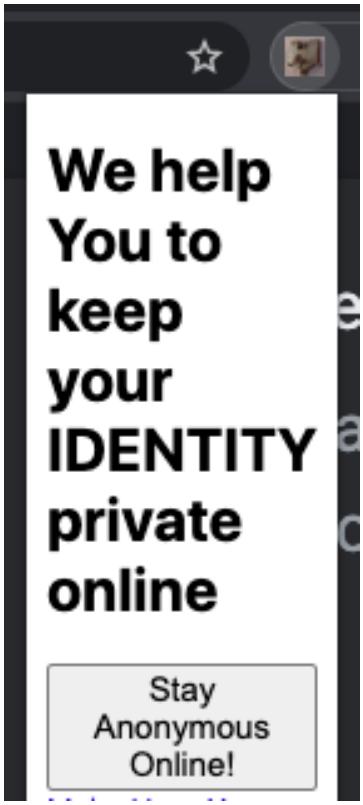
[Figure 11. Comparison of user accounts being compromised based on if they connect to public Wi-Fi network, to gain internet access]

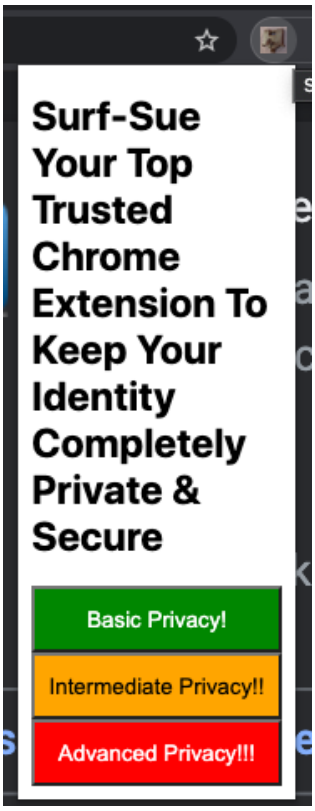
```
> my_table_8 <- table(MSc_Survey_Responses_$`7. Are you using a Password
 Manager (e.g. LastPass, Dashlane etc)`, MSc_Survey_Responses_$`5. Do yo
u auto-save passwords`)
> print.table(my_table_8)

                          No Yes
  No                     170 374
  What is a Password Manager  85 102
  Yes                     51  68
```

[Figure 12. Amount of users that make use of Password Managers, compared to if they auto-save their passwords]

The main functionality behind Surf-Sue is after enabling the Extension, users will need to choose the level of privacy they desire. On some platforms no private information should be

disclosed at all for which we have the advanced privacy while some website's that require or have authentication measures in place, needs to verify data.





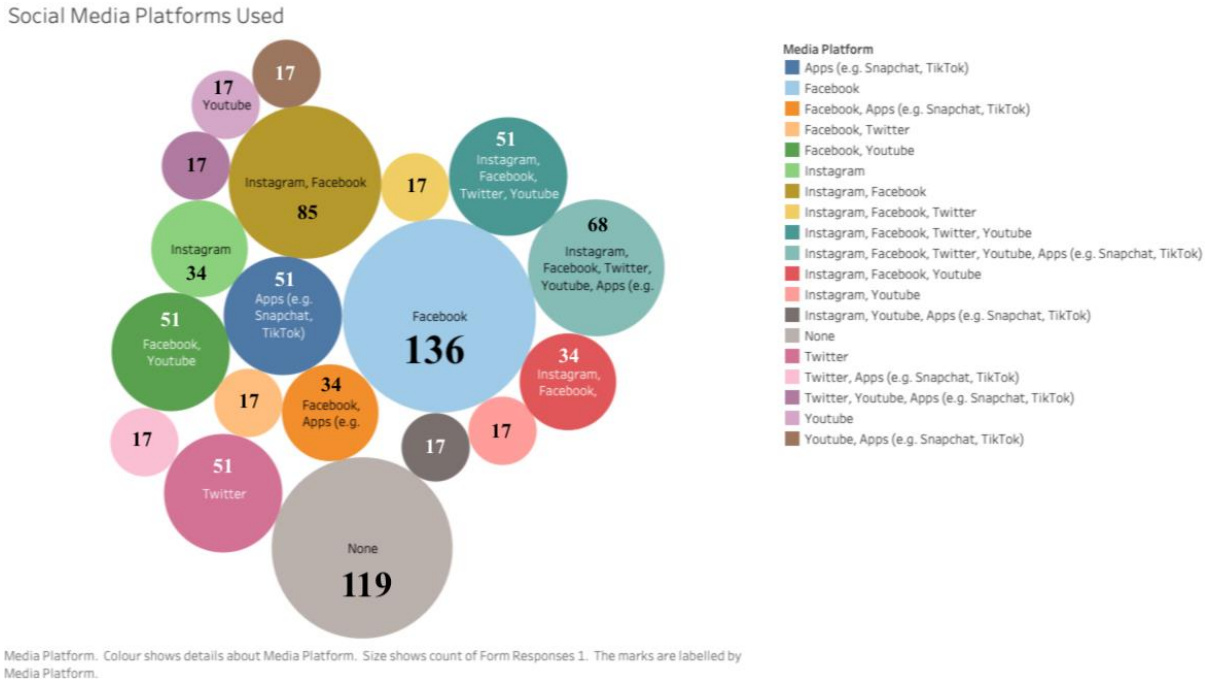
Another reason why we included multiple Privacy levels, is that on some platforms users are required to use their real details. In our Survey we noticed that majority of the responders use their Real Details when registering or subscribing to a certain service online see figure 13. For some it depends on the platform on which they are registering, we keep in mind that on some platforms it's a must like when you apply online for a study, but in other instances this results in that individuals are willing to disclose their own identities online.

6 Evaluation / Discussion

During our ongoing research we had to understand the fact that on Global level humans had to adapt a different working style, by being more online present than ever. This has been reflected in our Survey questionnaire, which was filled by the following audience as can been seen in Figure 14 & 15.

6.1 Analysis of Social Presence

Social Media usage is increasing, as it remains one of the quickest ways to connect with people and the outer world, this is reflected in figure 16. Please note that some individuals are using more than one platform.



[figure 16. Social Media Platforms being used]

Approx. 120 people responded none which is ±15% of our responders, this may be due to the fact of the exit movement, age or time constraints, as social media platforms may consume plenty of time to interact with others.

We also checked the following behaviours, are the Social Media users, using their real identity's online [figure 17] and which devices are mostly being used for social media purposes [figure 18] this was the response.

6.2 Analysis on Cyber Attacks

We got the following insight or knowledge outcomes, after having perform a scrutinized analysis on the data. Among the occurred attacks, majority where done Generic [figure 19] whereby no specific knowledge is needed of a certain process or functionality, for examples the attacker doesn't need to know the encryption algorithm that's being used.

Figure 20 shows us a plot chart with the duration of each attack, whereby we clearly see that no attack lasted longer than 60 Seconds, please note that we only included browser attacks that occurred over the internet.

It's hereby also important to get an understanding of the amount of packets that are being transferred, figure 21 provides a plotted graph with the Mean values of the packet size transmitted by the source to and from destination.

If we look at the vulnerability rate, the User Datagram Protocol (UDP), which provides low-latency and loss-tolerating connection[16] is causes the most security issues, this is followed by the Transmission Control Protocol (TCP) that is connection orientated between the sender and receiver prior to data transmission [17]. As [Fig. 22 indicates, after UDP & TCP the most exploitable protocols are Universal Network Architecture Services (UNAS), Open Shortest Path First (OSPF) and Stream Control Transmission Protocol (STCP).

As we mentioned the Term latency above, we also need to look at the jitter. To be more precise latency is the travel time for a packet to reach its destination, Jitter is the fluctuation in this latency, one packet can be transmitted in 45msec while another may take 250msec [18]. Figure 30 shows us information with the Source Jitter which is much higher compared to the Destination Jitter. Reason for this delay are given as follows: network congestion, improper queuing, or due to configuration errors, the steady stream becomes lumpy, or the intervals between each packet varies instead of remaining constant [19].

6.3 Analysis of Attack Assumption v/s Compromise Reality

In order for us to get a better insight on which attack type is considered the most frequent among our participants. We see [figure 24] an equal response between Hacking & Phishing and Identity Fraud & Scamming.

We include the response on our questionnaire to display the reality how our responders interact with technology and willingly or unwillingly make themselves including their data vulnerable to malicious attackers. Figures 25, 26, 27 and 28 visualize their outcomes.

The Table in Figure 29 provides a comparison of respondents that connect to public Wi-Fi networks and have ever one of their accounts compromised.

Interesting to see is that 560 responders acknowledged that their account has been compromised at least once. At that the majority compared to other variables are willingly connecting to public networks, either to save their Data usage or they have other internet connectivity option.

6.4 Statistical Tests

Below we performed a Two-Sample t-Test to compare the mean values, in Fig. 35 we took the mean values of Synack (the time between the SYN and SYN_ACK of packets) and Ackdat (the time between the SYN_ACK and ACK of packets. Both belong the TCP/IP protocol and happen according the 3-way handshake process.

The hypothesis is set as follows,
H^0 Duration between SYN and SYNACK = duration between SYN_ACK and ACK
H^A Duration between SYN and SYNACK ≠ duration between SYN_ACK and ACK

[16] UDP (User Datagram Protocol)
[17] https://www.privateinternetaccess.com/blog/tcp-vs-udp-understanding-the-difference/
[18] https://haste.net/2017/08/23/what-is-jitter/
[19] https://www.cisco.com/c/en/us/support/docs/voice/voice-quality/18902-jitter-packet-voice.html

```
> t.test(Attack$synack, Attack$ackdat)

        Welch Two Sample t-test

data:  Attack$synack and Attack$ackdat
t = -1.695, df = 177644, p-value = 0.09008
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -4.761642e-04  3.451989e-05
sample estimates:
 mean of x  mean of y
0.01001128 0.01023210
```

[Figure 30. Two Sample t.test of synack & ackdat mean values]

Based on the outcome we fail to reject the null hypothesis.

Next in fig. 31 we perform the same on the mean values of the Source & Destination jitter values, as an discrepancy was already visibly clear in section 6.2 of this paper, this test was performed for confirmation purposes.

```
> t.test(Attack$Djitter, Attack$Sjitter)

        Welch Two Sample t-test

data:  Attack$Djitter and Attack$Sjitter
t = -19.674, df = 89576, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1869.332 -1530.619
sample estimates:
mean of x mean of y
 201.8373 1901.8125
```

The final performed Two Sampled t-test [figure 32], includes the mean values of the transmitted packet size, by the source and destination.

Our Hypothesis for this is set as follows:
H^0 Transmitted packet size by source = transmitted packet size by destination
H^A Transmitted packet size by ≠ transmitted packet size by destination

```
> t.test(Attack$smean, Attack$dmeans)

        Welch Two Sample t-test

data:  Attack$smean and Attack$dmeans
t = 91.366, df = 177610, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 61.43333 64.12685
sample estimates:
mean of x mean of y
 94.05957  31.27948
```

[Figure 32. Two Sample T.Test mean values of packet sizes.]

Seeing above outcome, we reject the null hypothesis in favour of the alternative hypotheses.

