Lost Packet Warehousing Service

I Burke, M Motlhabi, R Netshiya and H Pieterse
Council for Scientific and Industrial Research (CSIR), Pretoria, South Africa
iburke@csir.co.za
mmotlhabi@csir.co.za
rnetshiya@csir.co.za
hpieterse@csir.co.za

Abstract: Recently, well-known and established South African organisations have experienced cyberattacks. South African Bank Risk Information Centre (SABRIC) confirmed in October 2019 that the industry had been hit by a wave of Distributed Denial of Service (DDoS) attacks targeting multiple banks. This happened shortly after the website of City of Johannesburg (CoJ) succumb to a ransomware attack. These attacks are a wakeup call for South African organisations and underline the essential need for suitable detection mechanisms to prevent cyberattacks. The detection of cyberattacks relies not only on understanding existing attacks but also being able to identify emerging threats. The continuous and strategic collection of relevant and valuable cybersecurity data sets can offer insight into ongoing threats or cyberattacks, while also assisting with the combatting of cybercrime. Although various third-party providers, such as Shodan and Have I Been Pwned (HIBP), exist and do provide access to cybersecurity data sets, these providers have little to no presence in South Africa (SA). Most of the available cybersecurity data sets are heavily slanted towards the United States and the identified trends might not be relevant to the South African context. Therefore, this paper introduces the Lost Packet warehousing Service, a technological solution that will function as the primary source for cybersecurity data collection within South Africa. The Lost Packet Warehousing Service will allow for the continuous but passive collection of cybersecurity data sets. Examples of such data sets could include network telescope, honeypot and NetFlow collectors. Data analysis and processing techniques are then applied to the collected cybersecurity data sets to identify, infer, detect and predict emerging trends and cyberattacks. Also discussed in this paper is the steps taken to maintain the security and privacy of the collected cybersecurity sets. The paper concludes by discussing the various benefits offered by the Lost Packet Warehousing Service.

Keywords: Cyberattacks, Cybersecurity, Threat Intelligence, Sensors, Data Analysis, Processing.

1. Introduction

It is not a question of if you will suffer a cyberattack but rather a question of when. A warning often promoted by cybersecurity experts to emphasise the need to deploy appropriate cybersecurity defensive mechanisms. However, cybersecurity is undergoing massive shifts in technology and its associated operations. The change is driven by the increasing dependency on Information Technology (IT) infrastructure, digitalisation, and the growth of the Internet of Things (IoT) devices (Sarkar, et al., 2020). Such changes increase the attack surface within the cyber domain, permitting the execution of cyberattacks. A cyberattack refers to the "deliberate actions to alter, disrupt, deceive, degrade, or destroy computer systems or networks or the information and/or programs resident in or transiting these systems or networks" (Owens, Dam, & Lin, 2009). Cyberattacks are threats that every system developer and administrators need to be familiar with.

Recently, well-known and established South African organisations have experienced cyberattacks. During October 2019, CoJ reported a network breach after receiving a ransom note from a group called the Shadow Kill Hackers. As a result, CoJ shutdown several customer-facing systems – including the website, e-services and billing system – as a precautionary measure (Moyo, 2019). Shortly afterwards, SABRIC reported that South African banks were targeted by DDoS attacks, which involved a ransom note being delivered via e-mail to all publicly available addresses (Moyo, 2019). With the arrival of the COVID-19 pandemic, attackers changed their focus to a new target. During June 2020, the second-largest private hospital operator in SA, Life Healthcare Group, fell victim to a cyberattack. Although the full extent of the attack is still unclear, the Life Healthcare Group did confirm the attack affected admissions systems, business processing systems and e-mail servers (Mungadze, 2020a). Roughly two months later, SA suffered a massive data breach when Experian, credit bureau agency, exposed personal information to a suspected fraudster. The exposed personal information include approximately 24 million South Africans, as well as 800 000 business entities (Moyo, 2020). Chief executive officer of Kaspersky Lab, Eugene Kaspersky, believed such sudden spikes in cyberattacks is a result of inadequate investment in cybersecurity and expect the attacks to continue in the near future (Mungadze, 2020b).

