Privacy Preserving Network Security Data Analytics: Architectures and System Design

Mark E. DeYoung[*†], Philip Kobezak[*†], David Raymond[†], Randy Marchany[*†], Joseph Tront[*]
[*]Bradley Department of Electrical and Computer Engineering, Virginia Tech, Blacksburg, VA
[†]Information Technology Security Lab, Virginia Tech, Blacksburg, VA
{mark.e.deyoung, pdk, raymondd, marchany, jgtront}@vt.edu

Abstract

An incessant rhythm of data breaches, data leaks, and privacy exposure highlights the need to improve control over potentially sensitive data. History has shown that neither public nor private sector organizations are immune. Lax data handling, incidental leakage, and adversarial breaches are all contributing factors. Prudent organizations should consider the sensitive nature of network security data. Logged events often contain data elements that are directly correlated with sensitive information about people and their activities -- often at the same level of detail as sensor data. Our intent is to produce a database which holds network security data representative of people's interaction with the network mid-points and end-points without the problems of identifiability. In this paper we discuss architectures and propose a system design that supports a risk based approach to privacy preserving data publication of network security data that enables network security data analytics research.

1. Introduction

Data-driven network security and information security efforts have decades long history. A deluge of logged events from network mid-points and end-points coupled with unprecedented temporal depth in data retention are driving an emerging market for automated Artificial Intelligence inspired cognitive security products. While network security data yields the most insight when it is aggregated from multiple vantage points in a network, aggregation is not without risk.

Additionally, there are broadly accepted ethical standards regarding collection and use of data about people and their behavior. A central ethical tenant is that data that can identify individuals should only be collected with the subject's voluntary consent.

Network security data often captures user's behaviors with surprising detail. As noted by Wright et.al., aggregated transactional and association data takes on many of the same properties as sensor data[1]. This means aggregated network security data has many of the same privacy concerns presented by sensor data -- it has the potential to reveal life-patterns that identify individuals. While this is not generally a concern when the data is used for correct and secure operation of a network, it must be addressed if we intend to produce generalizable knowledge. In addition to information security threats we must consider threats to privacy. Even when the information security threat model for a network allows trusted system administrators access to aggregated security data for security-relevant purposes it has the potential for privacy-invasive misuse by the 'honest-but-curious'. Additionally, aggregated data that provides 'sensor data' level observations could be of great interest to adversarial or malicious parties.

Because network security data contains data elements that identify individuals and the end-point devices they interact with we cannot ethically produce generalizable knowledge unless we implement procedural and technical controls that reduce the risk of exposing confidential information. We must consider means to reduce privacy vulnerabilities and provide appropriate countermeasures. Judicious implementation of control mechanisms should reduce privacy risks when network security data is used to produce generalizable knowledge.

Privacy and anonymization for data sets are encoded into United States (US) and European Union (EU) law. In the US much of this extends from legally mandated privacy requirements for census data where early privacy work focused on using statistical disclosure controls to limit

Cyber physical systems and smart infrastructures further enforce the need to address privacy concerns in systems design. The end-points and mid-points participating in a network produce detailed logs of human interaction to help administrators operate and secure the services they provide. Sensor level data (i.e.

URI: http://hdl.handle.net/10125/50458
ISBN: 978-0-9981331-1-9

logged transaction and interaction data) is often essential to a systems operation but can reveal people's location, presence, patterns of resource usage, and mobility traces[1]. Sensor level data, like logged events collected for network security purposes, is particularly concerning because it resists naive labeling approaches (i.e. data is labeled for specific uses). People interacting with the network end-point and mid-points are frequently unaware their transactions and interactions are recorded.

2. Network security data analytics

Network Security Data Analytics (NSDA) is the application of data science approaches to the problem domain of network security. Data science can be viewed as a collaboration between experts in a specific problem domain, statisticians and computer scientists with the additional requirement of analytic infrastructure supporting big data techniques. The specific problem domain in this case is network security. The relationship of these disciplines with the problem domain and solution domain are shown in Figure 1.