7 Conclusion and Future Work

In this research we started with investigating user online behaviour, and analysed Data of occurred Cyber Attack. This resulted in the development of a Chrome extension named Surf-Sue to keep one's identity private.

Despite all steps being taken till date. It's very easy to request data from people, just by posing a fake or deceptive vacancy online, and requesting applicant to send their CV/Resumes all their major personal information is obtained. As discussed earlier in the paper, a General Online Presence Regulation should be established with how organization can represent themselves online, which documents can be requested through a digital medium and if needed this can only be done by those who are licensed or authorised to do. As majority of data incidents are connected to an online website, service or subscription, for now consumers have very limited options and even if they delete their accounts, data may still be stored in a back-end server [20]. In order for us to answer our research question, is Online Data theft and leakage Preventable by implementing security features, the answer is party yes, but with the proper privacy regulations applied to it as well, which are as for now not yet in place, at least not for the individual consumers. Our further work is to focus on this, and implement it unto our developed application whereby users will be notified if they haven't used a specific online platform or website for certain amount of time, to delete their details to keep their online information limited, also as a part of this work in our pipeline is to focus on Biometric authentication as that may prevent unauthorized access completely.

[20] https://www.pcworld.com/article/3391916/mystery-data-breach-reportedly-exposes-80-million-names-addresses-and-income-info.html

References

[1] Piccoli, Gabriele; Pigni, Federico (July 2018). Information systems for managers: with cases (Edition 4.0 ed.). Prospect Press. p. 28. ISBN 978-1-943153-50-3. Retrieved 20 May 2020.

[2] [Anti-Phishing Working Group (APWG), Phishing Activity Trends Report 1st Quarter 2020, published 11 may '20, https://docs.apwg.org//reports/apwg_trends_report_q1_2020.pdf] Accessed 22 May 2020.

[3] Amirudin, N.R., Nawawi, A. and Salin, A.S.A.P. (2017), "Risk management practices in tourism industry – a case study of resort management", Management and Accounting Review, Vol. 16 No. 1, pp. 55-74]. Retrieved 21 May 2020

[4] Mortaza Shoae Bargh Realising Secure and Privacy-Protecting information Systems: Bridging the Gaps. ISBN: 9789493012080 1st edition, 2019 Publication by Hogeschool Rotterdam Uitgeverij Retrieved 5 June 2020

[5] Nawawi, A. and Salin, A.S.A.P. (2018), "Employee fraud and misconduct: empirical evidence from a telecommunication company", Information and Computer Security, Vol. 26 No. 1, pp. 129-144.

[6] Wallis, Jillian C., Borgman, Christine L. (2011), Who is Responsible for Data? An Exploratory Study of Data Authorship, Ownership, and Responsibility, https://doi.org/10.1002/meet.2011.14504801188. Retrieved in 5 June 2020

[7] Pranav Shrivastava , Prerna Agarwal, WiFi Data Leakage Detection, International Symposium on Fusion of Science and Technology (ISFT 2020), IOP Conf. Series: Materials Science and Engineering 804 (2020) 012042, doi:10.1088/1757-899X/804/1/012042,

[8] Manisha J. Nene, Data Leakage Prevention for Data in Transit using Artificial Intelligence and Encryption Techniques Mohammed Ghouse DM, Network and Cyber Security, Bharat Electronics Limited, Bangalore, India mohammedghouse@bel.co.in Dept. of Computer Science and Engineering, Defense Institute of Advanced Technology, Pune, India mjnene@diat.ac.in VembuSelvi C MRS, Cyber Security Central Research Laboratory, BEL, Bangalore, India vembuselvic@bel.co.in

[9] Lv Xuming, Cao Lina, Ji Peng, Gao Xiao, Chen Shuo, Current status and future prospects of data leakage prevention technology: A brief review, , CISAT 2019, Journal of Physics: Conference Series 1345 (2019) 022010, doi:10.1088/1742-6596/1345/2/022010

[10] Weifeng Xu, A Forensic Evidence Acquisition Model for Data Leakage Attacks School of Criminal Justice University of Baltimore Baltimore, MD, USA wxu@ubalt.edu Jie Yan Department of Computer Science Bowie State University Bowie, MD, USA jyan@bowiestate.edu Hongmei Chi Department of Computer & Infor. Sciences Florida A&M University Tallahassee, FL, USA hongmei.chi@famu.edu

[11] Entao Luo, Md Zakirul Alam Bhuiyan, Guojun Wang, Md Arafatur Rahman, Jie Wu, and Mohammed Atiquzzaman, HUMAN-DRIVEN EDGE COMPUTING AND COMMUNICATION, PrivacyProtector: Privacy-Protected Patient Data Collection in IoT-Based Healthcare Systems, , IEEE Communications Magazine • February 2018, 10.1109/MCOM.2018.1700364]

[12] Chaoshun Zuo, IEEE Symposium on Security and Privacy, Why Does Your Data Leak? Uncovering the Data Leakage in Cloud from Mobile Apps, 2019, The Ohio State University Zhiqiang Lin The Ohio State University Yinqian Zhang The Ohio State University

[13] [22 J. Zhu, P. He, Q. Fu, H. Zhang, M. R. Lyu and D. Zhang, "Learning to log: Helping developers make informed logging decisions", Proceedings of the 37th International Conference on Software Engineering - Volume 1, pp. 415-425, 2015.].

[14] [23. T. Spring, “Insecure backend databases blamed for leaking 43tb of app data,” https://threatpost.com/ insecure-backend-databases-blamed-for-leaking-43tb-of-app-data/ 126021/, June 2017]

[15] B. Chen and Z. M. J. Jiang, "Characterizing and detecting anti-patterns in the logging code", Proceedings of the 39th International Conference on Software Engineering, pp. 71-81, 2017

[16] Rui Zhou, Mohammad Hamdaqa, Haipeng Cai, Abdelwahab Hamou-Lhadj, MobiLogLeak: A Preliminary Study on Data Leakage Caused by Poor Logging Practices, Dept. Electrical and Computer Engineering, Concordia University, Montreal, Canada †School of Computer Science, Reykjavik University, Iceland School of Electrical Engineering and Computer Science, Washington State University, Pullman, USA r_ou@encs.concordia.ca, mhamdaqa@ru.is, haipeng.cai@wsu.edu, wahab.hamou-lhadj@concordia.ca

[17] W.A. Lombard, The 4th industrial revolution : is it here? Source: FarmBiz 3, pp 6 –7 (2017), Publisher: Plaas Media, Persistent Link : https://hdl.handle.net/10520/EJC-c05137ec9]

[18] Moustafa, Nour, and Jill Slay. "UNSW-NB15: a comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set)."Military Communications and Information Systems Conference (MilCIS), 2015. IEEE, 2015.

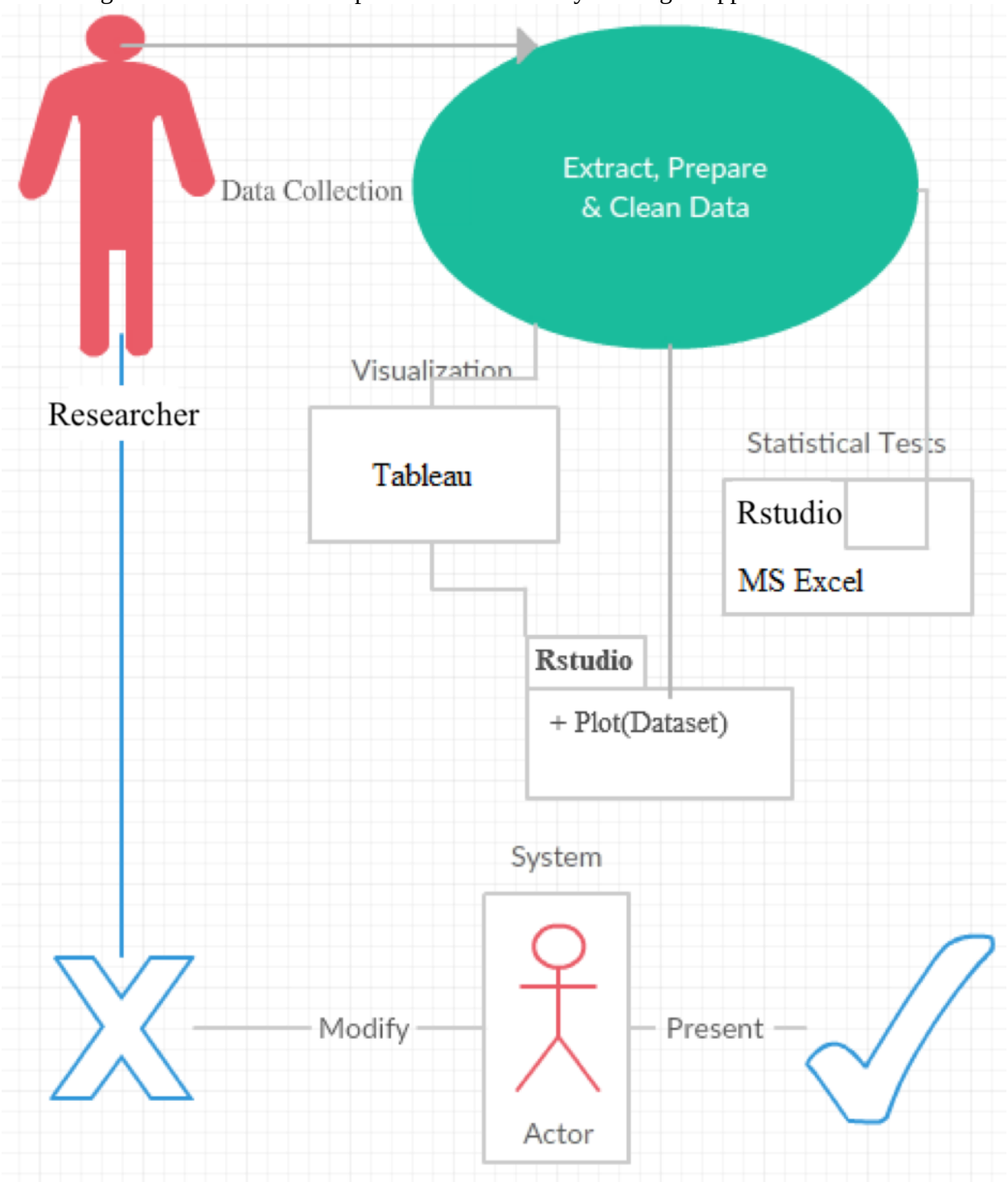
[19] Moustafa, Nour, and Jill Slay. "The evaluation of Network Anomaly Detection Systems: Statistical analysis of the UNSW-NB15 data set and the comparison with the KDD99 data set."Information Security Journal: A Global Perspective(2016): 1-14.