Such attacks must be a wakeup call for South African organisations and underline the need for suitable defences to proactively detect and prevent cyberattacks. Cyberattacks have been identified to follow a pattern (Mayfield, et al., 2018) and the answer to defend against such cyberattacks lies within the available cybersecurity data sets. Cybersecurity data sets are data collected by various technologies and processes used to protect IT infrastructure from cyberattacks. The continuous and strategic collection of relevant and valuable cybersecurity data sets can offer insight into ongoing threats or cyberattacks. Although various third-party providers, such as Shodan and HIBP, exist and do provide access to cybersecurity data sets, these providers have little to no presence in South Africa. The recent increase of cyberattacks affecting South African citizens and organisations, as well as the limited insight into these cyberattacks, emphasise the need for a technological solution that will enable the collection of cybersecurity data sets.

This paper introduces the Lost Packet Warehousing Service, a technological solution that will function as the primary source for cybersecurity data set collection within SA. The Lost Packet Warehousing Service will enable the continuous but passive collection of cybersecurity data sets. Examples of such data sets include network telescope, honeypot and NetFlow data. Data analysis and processing techniques are then applied to the collected cybersecurity data sets to identify, infer, detect and predict emerging trends and cyberattacks. With the collection of such large amounts of cybersecurity data sets, maintaining the security and privacy of the collected data sets are of critical importance. It is, therefore, necessary to consider the steps that must be taken to maintain the security and privacy of the collected cybersecurity sets from a South Africa legislative perspective.

The remainder of this paper is structured as follows. Section 2 introduces and discuss the various data sensors to be deployed as part of the Lost Packet Warehousing Service. In Section 3, the deployment and architectural design of the Lost Packet Warehousing Service are presented. Recommended data processing and analysis methodology for the collected cybersecurity data sets are deliberated in Section 4. Section 5 discusses the steps to consider to maintain the security and privacy of the collected cybersecurity data sets. The paper is concluded in Section 6.

2. Identification of Data Sensors

Collection of cybersecurity data sets on the Internet is a formidable undertaking. The Internet is expanding at a rapid pace, which causes data readily available for collection to grow exponentially. For sufficient collection of cybersecurity data sets, a variety of distributed data sensors will be required. Therefore, the Lost Packet Warehousing Service must deploy appropriate data sensors to enable the adequate collection of cybersecurity data sets. The recommended data sensors for the Lost Packet Warehousing Service are honeypots, network telescopes and NetFlow collectors. This section outlines the concept and use of each sensor with regards to the collection of cybersecurity data sets.

2.1 Honeypots

Honeypots masquerade as a vulnerable service or host to solicit bi-directional interaction (Hunter, et al., 2013). Since honeypots do not offer any legitimate services, all activity is deemed unauthorised and potentially malicious. Honeypots enable the capturing of malware, vulnerability exploitation, active network defences, credentials used in attacks, as well as the compromise of a system (Hunter & Irwin, 2011). Recommended honeypot implementations to be included and form part of the Lost Packet Warehousing Service include the following:

- **Honeyd**[1]: is a low-interaction virtual honeypot that creates and simulates hosts at the network level (Provos, 2004)
- **Kippo**[2]: emulates a Secure Shell (SSH) service designed to capture authentication credentials used in brute-force attacks, shell interaction or keystrokes perform by an attacker, as well as file download attempts (Hunter, et al., 2013).
- **Glastopf**[3]: emulates a vulnerable web server hosting web applications and web pages with various vulnerabilities.

[1] http://www.honeyd.org/
[2] https://www.honeynet.org/projects/old/kippo/
[3] https://www.honeynet.org/projects/old/glastopf/

- **Lyrebird[4]:** is a high-interaction honeypot framework that exposes real vulnerable applications.

The ability of honeypots to characterise attacks in great detail while also providing the ability to capture events requiring asynchronous communication (Hunter, et al., 2013) make them a fundamental data sensor for the Lost Packet Warehousing Service.

2.2 Network Telescopes

A network telescope refers to a portion of routed but unallocated Internet Protocol (IP) address space which is not being used for running services (Moore, et al., 2004). Therefore, very little, if any, legitimate network traffic is expected to be captured by a network telescope. There are, however, the following exceptions: (i) backscatter traffic, (ii) traffic created by misconfigured hosts, and (iii) aggressive or potentially hostile traffic (Irwin, 2011; Hunter, et al., 2013). By using a network telescope, it becomes possible to monitor unexpected traffic of events such as various forms of DDoS attacks, as well as the automated propagation of Internet-based worms or viruses (Moore, et al., 2004).