NSDA draws Artificial Intelligence (AI) inspired techniques such as Data Mining (DM), Machine Learning (ML), and Natural Language Processing (NLP) from the Computer Science (CS) and Statistics disciplines. It also requires an underlying Information Technology (IT) infrastructure capable of 'Big Data' techniques. In section 0 we overview our operational

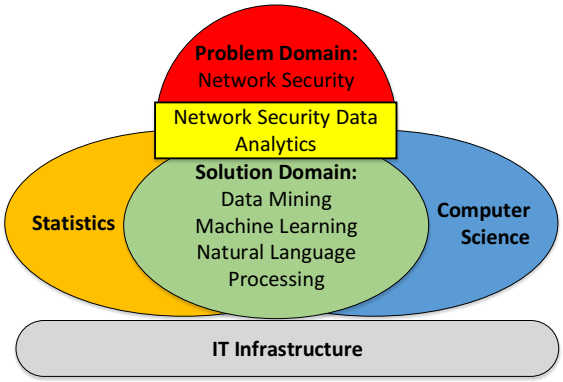

Figure 1: Network security data analytics

model and existing NSDA system design. In section 0 we discuss the threat model for our privacy concerns and in section 0 the data model and data elements with explicit privacy concerns.

2.1. Operational model

Here we frame the problem domain in the context of operational requirements and considerations in the operational model shown in Figure 2. The operational model is a five-phase intelligence cycle. In this paper we only cover the collection, processing and analysis phases which are implemented in our current system.

Collection is the gathering of raw data from the operational environment. For the purposes of NSDA this is the timestamped series of semi-structured data