Figures & Images

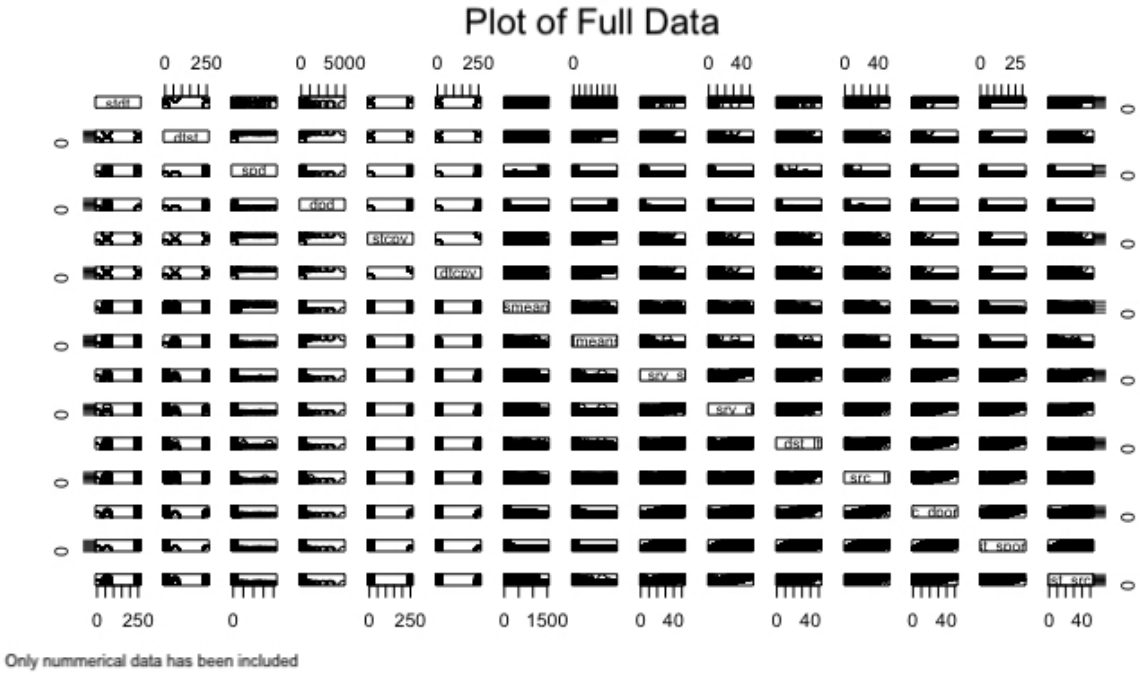
1. Figure 3, Statistical overview of Attack Dataset

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
$sport	0	1043	11653	23239	47439	65535	687
$dsport	0	53	53	667	53	65535	687
$duration	0	0	0	0.7711	0	59.9988	687
$stdkb	0.06	0.114	0.114	5.175	0.2	11063.467	687
$dtskb	0	0	0	3.501	0	12838.547	629
$stdt	0	254	254	242	254	255	687
$dtst	0	0	0	39.05	0	252	687
$spd	0	0	0	2.241	0	4158	687
$dpd	0	0	0	1.663	0	4829	687
$sload	0	50.67	57	100	114	5600	687
$dload	0	0	0	0.0092	0	2.3195	687
$stdpc	1	2	2	8.247	2	8324	687
$dtspc	0	0	0	4.654	0	9660	687
$stcpv	0	0	0	39.2	0	255	687
$dtcpv	0	0	0	39.2	0	255	687
$stcpn	0.00E+00	0.00E+00	0.00E+00	3.30E+08	0.00E+00	4.30E+09	687
$dtcpn	0.00E+00	0.00E+00	0.00E+00	3.27E+08	0.00E+00	4.30E+09	687
$smean	38	57	57	94.06	70.75	1504	687
$dmeans	0	0	0	31.28	0	1420	687
$trans_depth	0	0	0	0.0591	0	172	687
$res_bdy_len	0	0	0	949	0	5242880	687
$Sjitter	0	0	0	1902	0	1181164	687
$Djitter	0	0	0	201.8	0	120773.4	687
$Sintpkt	0	0	0.01	47.43	0.01	84371.5	687
$Dintpkt	0	0	0	29.27	0	15501.77	687
$synack	0	0	0	0.01	0	0.4642	687
$ackdat	0	0	0	0.0102	0	0.4566	687
$ct_srv_src	1	5	20	18.69	30	52	687
$ct_srv_dst	1	4	20	18.6	30	52	687
$ct_dst_ltm	1	2	14	13.04	18	51	687
$src_ltm	1	3	15	13.38	18	51	687
$src_dp_ltm	1	2	14	12.76	17	51	687
$dst_sp_ltm	1	2	12	9.976	16	31	687
$dst_src_ltm	1	4.25	20	18.69	30	52	687
$Label	1	1	1	1	1	1	687
$srcip	89581 character			$dstip	89581	character	
$protocol	89581 character			$attack_typ	89581	character	
$state	89581 character			$service	89581	character	

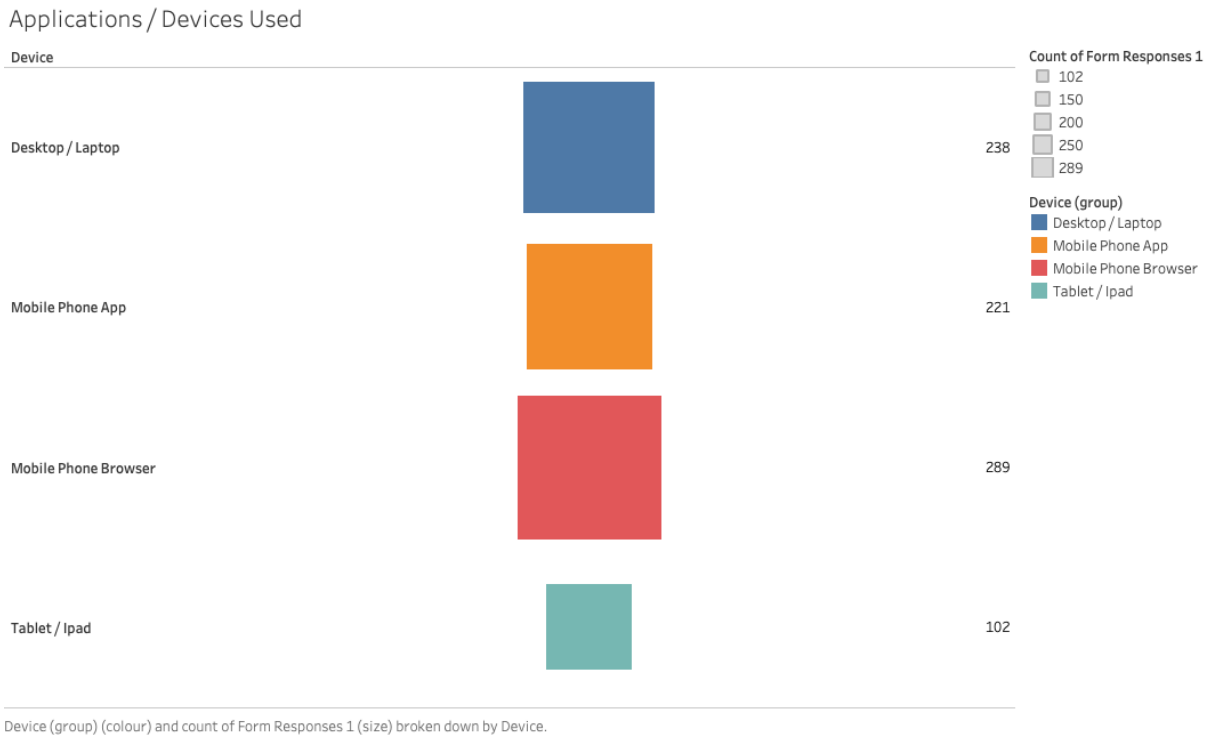
2. Figure 4. Overview of our performed Data Analysis/Insight Approach