The operation and deployment of a network telescope require thoughtful planning. The simplest form of a network telescope can be constructed using a system with a single network interface card and relatively low processing specifications. The design of a basic network telescope is presented in Figure 1 and in this configuration, the router routes an assigned block of addresses (e.g. 192.168.254.0/24) to the network telescope. However, the most significant design consideration from a hardware perspective is to equip the network telescope with sufficient disk storage for the captured data (Irwin, 2011).

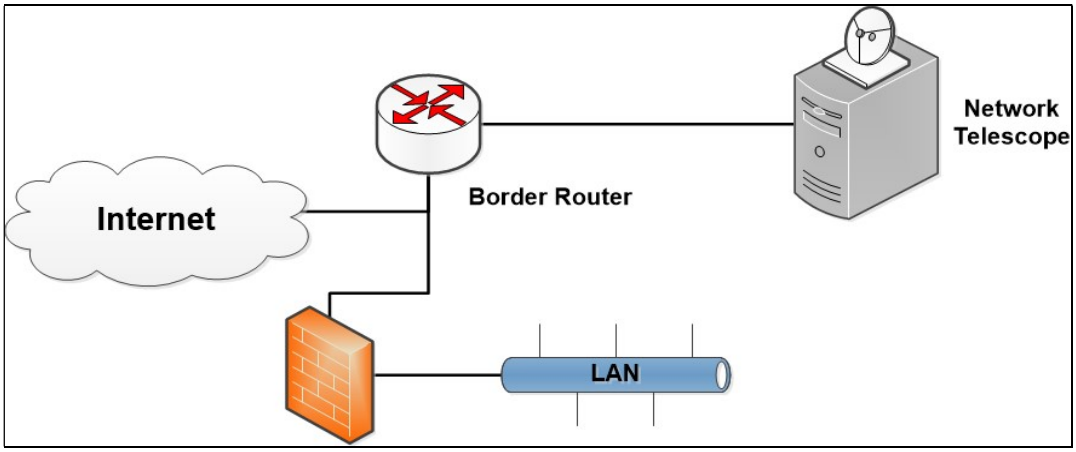

Figure 1: Basic design of a network telescope

Even though the volume of data captured by a network telescope may seem insignificant, the value of the captured data is immeasurable and required for the detection and prevention of cyberattacks. It is, therefore, imperative that a network telescope forms part of the Lost Packet Warehousing Service to ensure adequate coverage of collected cybersecurity data sets.

2.3 NetFlow Collectors

NetFlow is a network protocol developed and patented by Cisco in 1996 and provides the ability to collect IP network traffic as it enters or exits an interface (Hofstede, et al., 2014). Visibility into a network is of great importance to network administrators and operators and, therefore, NetFlow has become a widely used monitoring tool to evaluate network behaviour. NetFlow records the flows (e.g., source IP, destination IP, source port, destination port, layer 3 protocol, and the class of service) and their properties (e.g., packet counters, and the flow starting and finish times) (Cisco, 2012). Once the flow finishes, the NetFlow data is exported to a NetFlow collector. A NetFlow collector is an application responsible for receiving, ingesting, pre-processing and storing the NetFlow data.

[4] https://hub.docker.com/r/lyrebird/honeypot-base/

As with the deployment of network telescopes, adequate planning is also required to enable the capturing of NetFlow data. A typical NetFlow monitoring arrangement is illustrated in Figure 2, which utilises a NetFlow-enabled router or NetFlow exporter. Also required is sufficient disk storage to store and make available the pre-processed NetFlow data for further analysis.