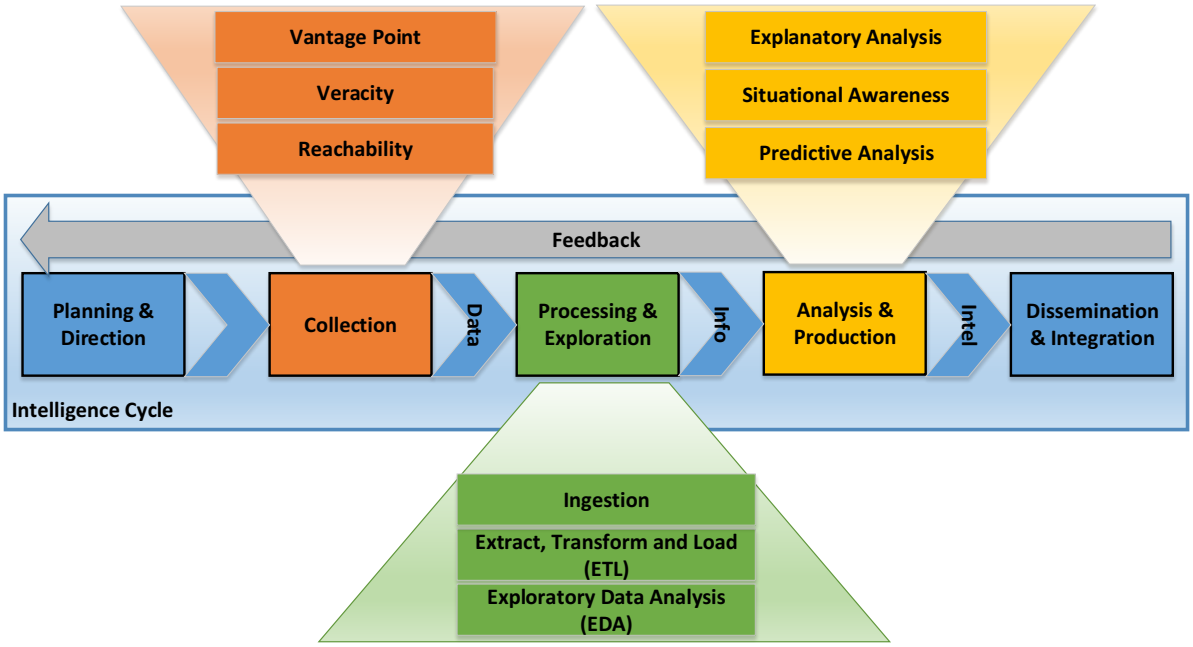

Figure 2: Network security data analytics

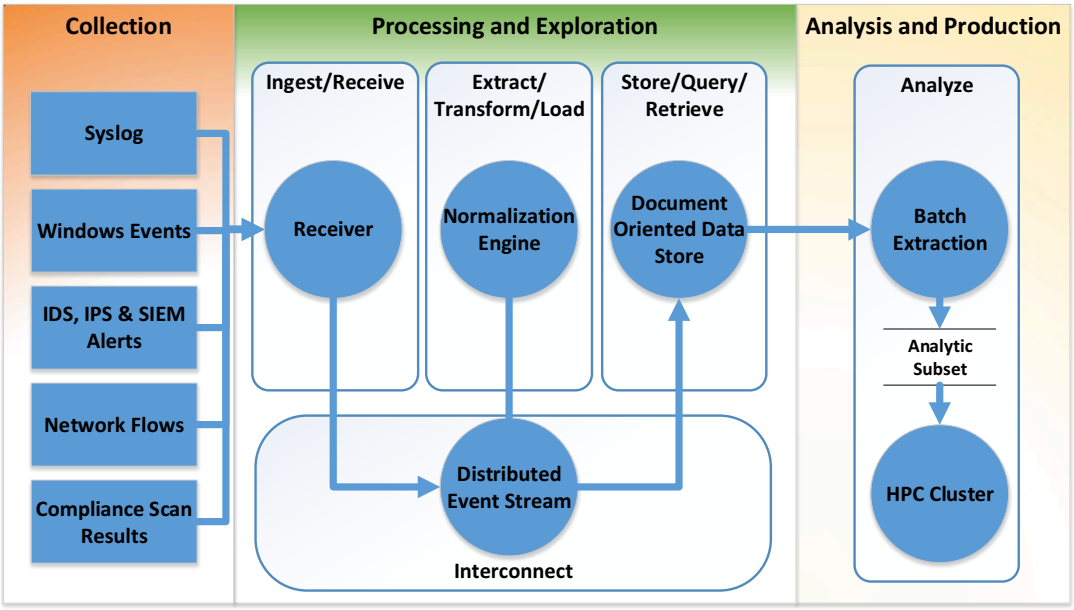

Figure 3: Log aggregation and analysis (LAA) system design

produced by mid-point and end-point systems. Several data sources (shown in Figure 3) include:

- End-point system events logged into syslog or Windows event logs
- Alerts from Intrusion Detection Systems (IDS), Intrusion Prevention Systems (IPS), and Security Information and Event Management (SIEM) systems
- Mid-point network flow data
- Administrative compliance scan results

In the collection phase operators must address the vantage point where data is collected (e.g. mid-point at a network boundary or at an end-point resource utilization). Additionally, they must address both veracity of data and reachability of the sensor.

During processing and exploration, data is normalized for preliminary exploratory data analysis. Because log events and logs are produced in a wide variety of formats we need methods to Extract, Transform, and Load (ETL) data. Exploratory Data Analysis (EDA) can range from classical statistical methods to automated processing with data mining techniques. Analyst often overlap ETL and EDA tasks and frequently lump the task set together as "data munging". Some data munging tasks include: renaming variables, data type conversion, recoding data, merging data sets, inputting missing data and handling missing values.

A primary concern in network security data analytics is accurate detection and reporting of anomalous conditions that have a negative impact on systems. Analysis and production has a range of temporal characterizations:

- **Retrospective** explanatory analysis; what happened?
- **Concurrent** situation awareness; what is happening?
- **Prospective** predictive analysis; what will happen?

The initial focus of our work was on aggregating disparate log data for retrospective analysis. To this end we designed and implemented a network security focused log aggregation and analysis (LAA) system shown in Figure 3. The system ingests a subset of the observational data from the operational network, normalizes data elements, and then stores normalized data in a document oriented store. The document store supports interactive data exploration via a web interface and batch extraction for analysis on external systems.