3. Figure 5. Plotted Graph of the Attack Dataset, conducted in R

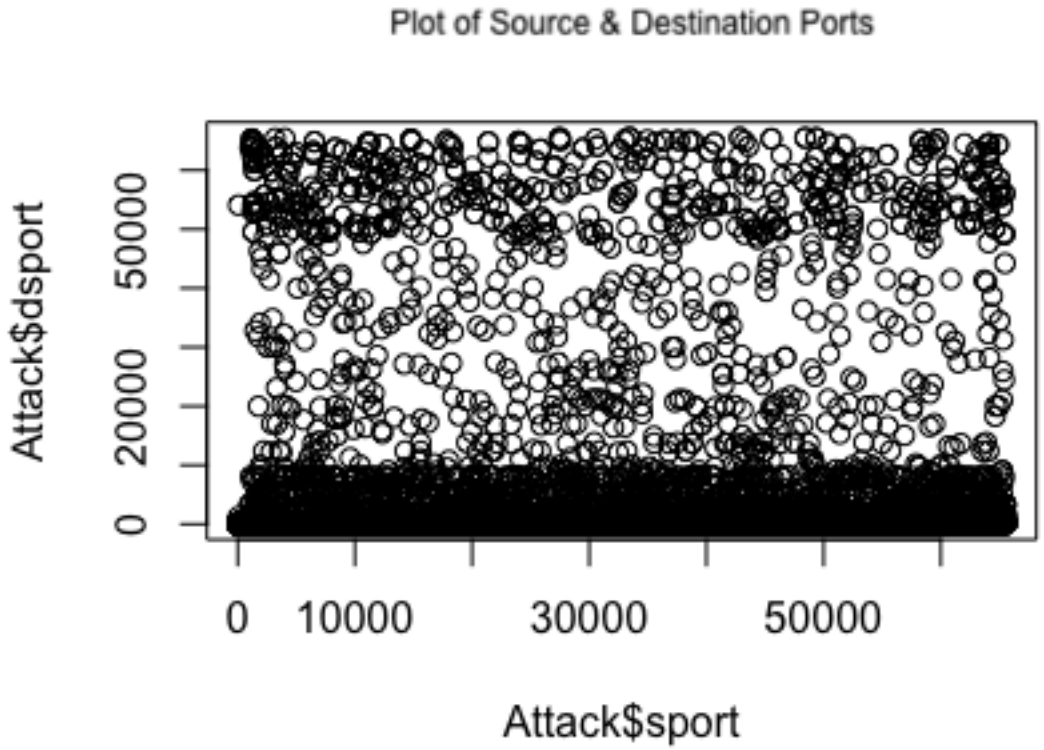


4. Figure 6. Device Type used by responders

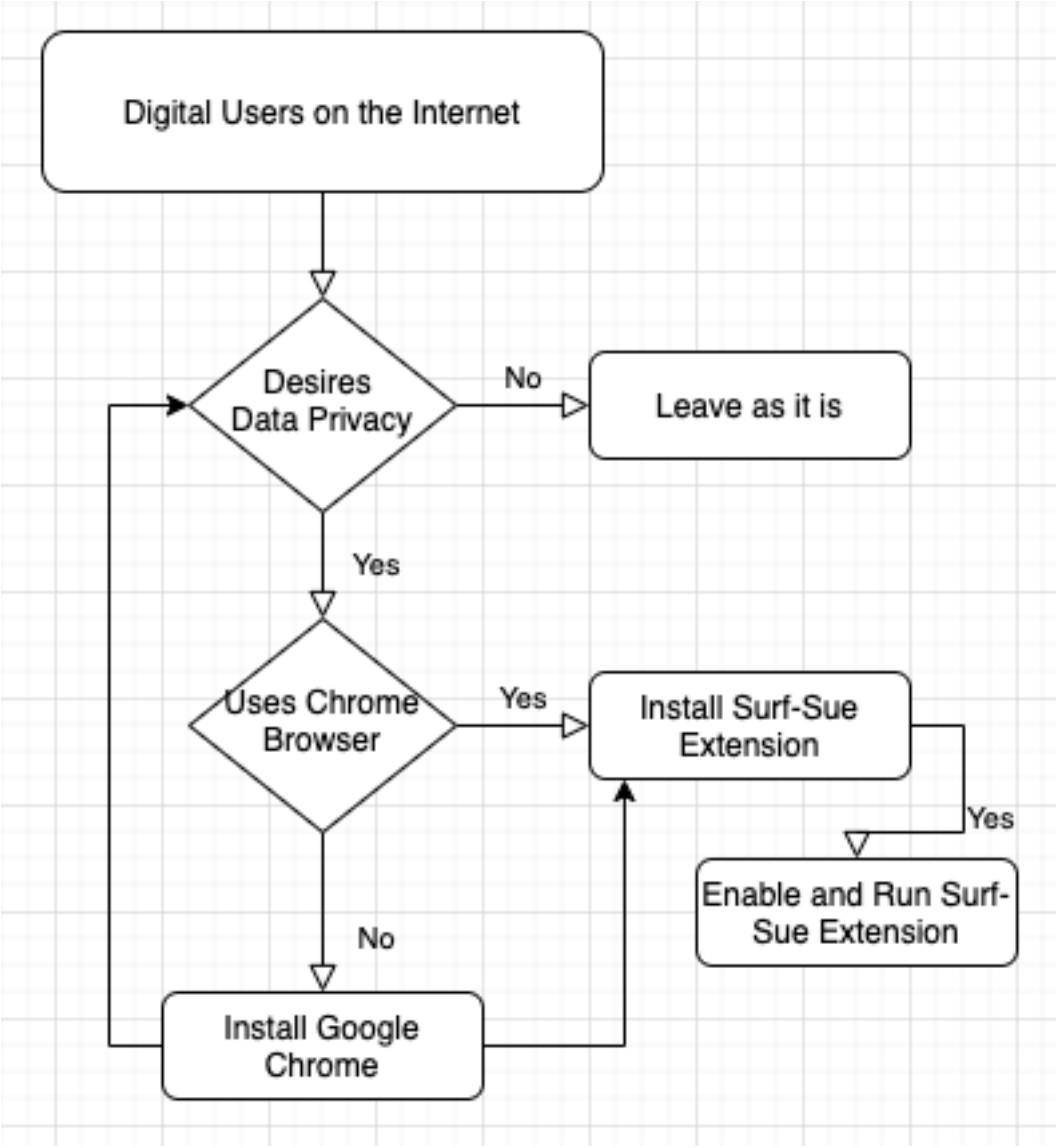


5. Figure 7. Services that victims were connected to during the exploit

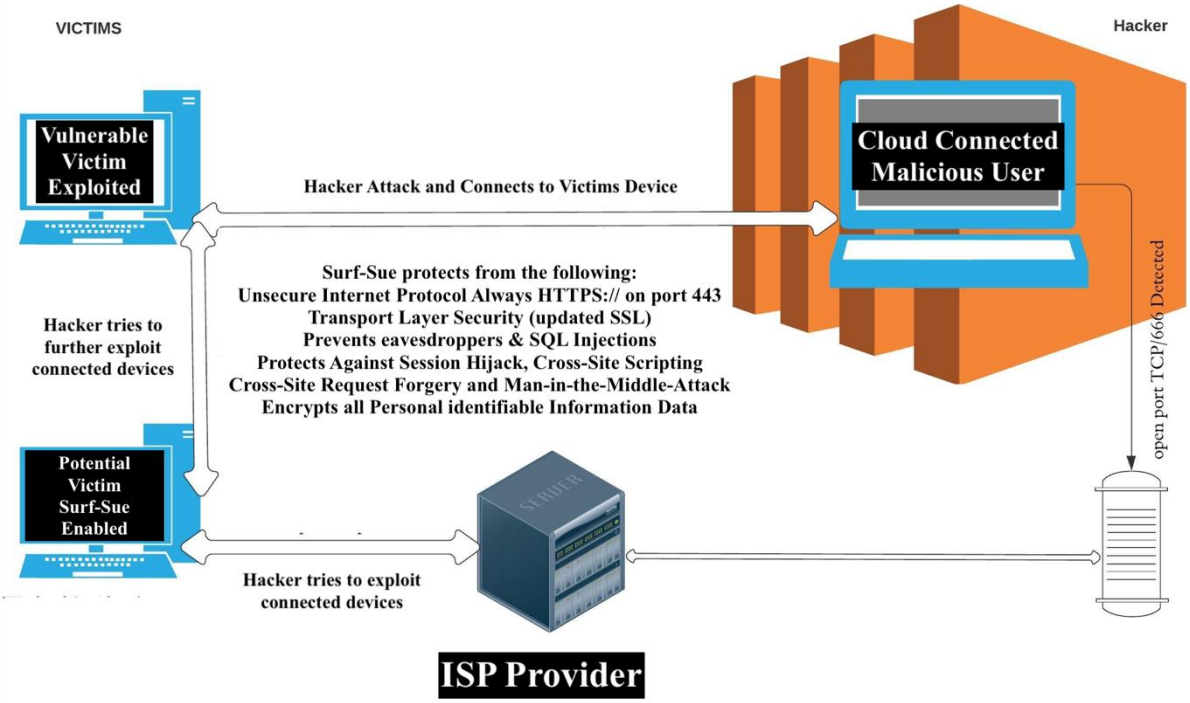
6. **Fig 7. Plot of Destination Ports v/s Source Ports**


Plot of Source & Destination Ports
Attack$dsport
0
20000
50000
0
10000
30000
50000
Attack$sport

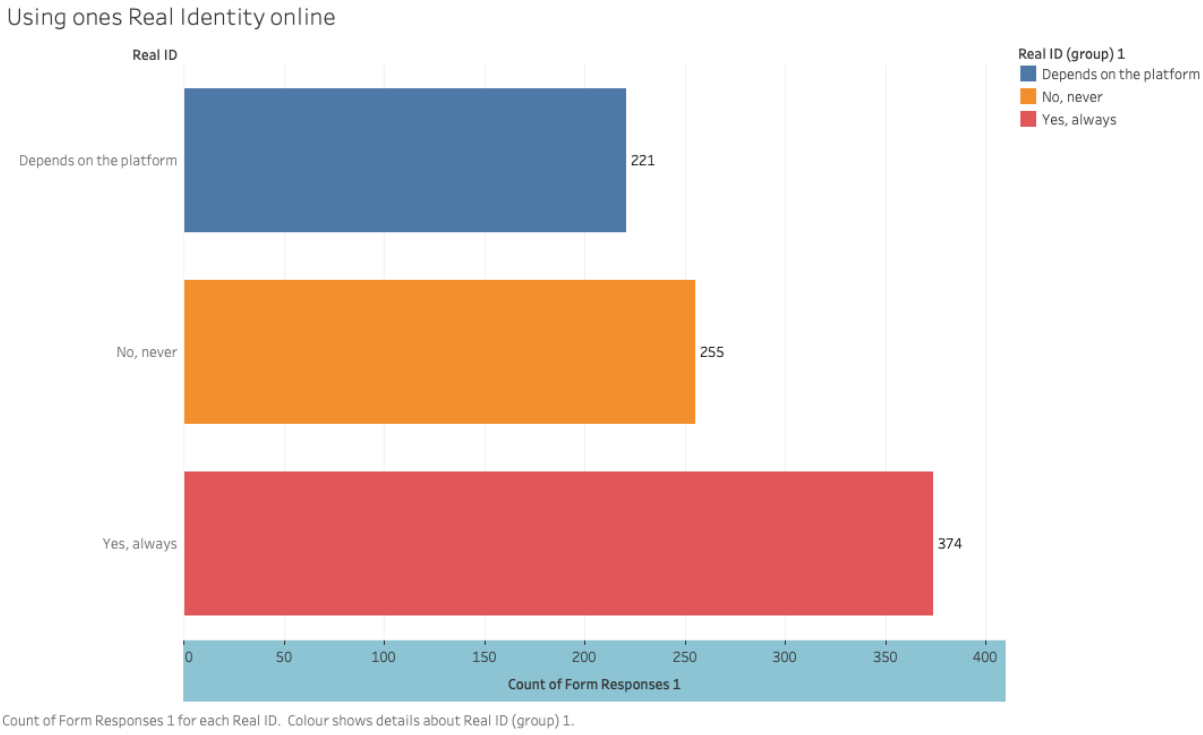
7. Figure 8. Business Process Model Notation of proposed Extension



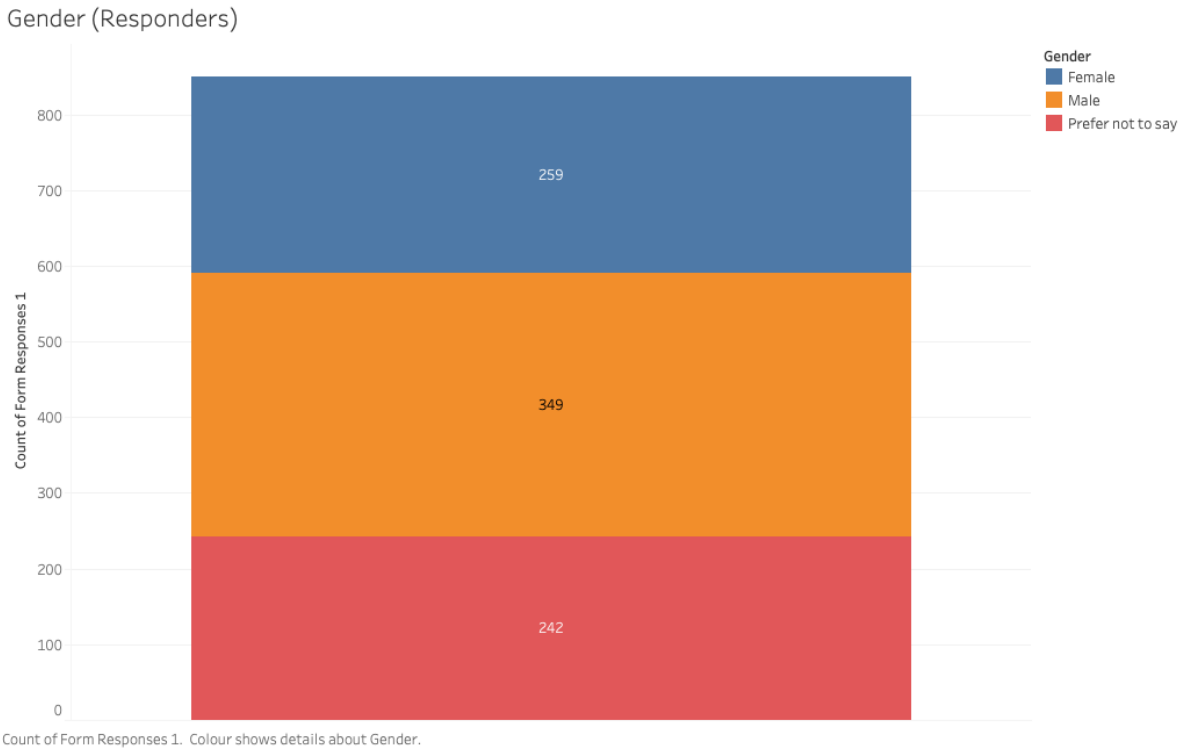
8. Figure 9. Exploitation Model for gaining malicious access