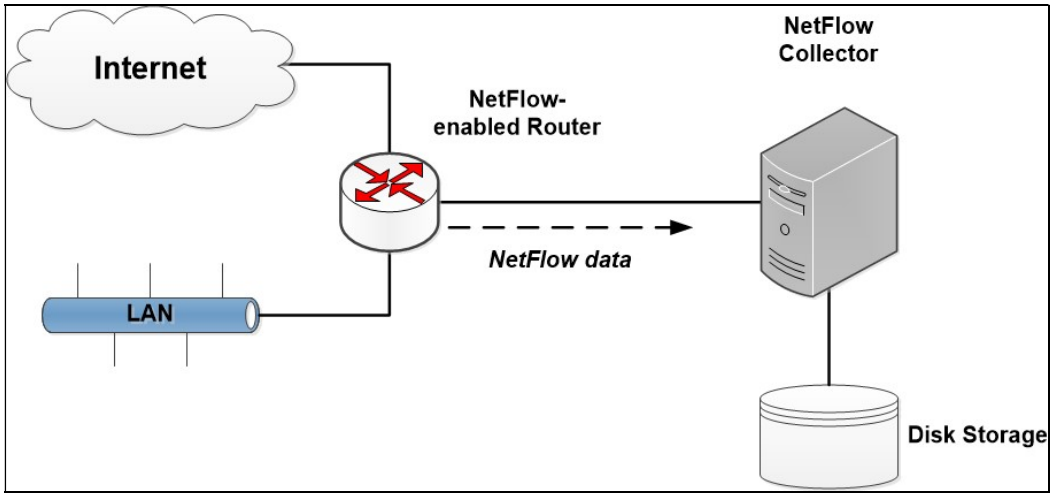

Figure 2: Basic infrastructure required for capturing NetFlow data

Analysis of NetFlow data offers insight into the behaviour of the network, such as monitoring network usage, validating the quality of service, conducting troubleshooting, and performing security and attack detection (Li, et al., 2016). The network awareness provided by NetFlow data emphasises the importance of such data to the Lost Package Warehousing Service.

3. Deployment and Architectural Design

The purpose of the Lost Packet Warehousing Service is to function as the primary source for cybersecurity data collection within SA. The data sensors identified and discussed during the previous section will thus form the foundation of the Lost Packet Warehousing Service. It is, however, important for these data sensors to operate at a national level to enable adequate coverage and collection of cybersecurity data sets. Therefore, the Lost Packet Warehousing Service will be deployed on the National Research and Education Network of South Africa (SA NREN). SA NREN provides the backbone infrastructure to enable network connectivity and services for all tertiary education networks and research councils within SA. The NREN forms part of SA's national integrated cyberinfrastructure, making it is a probable target for cyberattacks (Burke & Herbert, 2020).

The recommended deployment of the Lost Packet Warehousing Service is outlined and illustrated in Figure 3. The architectural design demonstrates the incorporation of the data sensors, as well as the transferal of the collected cybersecurity data sets from the sensors to the data processing and analysis infrastructure.