We use High Performance Computing (HPC) clusters to run batch analytic tasks on data. The data is loaded into temporarily instantiated Apache Hadoop clusters (for storage) and Apache Spark (for analysis) similar to the method used in [2]. The Hadoop and Spark clusters are overlaid in the HPC cluster environment as batch jobs. While the existing system is appropriate for network security focused research it does not address privacy requirements. This severely limits the use of observational data drawn from aggregated events. We must establish a threat model that evaluates privacy requirements and establish means to ensure an adequate level of privacy protection that supports analytic utility. The privacy requirements, technical controls and procedural controls must be documented in a research protocol.

2.2. Threat model

Privacy, usability, and security are complementary and overlapping in some methodology. Even so they are arguably distinct disciplines with competing requirements. Network security systems are inherently privacy invasive and have the potential to restrict or impede usability -- even for authorized users. A network will not function without data attributes that can directly identify real world objects such as the people associated with user accounts and device identifiers. Additionally, the ability to produce actionable network security intelligence often requires identifying information for people and the mid-point and end-point systems they interact with. This naturally leads to tension between functional requirements for usability, information security, and privacy. For our system we seek to prevent casual re-identification of users by researchers.

An identifier (ID) has several desirable characteristics such as no-reuse, immutability, one-to-one correspondence with identified entity (e.g. person, place, or thing). A *direct identifier* is a data value that uniquely corresponds to a real world entity within in a defined domain. Additionally, we must consider other data elements that have the potential to re-identify when extrinsic information (or adversarial knowledge) is used to infer identity from other data characteristics. Formally, an *indirect identifier* or Quasi-Identifier (QID), is the minimum set of data elements $QI = < Q_1, Q_2, \ldots, Q_d >$ that can be used to identify individual records when linked with an attacker's knowledge which is extrinsic to the released data set.

Privacy preservation must be considered because the ethical production of generalizable knowledge from data that incorporates human behaviors requires informed voluntary consent. We must at minimum de-identify data records so that there is not one-to-one correspondence between an identifier and a real world person. In the case where informed voluntary consent is not possible and the data is from existing records direct identifiers must be de-identified in accordance with a research protocol. The de-identification for direct identifiers can be as simple as a re-encoding of the values. A key that maps between the direct identifier and its encoding can be maintained by trusted agents (e.g. the principle investigator or senior researchers). Our organizations Internal Review Board (IRB) specifies data elements that must be de-identified. A partial list of data element types that must be de-identified when they correspond to a real world person is shown in Table 1.

Personnel who routinely operate network systems can become familiar with information such as IP numbers, computer host names, and the people who use specific systems. This internalized background knowledge presents the possibility of casual, "over-the-shoulder" re-identification. In our threat model the primary adversary is the researcher. Re-identification of people from de-identified study data must require non-trivial effort. Because much of this research is conducted in concert with network operators we should take steps to prevent re-identification from data elements that are commonly associated with specific people.

Table 1: Direct Identifiers

Data Element Types	Examples
Names	Persons name, Employer's name, Relative's names
Dates	Birthdate, Date of death, Appointments
Addresses	Home or work addresses, Relative's addresses
Account numbers	Telephone numbers, Social Security numbers, Member account numbers
Features	Voiceprints, Fingerprints, Full face photos and comparable images

2.3. Data model

Networked mid-points and end-points produce observational data with data elements in many syntactic formats. Several of the data sources are shown in Figure 3 in the left hand column labeled collection. Below we discuss a subset of data elements available from the operational network which have explicit privacy concerns. Many other data elements are available in the observational data collected for network and systems operations. Some potentially identifying data elements that are fundamental to our research are usernames, media access codes, and internet protocol numbers.

While IRB-specified data elements should not intentionally be collected in observational network security data we consider user account names to be equivalent to a member account number. User account names and e-mail addresses are assigned when the user is enrolled into the administrative domain. While account names and e-mail addresses are not required to uniquely correspond with a real person it is not uncommon for people to select account names similar to their actual names. Regardless, because the account names (and e-mail addresses derived from account names) can be closely associated with specific people we treat them as confidential like member account numbers.