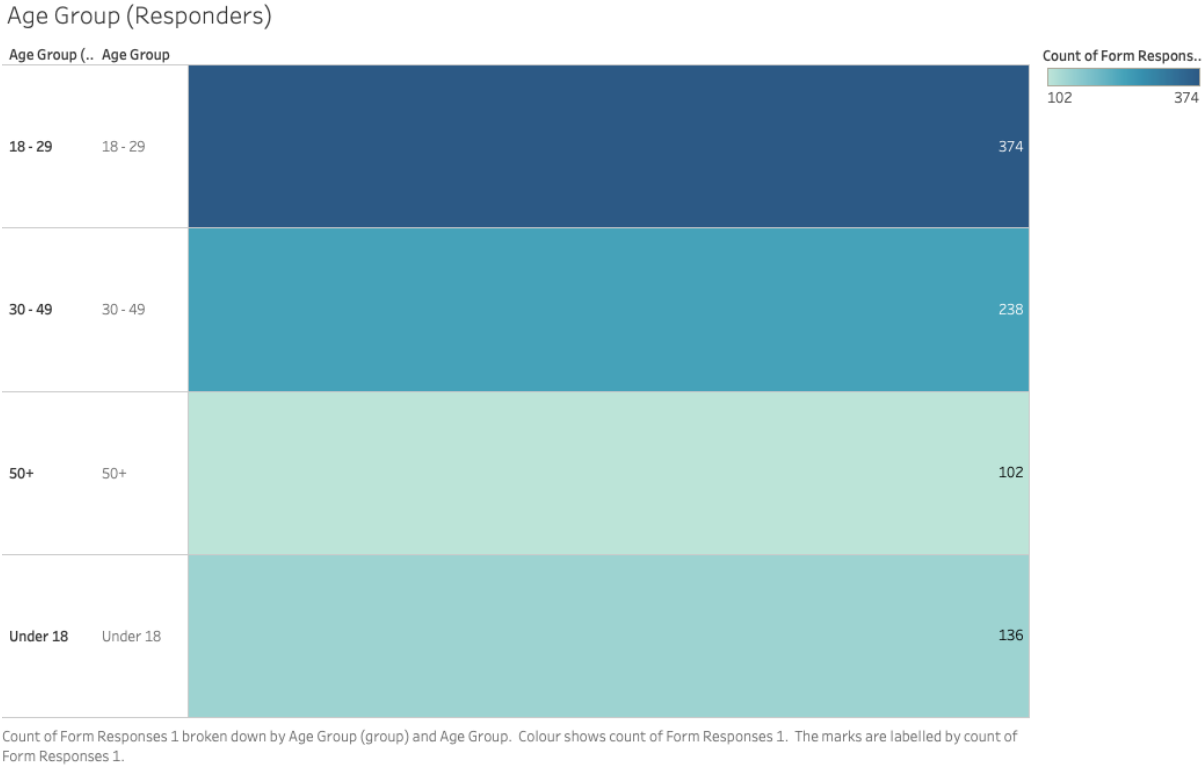
9. Figure 13. Survey outcome are you registering with your "Real" identity details.



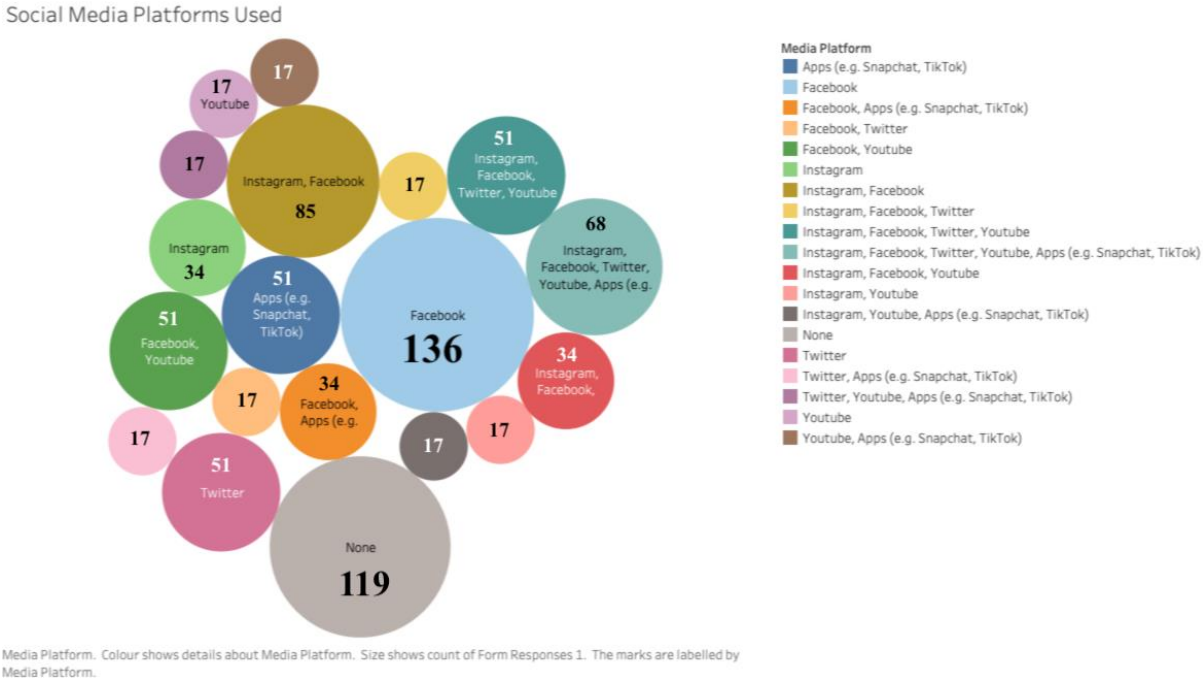
10. Figure 14. Survey responses sorted by Gender



11. Figure 15. Survey responses sorted by Age Group



12. figure 16. Social Media Platforms being used



13. **Figure 17. Outcome of Real identity being used on Social Media Platforms**

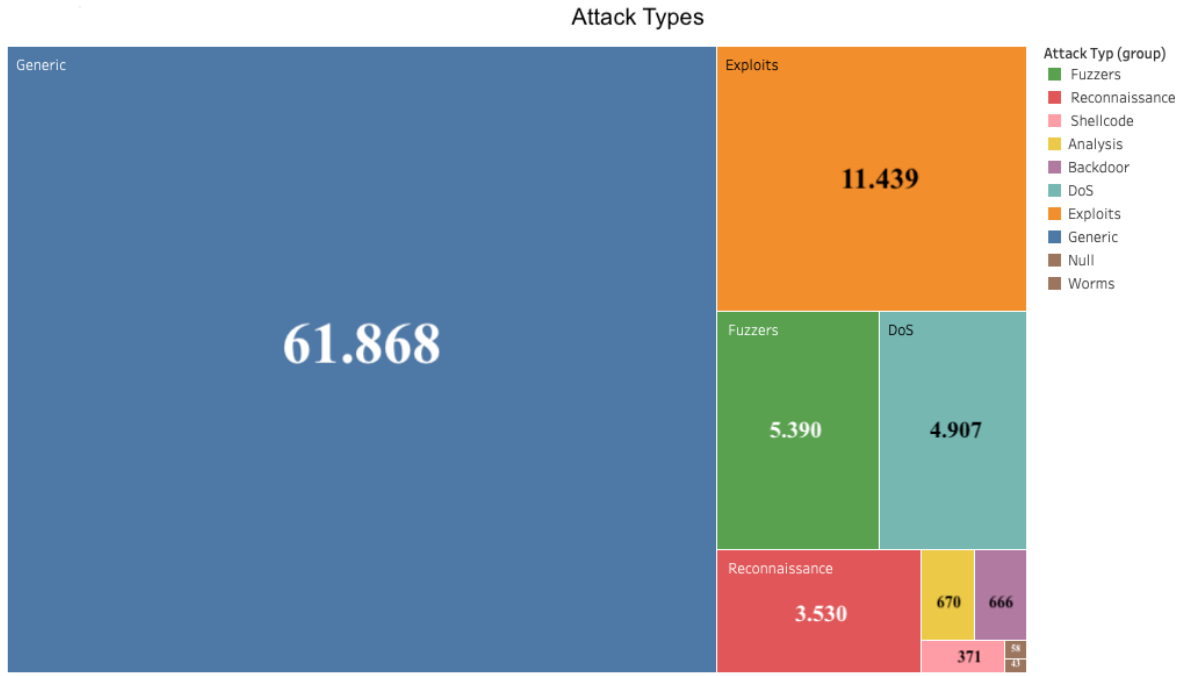
	Depends on the platform	No, never	Yes, always
Apps (e.g. Snapchat, TikTok)	0	17	34
Facebook	17	85	34
Facebook, Apps (e.g. Snapchat, TikTok)	0	17	17
Facebook, Twitter	17	0	0
Facebook, Youtube	17	0	34
Instagram	0	0	34
Instagram, Facebook	0	17	68
Instagram, Facebook, Twitter	17	0	0
Instagram, Facebook, Twitter, Youtube	34	17	0
Instagram, Facebook, Twitter, Youtube, Apps (e.g. Snapchat, TikTok)	0	17	51
Instagram, Facebook, Youtube	0	0	34
Instagram, Youtube	0	17	0
Instagram, Youtube, Apps (e.g. Snapchat, TikTok)	0	17	0
None	68	34	17
Twitter	34	0	17
Twitter, Apps (e.g. Snapchat, TikTok)	0	0	17
Twitter, Youtube, Apps (e.g. Snapchat, TikTok)	0	0	17
Youtube	0	17	0
Youtube, Apps (e.g. Snapchat, TikTok)	17	0	0