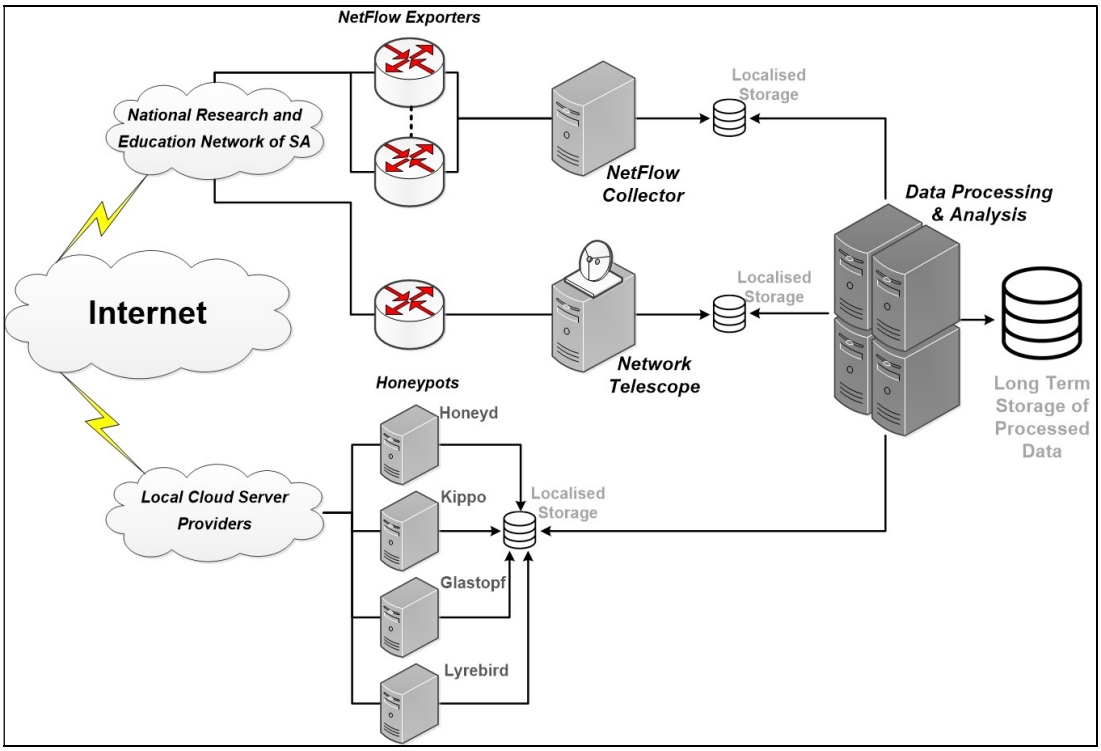

Figure 3: Deployment and architectural design of the Lost Packet Warehousing Service

The NetFlow data sensor relies on NetFlow Exporters, which are NetFlow-enabled border routers forming the backbone of the National Research and Education Network of South Africa. The NetFlow capture daemon (*nfcapd*) running the NetFlow Exporters capture the NetFlow data and store it files using IP Flow Information Export (IPFIX) data format. The files are then transferred to the NetFlow collector, which receives, ingests and conducts pre-processing on the NetFlow data. Once completed, the pre-processed NetFlow data is retained in localised storage.

As discussed in Section 2.2, a Network Telescope monitors a portion of routed but unallocated IP address space. The Lost Packet Warehousing Service will require such IP address space to form part of the National Research and Education Network of South Africa. Monitoring the IP address space will require a redirection on the Address Resolution Protocol (ARP) level. When network traffic arrives for a specific IP address, the router broadcasts an ARP request to discover the host. Once the host responds to the ARP request as the owner of the corresponding IP address, the router directs all traffic to this host. To enable this method of traffic direction, a Fake ARP Daemon (*farpd*) is deployed and assigned the IP address space to monitor. The captured traffic is also retained in localised storage.

The final data sensor is the large-scale deployment of honeypots. An assortment of honeypots (see Section 2.1) will be deployed across the Internet. Various honeypots are utilised to ensure collection of cybersecurity data sets that involve various cyberattacks. To support the deployment of the honeypots, the Lost Packet Warehousing Service will rely on local cloud server providers to host the implementation and deployment of the honeypots. Such deployment of the honeypots will ensure wide reach and increased coverage, enabling the collection of the cybersecurity data sets at a national level. The activities captured by all of the honeypots will be retained in localised storage.

Each data sensor is equipped with localised storage, which is responsible to retain the cybersecurity data sets captured by the sensor. The availability of the collected cybersecurity data sets for each data sensors directly depends on the size (disk space) of the localised storage, as well as the rate at which the data sets are captured. Therefore, the Lost Packet Warehousing Service will provide each data sensor with adequate localised storage to ensure a sufficient collection of cybersecurity datasets over time.

Systematic processing and analysis of the captured cybersecurity data sets are highly recommended to ensure continuous insights relating to emerging trends and cyberattacks are produced from the data. The Lost Packet Warehousing Service will rely on the data processing and analysis infrastructure, further discussed in the following section, to retrieve the various data sets collected by the data sensors. The output produced by the data processing and analysis components is captured and retained in long term storage, readily available for further interrogation to identify emerging trends and on-going cyberattacks.

4. Data Processing and Analysis

The deployment and architectural design of the Lost Packet Warehousing Service will enable the continuous collection of relevant and up to date cybersecurity data sets. Within these collected cybersecurity data sets lie the solutions to identify, detect and possibly prevent occurring cyberattacks. As the collected cybersecurity data sets are obtained using various data sensors, appropriate data processing and analysis techniques are applied to obtain insight that will allow for the identification, inferring, detection and prediction of emerging trends and cyberattacks.

4.1 Methodology

Open-source data processing and analysis tools will be used to analyse the cybersecurity data sets and provide visibility to the structure of collected data from the different data sensors. The Lost Packet Warehousing Service will rely on inductive methods, using few preconceptions to allow intelligence to emerge from the collected cybersecurity data sets (Vasquez, 2018). Given the nature of the cybersecurity data sets collected from varying data sensors, as well as the size of the collected data sets, automation will be required.