A username, or user account name, is used for authentication and authorization to use end-point and

mid-point systems within a local administrative domain. It is typical for an administrative domain to require a standard for the format of a username [3], [4]. As an example, a personal identifier (PID) could be constrained to a syntactic structure with specific requirements: it must be between 3-8 characters, it must start with a letter (not a number) and must only contain letters and numbers (no spaces or special characters).

Usernames can lead to unintended confidentiality exposure. It is not uncommon for a person to use multiple services from different vendors but use the same username (or very similar username) among different systems, resulting in traceability and linkability between services [5]–[7]. Also, it is not uncommon for usernames to support localized language encodings which can reveal a person's language preferences[8]. Unprotected user names have also been shown to have sex categorization which can reveal gender of the person associated with the username[9].

A Media Access Control (MAC) address, is a unique identifier assigned to a network interface device to enable communication at the data link layer. MAC addresses come in several formats. Both 48-bit and 64-bit are commonly used in contemporary network standards like Ethernet, 802.11 wireless networks and Bluetooth [10], [11]. The IEEE Guidelines on use of organizationally unique identifiers explain the breakdown of the address space between the Organizationally Unique Identifier (OUI) and Network Interface Controller (NIC) specific[12]. While the intent of a MAC is to provide a unique identifier for each NIC we do not consider a MAC to be a strict identifier. One reason is that it is possible to override the hardware MAC in software configurations meaning there can be reuse of a MAC. Other causes of reuse are: network interface hardware can be moved between computers, perhaps for maintenance or repair, and between organizations. Additionally, an end-point device can be shared and used by several people breaking strict one-to-one correspondence. Although we do not consider a MAC address to be a direct identifier we handle it as a quasi-identifier with one field.

Like MAC addresses we treat Internet Protocol (IP) numbers as a quasi-identifier with one field. It is not a direct identifier because it can be reused. It can be re-assigned alternate end-points. Also, globally routable IP numbers are a salable commodity -- ownership of the IP number can change over time. Some roles involved in IPs include: owner, custodian/operator, and end user. Additionally, some IP ranges are used for link-local (non-routable) connectivity. Another complicating factor is that there is not guaranteed one-to-one correspondence between an end-point (host computer) and an IP. A single end-point device can have multiple network interfaces, each assigned an IP number.

IP assignment can be static or dynamic. A statically assigned IP number is a temporary indirect syntactic identifier. We describe it as temporary because its period of stability (in terms of owner, custodian, and user) is typically long term. We consider dynamically assigned IPs ephemeral because it is used for relatively short periods of time. Dynamic IPs can be acquired by an end-point device through several methods: Dynamic Configuration of IPv4 Link-Local Addresses or Automatic Private IP Addressing (APIPA)[13] IPv6 Stateless Address Autoconfiguration (SLAAC)[14], or Dynamic Host Control Protocol (DHCP)[15], [16].

3. Architectures

While there is a large body of relevant research that has produced privacy enhancing methodology a comprehensive survey and comparative analysis is beyond the scope of this work. General approaches to privacy preservation can be described in several dimensions. Here we frame architectural approaches within in the context of an intelligence cycle, which is the bases of our operational model, and informally discuss some of the operational tradeoffs of each approach. We group approaches to privacy preservation by the systems trust boundary: access controlled sharing within a protected enclave, privacy preserving data publication, privacy preserving analysis, and statistical data controls. Another dimension is data state, which is commonly categorized as: in-motion, at-rest, or in-use. Where in-motion is data that in transmission via a communication system; at-rest is data stored as files; and in-use describes data that is actively being processed (i.e. the data is resident in a devices main memory, caches, or registers).

3.1. Enclave

The use of enclaves for isolation is well established in the information security discipline. In the realm of privacy it is the most restrictive form of limited data release. In the enclave based approach data is intentionally shared only with those who are explicitly authorized access to the system. Enclaves provide restricted access to data by imposing conditions on that access[17]. This is unlike the other approaches which restrict the data released by limiting or adjusting data before it is published. We group other methods that require authorization into the enclave category. This includes remote access and data licensing approaches which require a signed non-disclosure agreement.