14. figure 18. Outcome of devices being used to access Social Media Platforms

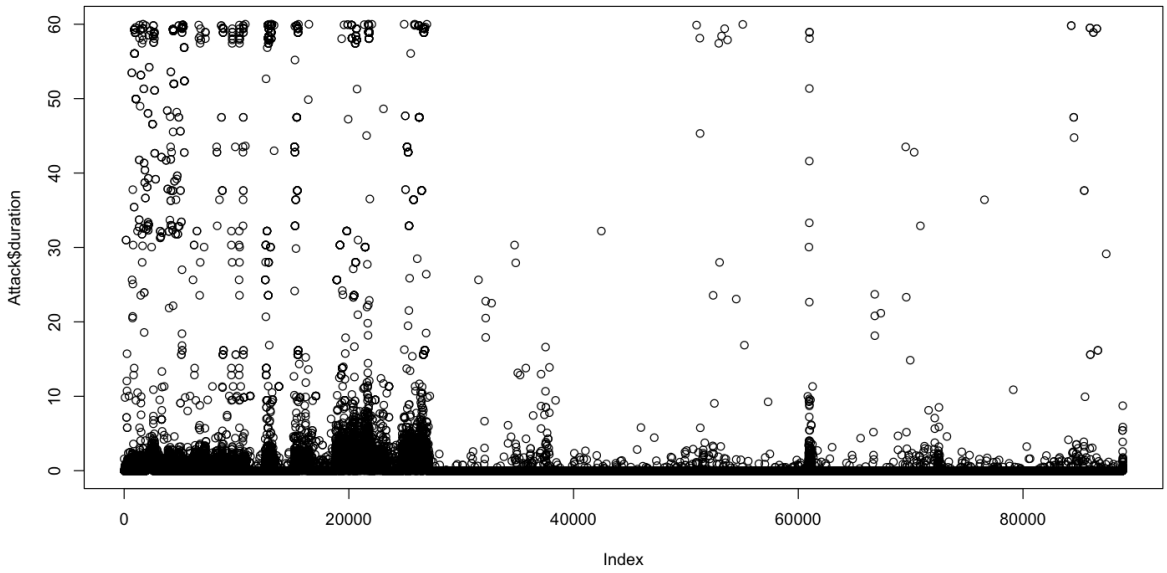
	Mobile Phone Browser	Tablet / Ipad
Apps (e.g. Snapchat, TikTok)	17	17
Facebook	85	17
Facebook, Apps (e.g. Snapchat, TikTok)	17	0
Facebook, Twitter	17	0
Facebook, Youtube	17	0
Instagram	0	0
Instagram, Facebook	34	17
Instagram, Facebook, Twitter	0	0
Instagram, Facebook, Twitter, Youtube	17	0
Instagram, Facebook, Twitter, Youtube, Apps (e.g. Snapchat, TikTok)	17	0
Instagram, Facebook, Youtube	0	0
Instagram, Youtube	0	0
Instagram, Youtube, Apps (e.g. Snapchat, TikTok)	17	0
None	17	0
Twitter	17	0
Twitter, Apps (e.g. Snapchat, TikTok)	0	17
Twitter, Youtube, Apps (e.g. Snapchat, TikTok)	0	17
Youtube	17	0
Youtube, Apps (e.g. Snapchat, TikTok)	0	17

	Desktop / Laptop	Mobile Phone App
Apps (e.g. Snapchat, TikTok)	17	0
Facebook	17	17
Facebook, Apps (e.g. Snapchat, TikTok)	0	17
Facebook, Twitter	0	0
Facebook, Youtube	34	0
Instagram	34	0
Instagram, Facebook	0	34
Instagram, Facebook, Twitter	0	17
Instagram, Facebook, Twitter, Youtube	17	17
Instagram, Facebook, Twitter, Youtube, Apps (e.g. Snapchat, TikTok)	17	34
Instagram, Facebook, Youtube	0	34
Instagram, Youtube	0	17
Instagram, Youtube, Apps (e.g. Snapchat, TikTok)	0	0
None	102	0
Twitter	0	34
Twitter, Apps (e.g. Snapchat, TikTok)	0	0
Twitter, Youtube, Apps (e.g. Snapchat, TikTok)	0	0
Youtube	0	0
Youtube, Apps (e.g. Snapchat, TikTok)	0	0

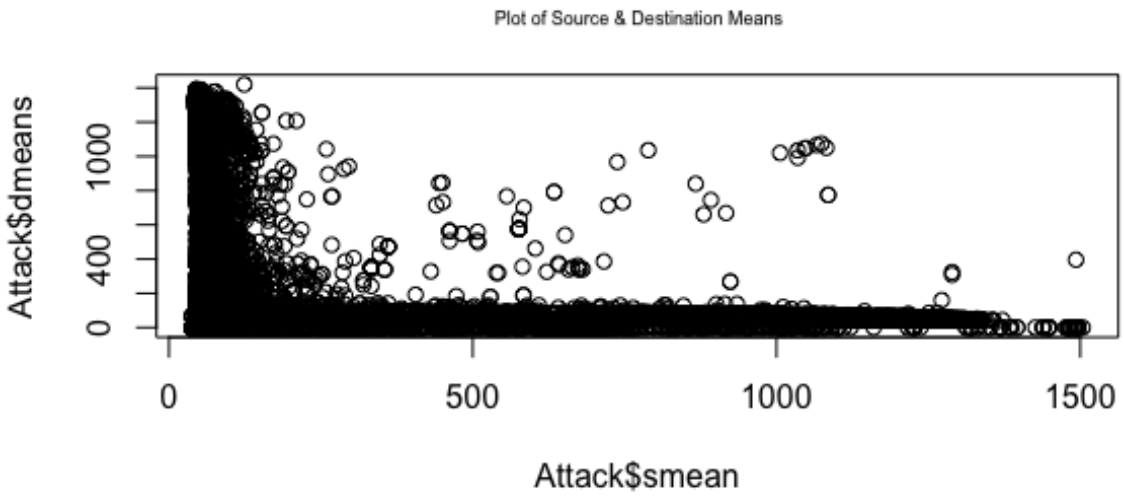
15. Figure 19. Most occurring Attack Types



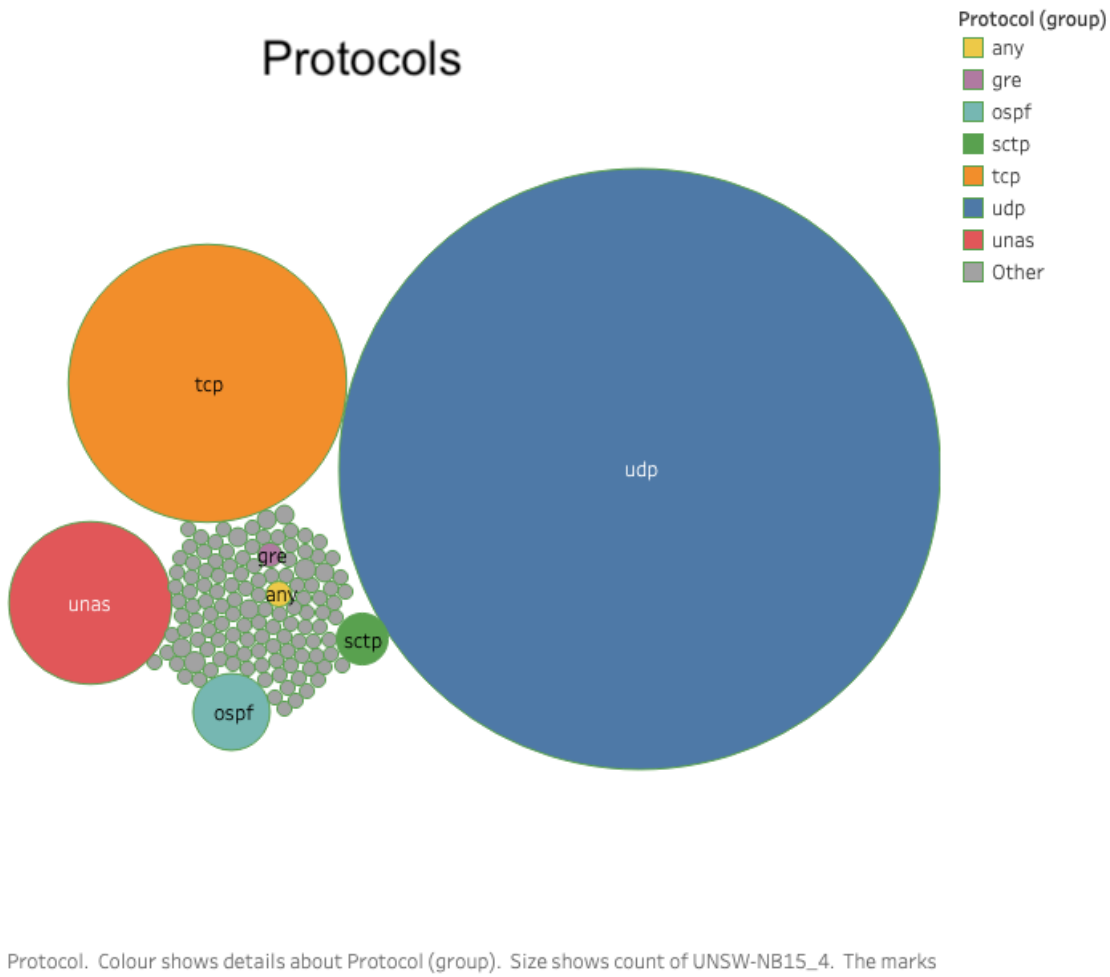
16. Figure 20. Plot of Attack Duration



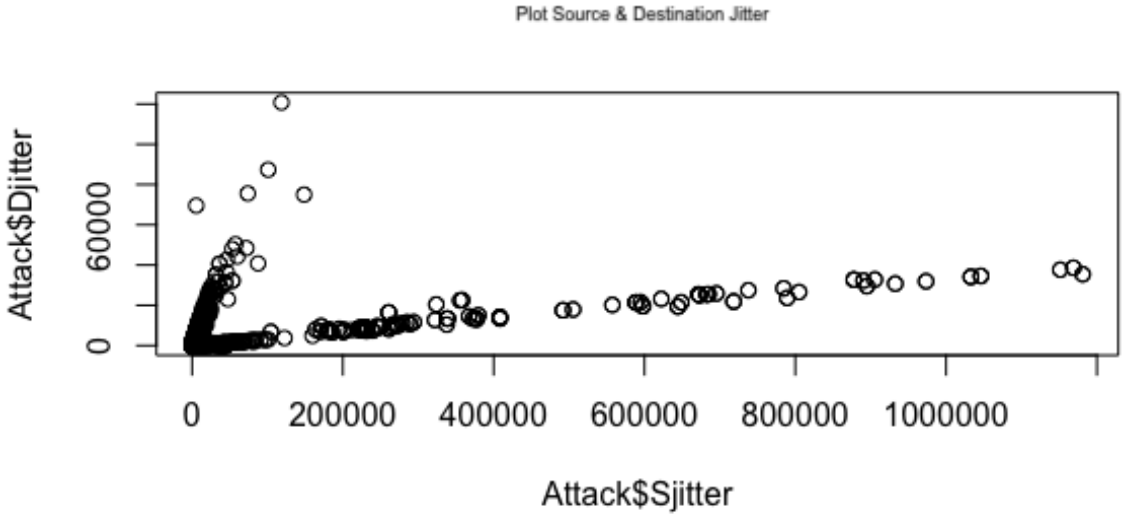
17. Figure 21. Plotted Graph with the mean values of transmitted packets from the source to the destination and from the destination to the source.