The Lost Packet Warehousing Service will automate the data processing and analysis aspects to reduce the complexity of the cybersecurity data sets to a manageable scope. The key enabler for the data processing and analysis of the Lost Packet Warehousing Service is, therefore, machine learning (ML), as it represents a collection of algorithms for dissecting the cybersecurity data sets and discovering insight from the available data (Li, Gunes, Bebis, & Springer, 2013). ML harness the processing power of machines to automate the analysis of large but highly complex data sets. Therefore, both hardware and software must be identified to support the implementation of ML. From a hardware perspective, adequate servers must be in place with a high specification to support the collection, processing and storage as per the architectural design presented in Section 3. Appropriate software to process the data and design ML models must also be accommodated. For data processing, software applications, such as Notepad++ for reading text files and Microsoft Excel to handle Comma Separated Values (CSV), are recommended. These applications have significant functions that support the following:

- Normalisation and correlation process.
- Rule creation for inclusion and/or exclusion of data points.
- Data conversion from one format (.txt) to another (.csv).

Suitable ML tools are deployed to drive the design and execution of ML models that will be used to generate usable intelligence from the collected cybersecurity data sets. For the development and design of the ML models, Jupyter Notebooks will be used, which is a flexible platform supporting the Python programming language, as well as various ML libraries (Perkel, 2018). The single integrated development environment will streamline the data processing and analysis aspects of the Lost Packet Warehousing Service.

With the appropriate ML tools in place, the ML models will be finalised. Firstly, unsupervised learning will be used to uncover meaning in the large unstructured cybersecurity data sets by developing models to arrange and structure the data. Two branches of unsupervised learning will be utilised namely Principal Component Analysis (PCA) and Latent Dirichlet Allocation (LDA). PCA is a non-parametric statistical technique that is primarily used for dimensionality reduction (Cao, Chua, Chong, Lee, & Gu, 2003) which can also be used to filter noisy data sets. While LDA will be used for the reduction of features in the pre-processing phase for pattern classification.

Once the data is understood and arranged in a structured way, supervised ML methods will be deployed to further extract intelligence from the cybersecurity data sets. Based on the nature of the collected data, the Lost Packet Warehousing Service will rely on the Decision Tree and Online Support Vector Machine (SVM)

supervised ML algorithms. Decision Trees use inductive inference to produce a tree-like structure which includes a root node, internal nodes, and leaf nodes. Classification is processed from the root node and moving down toward some leaf nodes (Ming, Wenying, & Xu, 2009). The strength of Decision Tree algorithms is to perform fast classification without requiring much computation and providing information on which features are most important for classification. Since the cybersecurity data sets will be collected and processed in real-time, the online SVM model can update the models over time. Online SVM is an extensive version that stores the model for each training and then adds new examples while removing the least relevant examples for the new data set. It saves memory and provides flexibility for scenarios involving stream data (Li, Gunes, Bebis, & Springer, 2013).

Further enhancement or altering of both the unsupervised and supervised models are not recommended and the models can be used unmodified to obtain intelligence from the collected cybersecurity data sets.

4.2 Process flow

Data collected from the varying data sensors is expected to arrive at a central location in different formats and arrangements. This is because there is no standardisation enforced by the data sensors. Therefore, a suitable phased-approach (see Figure 4) must be followed to transfer the collected cybersecurity data sets into meaningful output. The first phase shall export the cybersecurity data sets to a single file such as a JavaScript Object Notation (JSON) or CSV file formats. The next phase will be to discard missing values or data that does not add to the information base needed for analysis. For example, duplicates, usernames, organisation-specific data, etc. will be removed. Normalising and linearization of the cybersecurity data sets will provide high performance in terms of memory and processing. This also involves the process of imputation, where some values are replaced with the mean or some other relevant measure in that row. Once the data is normalised, data correlation can take place. Varying source origins will arrange data in unique ways that do not match with other sources. In this phase, an appropriate feature that appears on all the data source records will then be used to correlate the data. This will make the data set manageable for deeper analysis. Simulation and visualisation are achieved by using Jupyter Notebook. Intelligence from the collected data will be simulated and visualised using Jupyter Notebook and the Pandas Python library. The Pandas library already contains ML models and thus the data can be applied directly from this tool. Pandas contain tools for data pre-processing, classification, clustering, association rules and visualisation. Both Jupyter Notebook and Pandas are free and open-source products under the GNU General Public License.