In the example, shown in Figure 4 below, a single enclave encompasses all phases of the intelligence cycle for the system under consideration. A more granular

approach could use multiple federated enclaves to limit access at specific phases or to control access to subsets of the underlying operational processes and technical infrastructure (i.e., mid-point and end-point devices that comprise the information systems). In practice, it is likely that multiple enclaves will be used to maintain separation of concerns and separation of duties among multiple participants in the intelligence cycle.

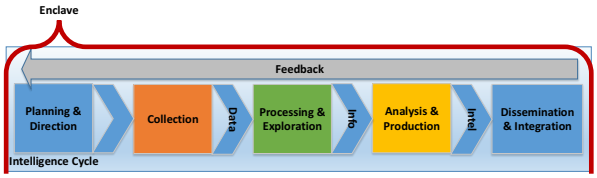

Figure 4: Operational model with enclave

Enclave approaches can be designed to protect data in all three states. Unfortunately, as shown by historical cases, enclaves are not immune to breach by adversarial actors, intentional leakage by insiders, or incidental spillage by failures in process or technical controls. Additionally, without privacy controls, an enclave approach does not prevent privacy-invasive data misuse by the 'honest-but-curious' data snooper – which is the main threat to privacy in our threat model.

3.2. Privacy preserving data publication (PPDP)

Privacy preserving data publication de-couples privacy from analytics. The intent is to produce and publish data sets with adequate privacy protection and sufficient analytic utility. This can be advantageous in situations where the data owner is unable to perform analytics and intends to outsource the effort to a third party.

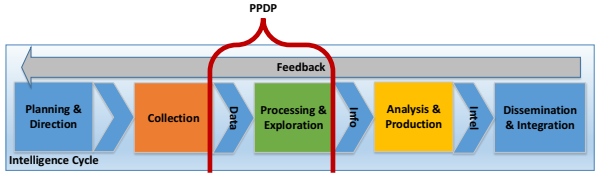

Figure 5: Operational model with PPDP

PPDP approaches have been proposed for a wide range of data sources including: social network data [18], trajectory stream publication[19]–[21], and big data publication[22]. When using a PPDP the impact of incremental data release must be considered when releasing additional data elements. An adversary could collect historical data and leverage the new data elements to derive quasi-identifiers.

3.3. Privacy preserving data analytics (PPDA)

Privacy Preserving Data Analytics (PPDA) use analytic techniques drawn from data mining and machine learning. PPDA is process oriented. The data, analytic procedure, or both must be modified so that the technique is oblivious to identifying data. Because of this PPDA techniques must be incorporated into both the process & exploration and analysis & production phases of the operational model (see: Figure 6).

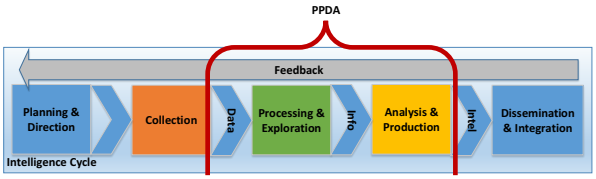

Figure 6: Operational model with PPDA

PPDA models require data holders to modify the original data in such a way that it is still possible to generate analytic results. While the inferred models and parameters could be published, the modified data (potentially randomized with cryptographic techniques) will have little utility. Cryptographic methods requires data owners to execute specifically designed analytic algorithms. Statistics based approaches allow data owners to release sanitized data sets (perturbation or generalization). There are many machine learning and data mining efforts that provide PPDA capabilities. PPDA techniques can potentially provide some protection for data-in-use.

3.4. Statistical Disclosure Controls (SDC)

Statistical Disclosure Controls (SDC), sometimes called Statistical Disclosure Limitation (SDL), methods can be implemented in the final dissemination & integration phases of the operational model as shown in Figure 7.