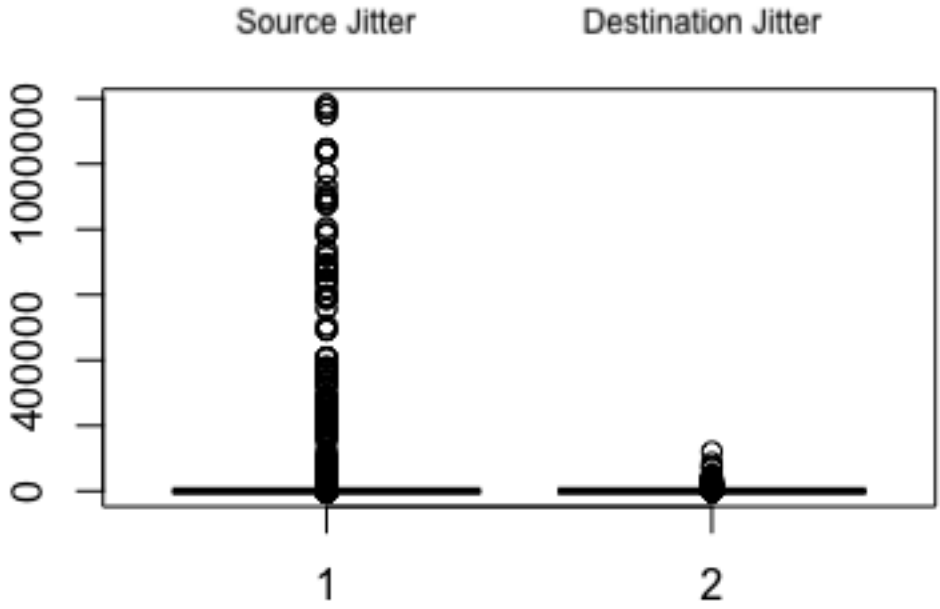


18. Figure 22 Affected Protocols leading to a successful attack

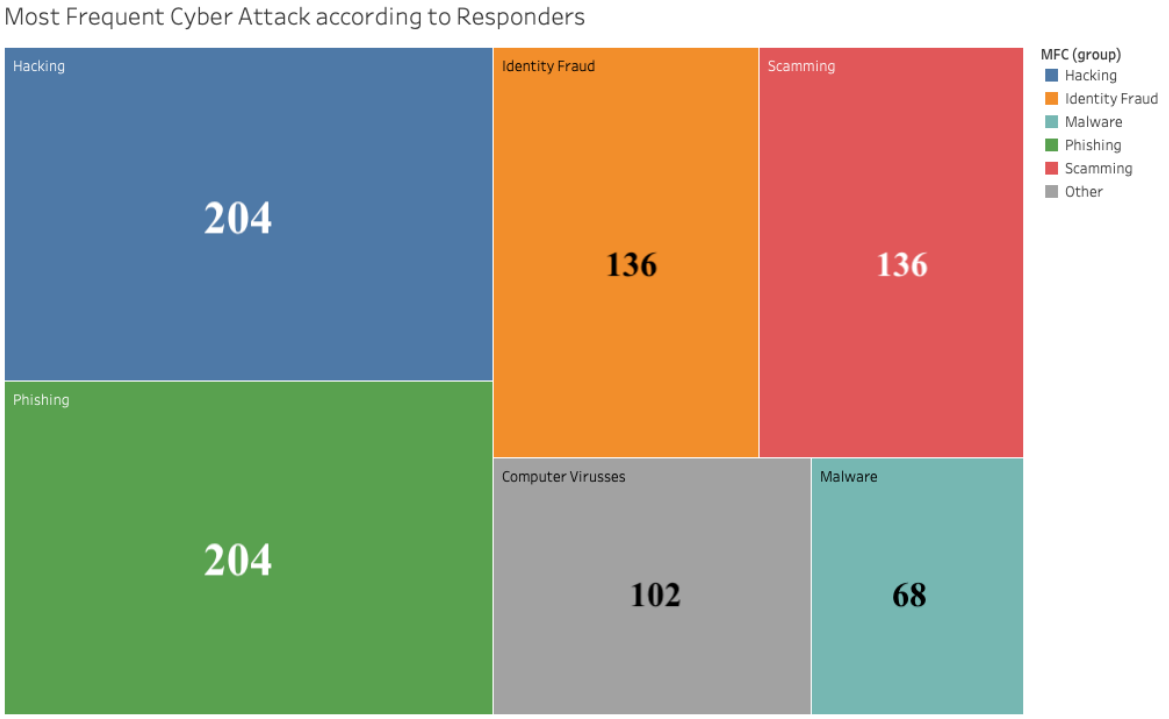


19. **Figure 23. plot of jitter values in Milliseconds**

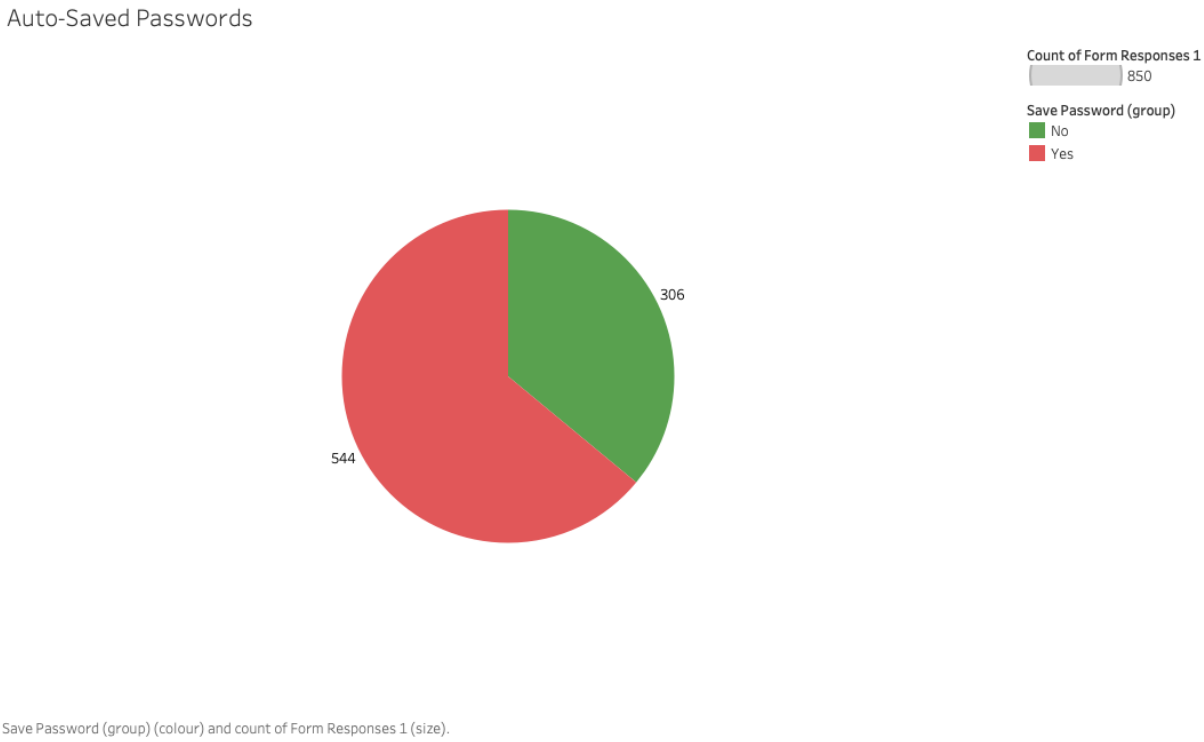




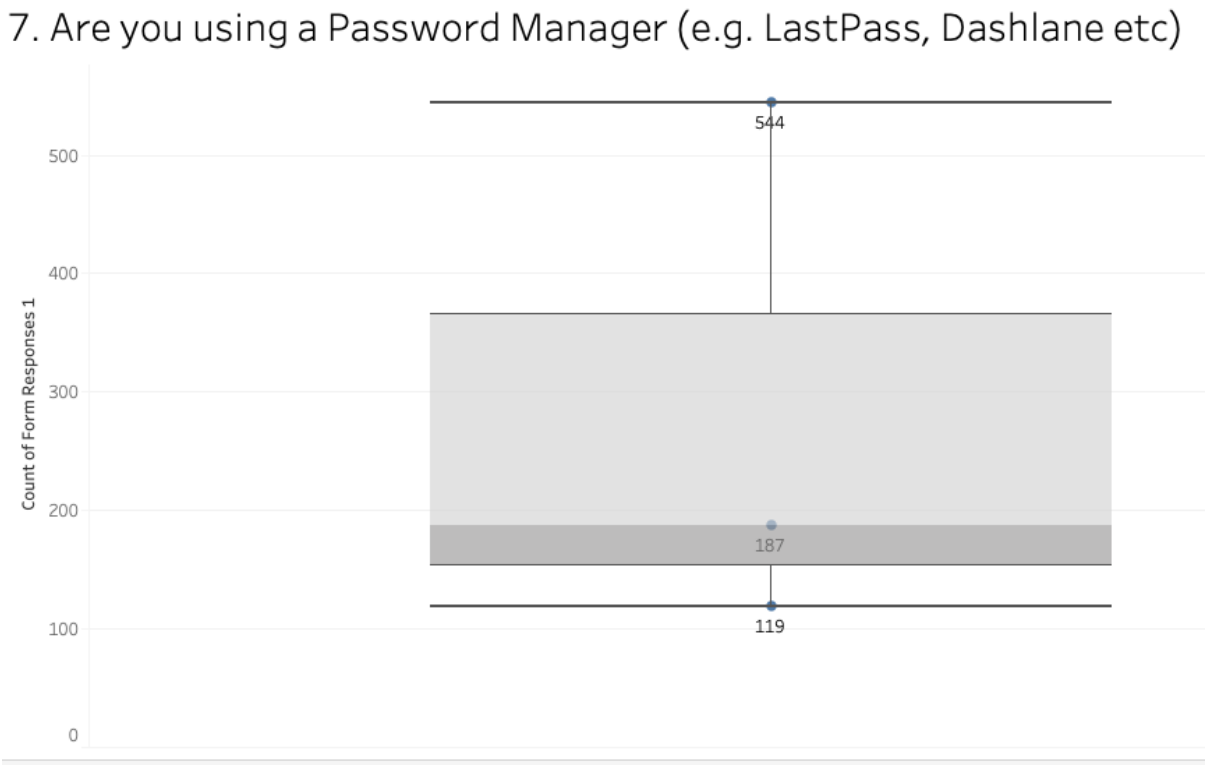
20. Figure 24. Most Frequent Occurring Cyber Attacks according our survey responders