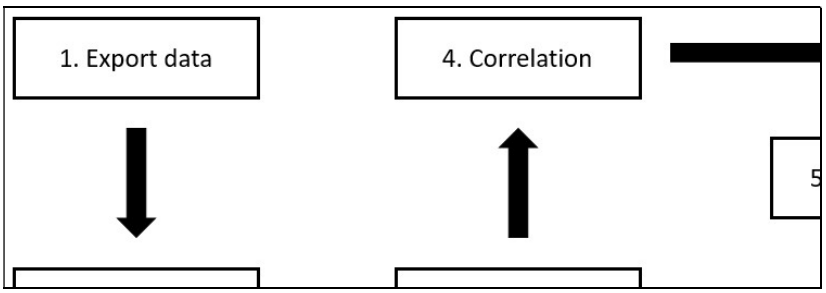

Figure 4: Data analysis process flow

4.3 Output from analysis and processing

The output produced by data processing and analysis will be represented using appropriate visualisation techniques, such as graphs and charts. The produced results will offer insight into the identification, inferring, detection and prediction of emerging trends and cyberattacks. Furthermore, the results will point to a historical period when the network was configured in a certain way. If those network configuration settings are captured during the data collection phase, then that configuration could be tied to the security vulnerabilities that are recorded for that period. Given the stream of data captured, time series analysis could be deployed to investigate the relationship of events from time to time. The time-series data is useful not only for historical analysis but also for future predictions and network behaviour analysis. The data collected can shed light immerging cyber threats and, therefore, be used to harden the network configuration settings.

5. Maintaining Security and Privacy

In SA legislation requires that compliance to Protection of Personal Information be observed as stipulated in the respective acts. Similar to the General Data Protection Regulation (GDPR), which stipulates that the processing of personal information must be legally justified, SA adheres to the Protection of Personal Information Act (PoPIA). Any data collected by the Lost Packet Warehousing Service falls under that PoPIA umbrella. This is because data such as IP, MAC addresses, workstation number etc. can be used to identify a specific user. In the event of a data breach or a loss of data by the Lost Packet Warehousing Service, legal action can be taken by the data owner under Act No 4 of 2013 Part 21(2) if consent is not given. POPI Act 4 of 2013 (10) provides guidelines about the use of personal data that are as follows:

- **Lawfulness and transparency**: data must be processed lawfully, fairly and transparently concerning individuals (POPI Act 4 of 2013 (8) (9)).
- **Purpose limitation**: personal data shall be collected for specified, explicit and legitimate purposes and not further processed in a manner that is incompatible with those purposes.
- **Data minimisation**: personal data shall be adequate, relevant and limited to what is necessary (POPI Act 4 of 2013 (10)).
- **Accuracy**: personal data shall be accurate and, where necessary, kept up to date.
- **Storage limitation**: personal data shall be kept in a form which permits identification of data subjects for no longer than is necessary.
- **Integrity and confidentiality**: personal data shall be processed in a manner that ensures appropriate security of personal data.

The guidelines discussed above are also part of the Cybercrimes and Security Bill in chapter two. The SA Cybercrimes and Security Bill is not yet in effect, however, some of the legislation is covered in broad terms in the PoPI Act of 2013. Since the cybersecurity data sets gathered by the Lost Packet Warehousing Service comes from consenting sources and will be used initially as research material with no personally identifiable data from data sources, there is no risk of violating the SA legislation.

6. Conclusion

The paper aimed to introduce the Lost Packet warehousing Service, a technological solution that will function as the primary source for cybersecurity data collection within South Africa. The Lost Packet Warehousing Service will rely on the deployment of a wide variety of data sensors to collect cybersecurity data sets, which are processed and analysed to offer insight into emerging security trends from a South African perspective. Therefore, the envisioned purpose of the Lost Packet Warehousing Service is to defend South African organisations against cyberattacks. Currently, the Lost Packet Warehousing Service is being deployed and implemented by a research team at the Council for Scientific and Industrial Research (CSIR). Future work will continue to focus on the expansion of the Lost Packet Warehousing Service by incorporating additional or extending the existing data sensors. Furthermore, the data processing and analysis techniques utilised will be enhanced to improve the identification, detection and/or prediction of trends and cyberattacks.