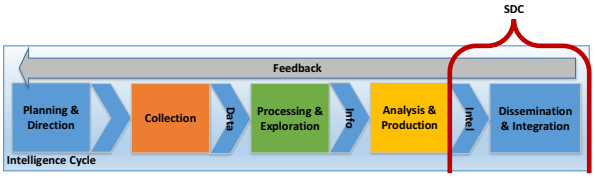

Figure 7: Operational model with SDC

SDC is a venerable approach with a long history. For example, the U.S. Census Bureau has released micro-data for several decades without reported disclosure[23]. SDC is of intense interest to governments which can make evidence based policy

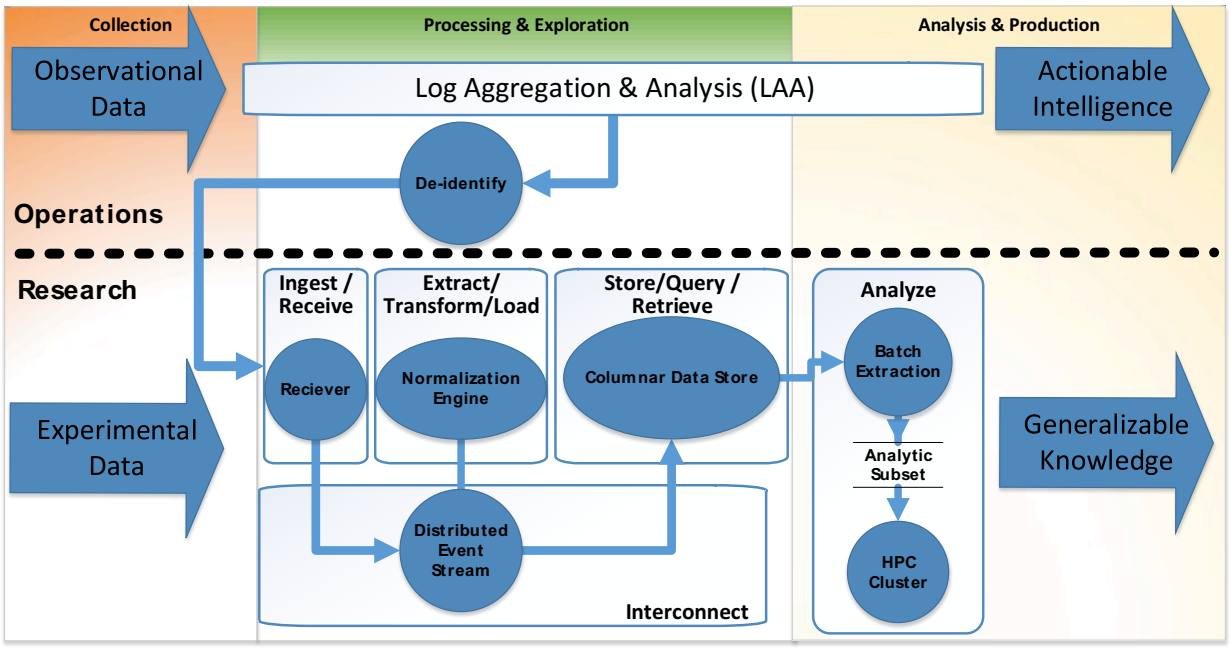

Figure 8: Data analytics system (DAS) design

decisions based on aggregated data like census information[17]. SDC is used to release accurate statistics about a population while preserving individual privacy [24], [25]. SDC is well studied, Abowd et. al. highlight several dozen appropriate methods in their 2001 survey of methods for longitudinal linked data[26]. Many SDC methods modify data values (by randomization). Because this does not preserve the truthfulness of data values, data release might not be useful for evidence based decision making. Instead, tables of summary statistics are typically disseminated.

4. System design

We separate the operations and research into separate enclaves and use a PPDP architecture to publish de-identified data from the operational systems into our research oriented system. Our overall design is shown as a research data flow in Figure 8. Essentially, observational data is extracted from a log aggregation system, the data is de-identified, and then used to develop and evaluate anonymization and generative synthesis techniques. The de-identified data is managed within a research specific enclave, where only those who have executed a non-disclosure agreement and IRB training are allowed access to the data sets. Data within the Data Analytics system (DAS) is de-identified, but not anonymized. This enables researchers with enclave access to develop anonymization and generative synthesis techniques. The techniques developed by enclave researchers are then used to produce sanitized data sets.