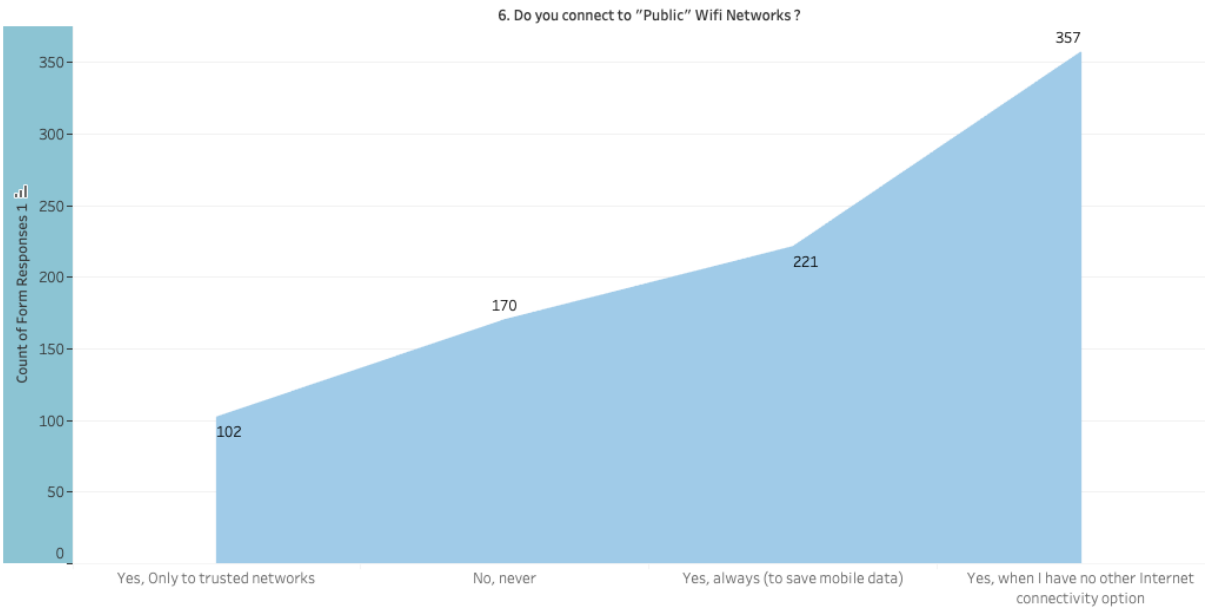
21. Figure 25 Amount of responders that auto-save their passwords



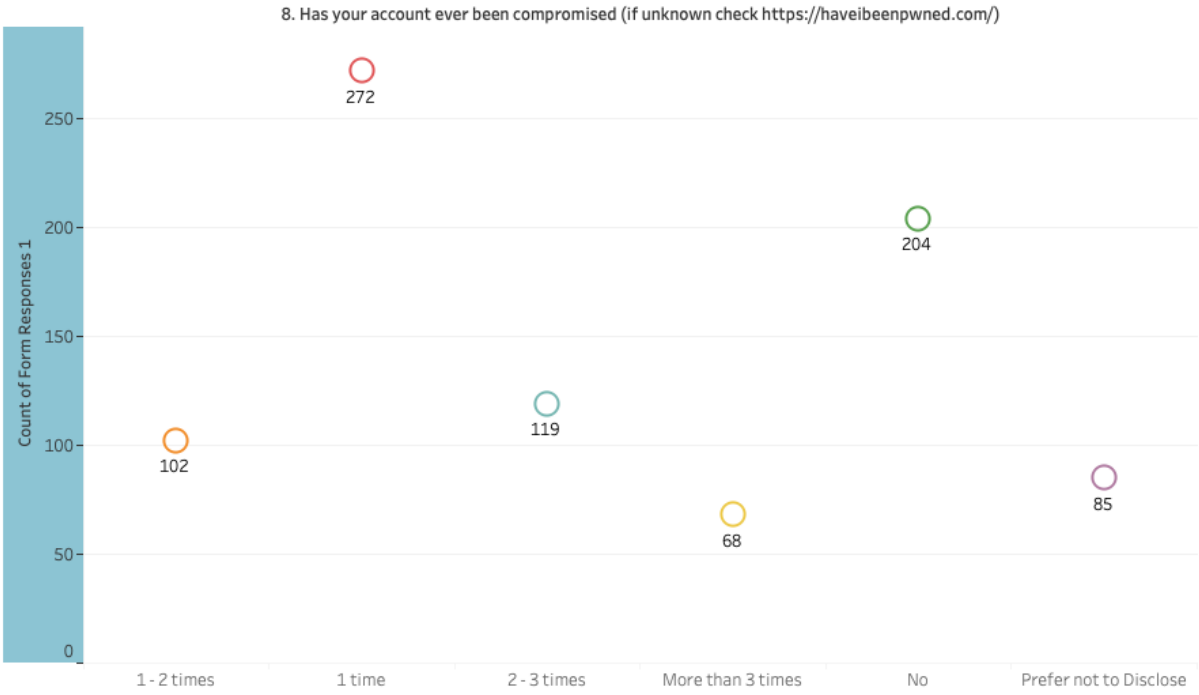
22. Figure 26 Use of Password Managers (to store passwords) by responders



23. Figure 27 Amount of Responders that connect to Public Networks



24. Figure 28. Amount of responders that got one of their accounts ever compromised



25. Figure 29. Left: people that connect to public networks v/s Right: amount of compromises

```
> print.table(my_table_3)

                                                    1 - 2 times 1 time
No, never                                                    17     17
Yes, always (to save mobile data)                            34    102
Yes, Only to trusted networks                                17     34
Yes, when I have no other Internet connectivity option       34    119

                                                    2 - 3 times
No, never                                                    0
Yes, always (to save mobile data)                           34
Yes, Only to trusted networks                                0
Yes, when I have no other Internet connectivity option      85

                                                    More than 3 times  No
No, never                                                           0 102
Yes, always (to save mobile data)                                  34   0
Yes, Only to trusted networks                                      17  34
Yes, when I have no other Internet connectivity option             17  68

                                                    Prefer not to Disclose
No, never                                                               34
Yes, always (to save mobile data)                                       17
Yes, Only to trusted networks                                            0
Yes, when I have no other Internet connectivity option                  34
```

Appendix

1. Glossary

Business Email Compromise: Targeting companies by posing to be one of their partners or vendors and request (outstanding) payments
Clickjacking: Redirect a genuine website visitor to perform an vulnerable act, unknown to them.
Confidence/Romance Fraud: Gaining someone's trust, though partnership, either business related of love affair and request frequent payments.
Extortion: Obtaining money by force
Non-Payment/Non-Delivery: Money transferred but no Product/Service received, or Product/service delivered by no money received.
Personal Data Breach: Obtaining Personal Identifiable information through unethical ways.
Pharming: redirecting a genuine website to a fraudulent website, without the user noticing it.
Phishing: Spam emails being sent, in the hope to either receive sensitive information, or that the recipient performs a certain action.
Smishing: Same as above, but with the use of text messages, like your parcel has been secured at our depot, provide your details HERE
Social Engineering: Investigate people, their background employment etc. to become a potential target during a Cyber Attack.
Spoofing: disguise communication, imposter trying to represent a trusted source.
TCP/IP: Transmission Control Protocol / Internet Protocol, connection is established prior to transferring data.
UDP: User Datagram Protocol, connectionless communication some data maybe lost, ideal for watching live stream.
Vishing: Similar to Scam calls, pretending to be from a reliable or trusted source, and ask for a payment etc.

2. Occurred Data Breaches

Dubsmash	LinkedIn	My Fitness Pal
Date: December 2018	**Date:** 2012 (and 2016)	**Date:** February 2018
Impact: 162 million accounts	Impact: 165 million accounts	Impact: 150 million accounts
Details: email addresses, usernames, PBKDF2 password hashes, and other personal data such as dates of birth stolen, all of which was then put up for sale	Details: As the major social network for business professionals, attackers looking to conduct social engineering attacks. In 2012, 6.5 million unassociated passwords (un salted SHA-1 hashes) were stolen by attackers. it wasn't until 2016 that the full extent of the incident was revealed. The data was found to be offering the email addresses and passwords of around 165 million LinkedIn users	Details: among the massive information dump of 16 compromised sites that saw some 617 million customers accounts leaked, the usernames, email addresses, IP addresses, SHA-1 and bcrypt-hashed passwords of around 150 million customers were stolen and then put up for sale a year later
Acknowledgement: https://dubsmash.com/user-notice	Acknowledgement: https://blog.linkedin.com/2016/05/18/protecting-our-members	Acknowledgement: https://content.myfitnesspal.com/security-information/FAQ.html

Sina Weibo	Zynga
Date: March 2020	**Date:** September 2019
Impact: 538 million accounts	Impact: 218 million accounts
With over 500 million users, Sina Weibo is China's answer to Twitter. However, in March 2020 it was reported that th e real names, site usernames, Details: gender, location, and -- for 172 million users -- phone numbers had been posted for sale	a hacker who goes by the name Gnosticplayer hacked into Zynga's database and gained access to t he 218 million accounts registered there. Zynga lat er confirmed that email addresses, salted SHA-1 hashed passwords, phone numbers, and user IDs for Facebook and Zynga accounts were stolen.
	Acknowledgement: https://investor.zynga.com/news-releases/news-release-details/player-security-announcement

3. Retrievable Data

A device is being booted, either desktop, portable or mobile, after which connected to the a network and able to access the Internet. A internet browser is usually required to see graphics instead of few lines with code, these browsers have the ability to store following capture information, and retrieve it as well at any desired time:

• Network: Mac Address, Device Id, Serial Number, Country and Package Manager.
• Account: Name and Size.
• Location: Latitude, Longitude, and Last Know Location.
• Database: ID, Password, Subdomain, Website Link Name, etc.

4. Header names of Attack Dataset

No.	Name	Type	Description
1	$ srcip	nominal	Source IP address
2	$ sport	integer	Source port number
3	$ dstip	nominal	Destination IP address
4	$ dsport	integer	Destination port number
5	$ protocol	nominal	Transaction protocol
6	$ state	nominal	Indicates to the state and its dependent protocol
7	$ duration	Float	Record total duration
8	$ stdkb	Integer	Source to destination transaction bytes
9	$ dtskb	Integer	Destination to source transaction bytes
10	$ stdt	Integer	Source to destination time to live value
11	$ dtst	Integer	Destination to source time to live value
12	$ spd	Integer	Source packets retransmitted or dropped
13	$ dpd	Integer	Destination packets retransmitted or dropped
14	$ service	nominal	http, ftp, smtp, ssh, dns, ftp-data ,irc
15	$ sload	Float	Source megabits per second
16	$ dload	Float	Destination megabits per second
17	$ stdpc	integer	Source to destination packet count
18	$ dtspc	integer	Destination to source packet count
19	$ stcpv	integer	Source TCP window advertisement value
20	$ dtcpv	integer	Destination TCP window advertisement value
21	$ stcpn	integer	Source TCP base sequence number
22	$ dtcpn	integer	Destination TCP base sequence number
23	$ smean	integer	Mean of packet size transmitted by the src
24	$ dmeans	integer	Mean of packet size transmitted by the dst
25	$ trans_depth	integer	Represents the depth into the connect of http transaction
26	$ res_bdy_len	integer	Actual uncompressed data size transferred from http service.
27	$ Sjitter	Float	Source jitter (mSec)
28	$ Djitter	Float	Destination jitter (mSec)
29	$ Sintpkt	Float	Source interpacket arrival time (mSec)
30	$ Dintpkt	Float	Destination interpacket arrival time (mSec)
31	$ synack	Float	The time between the SYN and the SYN_ACK packets.
32	$ ackdat	Float	The time between the SYN_ACK and the ACK packets.
33	$ srv_src	integer	No. of connect that contain the same service and source address
34	$ srv_dst	integer	No. of connect that contain same service and destination address
35	$ dst_ltm	integer	No. of connections of the same destination address
36	$ src_ ltm	integer	No. of connections of the same source address
37	$ src_dp_ltm	integer	No of connect. of same source address and the destination port
38	$ dst_sp_ltm	integer	No of connect of the same dest. address and the source port
39	$ dst_src_ltm	integer	No of connect of the same source and the destination address

40	$ attack_typ	nominal	The name of each attack category. In this data set.
41	$ Label	binary	0 for normal and 1 for attack records