References

Burke, I. D., & Herbert, A. (2020). Tracking Botnets on Nation Research and Education Network. *Proceedings of the 19th European Conference on Cyber Warfare and Security*, (pp. 61-71). Chester.

Cao, L., Chua, K., Chong, W., Lee, H., & Gu, Q. (2003). A comparison of PCA, KPCA and ICA for dimensionality reduction in support vector machine. *Neurocomputing* (pp. 321 - 336). Elsevier.

Cisco. (2012). *Introduction to Cisco IOS.* San Jose: Cisco.

Hofstede, R., Čeleda, P., Trammell, B., Drago, I., Sadre, R., Sperotto, A., & Pras, A. (2014). Flow Monitoring Explained: From Packet Capture to Data Analysis with NetFlow and IPFIX. *IEEE Communications Surveys & Tutorials, 16*(4), 2037-2064.

Hunter, S. O., & Irwin, B. (2011). Tartarus: A honeypot based malware tracking and mitigation framework. (pp. 1-8). Johannesburg: Information Security for South Africa.

Hunter, S. O., Irwin, B., & Stalmans, E. (2013). Real-time Distributed Malicious Traffic Monitoring for Honeypots and Network Telescopes. (pp. 1-9). Johannesburg: Information Security for South Africa.

Irwin, B. (2011). *A framework for the Application of Network Telescope Sensors in a Global IP Network.* Grahamstown, South Africa: Rhodes University.

Li, B., Gunes, M. H., Bebis, G., & Springer, J. (2013). A Supervised Machine Learning Approach to Classify Host Roles On Line Using sFlow. *Proceedings of the first edition workshop on High performance and programmable networking* (pp. 53-60). HPPN '13.

Li, Y., Miao, R., Kim, C., & Yu, M. (2016). FlowRadar: A Better NetFlow for Data Centers. Santa Clara: Proceedings of the 13th USENIX Symposium on Networked Systems Design and Implementation.

Mayfield, K. P., Petty, M. D., Bland, J. A., & Whitaker, T. S. (2018). Composition of cyberattack models. *Proceedings of the 31st International Conference on Computer Applications in Industry and Engineering*, (pp. 3-8). New Orleans.

Ming, H., Wenying, N., & Xu, L. (2009). An improved Decision Tree classification algorithm based on ID3 and the application in score analysis. *Chinese Control and Decision Conference.* Guilin, China: IEEE.

Moore, D., Shannon, C., Voelker, G. M., & Savage, S. (2004). *Network Telescope: Technical Report.* San Diego: Department of Computer Science and Engineering, University of California.

Moyo, A. (2019, October 25). *Bad day for SA's cyber security as banks suffer DDoS attacks*. Retrieved September 18, 2020, from ITweb: https://www.itweb.co.za/content/LPp6V7r4OVzqDKQz

Moyo, A. (2020, August 19). *Experian hacked, 24m personal details of South Africans exposed*. Retrieved September 18, 2020, from ITweb: https://www.itweb.co.za/content/rxP3jqBmNzpMA2ye

Mungadze, S. (2020a, June 9). *Life Healthcare Group hit by cyber attack amind COVID-19*. Retrieved September 18, 2020, from ITweb: https://www.itweb.co.za/content/JBwErvnBK4av6Db2

Mungadze, S. (2020b, June 12). *Security gurus weigh in on SA's latest cyber attacks*. Retrieved September 18, 2020, from ITweb: https://www.itweb.co.za/content/WnpNgM2KPz5qVrGd

Owens, W. A., Dam, K. W., & Lin, H. S. (2009). *Technology, Policy, Law, and Ethics Regarding U.S. Acquisition and Use of Cyberattack Capabilities.* Washington, USA: National Academy of Science.

Perkel, J. M. (2018). Why Jupyter is data scientists' computational notebook of choice. *Nature Publishing Group*.

Provos, N. (2004). A virtual honeypot framework. (pp. 1-14). Berkeley: Proceedings of the 13th conference on USENIX Security Symposium.

Sarkar, I. H., Kayes, A., Badsha, S., Alqahtani, H., Watters, P., & Ng, A. (2020). Cybersecurity data science: an overview from machine learning perspective. *Journal of Big Data, 7*(41), 1-29.

Vasquez, V. E. (2018). *Classification of logs using Machine Learning Technique.* Trondheim, Norway: Norwegian University of Science and Technology.