One objective of this research is the design and implementation of privacy preserving methods to support the risk based release of network security data so it can be utilized to develop new analytic methods in an ethical manner. Another objective is to preserve analytic utility by maintaining semantic meaning and syntactic structure. Protected data elements should be parsable by standard approaches (i.e. data should be syntactically correct and preserve semantic meaning) so that the techniques are readily brought back into a production environment without excessive process revisions and re-tooling.

Our design is a system-of-systems that reflects the five-phase intelligence cycle previously discussed operational model. Our proof-of-concept system relies on the operational log aggregation system for observational data. We extract batches of tabular data, apply de-identification to direct identifiers and quasi-identifiers, and then store de-identified data in our research-focused Data Analytics System (DAS). We limit our study to extracted columnar formats because we can reliably use automated processes to ensure that tabular data elements are de-identified. This does not eliminate the general problem of regular expressions that are used to parse semi-structured text from event messages but does somewhat constrain the problem space. The use of a distributed event stream as interconnect allows us to substitute sub-systems for scalability purposes. We can leverage substitutable components and vary configurations to tune the system for our expected performance requirements. This is particularly important for the storage subsystem because

we foresee rapid growth as we ingest additional observational data sources.

5. Related work

Here we highlight some other system-oriented privacy preservation work. The synthetic data vault described in [27] uses a PPDP architecture to evaluate the analytic utility of synthetic relational data. The authors generated synthetic data from models learned from public data sets. PRACIS is a cybersecurity information sharing platform using a hybrid of PPDP and PPDA architecture. It is focused on sharing cybersecurity in in Structured Threat Information Expression (STIX) format. The PPDA operations are limited to aggregation of encrypted values and generation of some summary statistics[28]. A proof of concept system focused on unstructured web query logs was implemented by Sedayao but, as noted by the authors, was not used in a production environment[29]. The proposed system is a PPDP architecture.

6. Conclusion

Privacy preserving data publication has the potential to support reproducibility and exploration of new analytic techniques for network security. Providing sanitized data sets de-couples privacy protection efforts from analytic research. De-coupling privacy protections from analytical capabilities enables specialists to tease out the information and knowledge hidden in high dimensional data. While, at the same time, providing some degree of assurance that people's private information is not exposed unnecessarily.

We hypothesize that for some network security use cases that generative synthesis will provide sufficient utility and privacy protection, such that it is possible to make an informed risk decision regarding data release. Information security programs could be enhanced by implementing data minimization practices and misuse prevention by removing identifying information that is not necessary for correct function, security purposes, or production of generalizable research results. The system under design can supply data that is statistically accurate to attributes and behavior of humans without the problems of identifiability.

7. References

[1] R. N. Wright, L. J. Camp, I. Goldberg, R. L. Rivest, and G. Wood, “Privacy tradeoffs: myth or reality?,” in *International Conference on Financial Cryptography*, 2002, pp. 147–151.

[2] M. E. DeYoung, M. Salman, H. Bedi, D. Raymond, and J. G. Tront, “Spark on the ARC: Big Data Analytics Frameworks on HPC Clusters,” in *Proceedings of the Practice and Experience in Advanced Research Computing 2017 on Sustainability, Success and Impact*, New York, NY, USA, 2017, p. 34:1–34:6.

[3] B. Aboba, M. Beadles, J. Arkko, and P. Eronen, “The Network Access Identifier.” IETF Network Working Group, Dec-2005.

[4] K. Zeilenga, “SASLprep: Stringprep Profile for User Names and Passwords,” IETF Network Working Group, RFC4013, Feb. 2005.

[5] B. M. Gross and E. F. Churchill, “Addressing Constraints: Multiple Usernames Task Spillage and Notions of Identity,” in *CHI ’07 Extended Abstracts on Human Factors in Computing Systems*, New York, NY, USA, 2007, pp. 2393–2398.

[6] D. Perito, C. Castelluccia, M. A. Kaafar, and P. Manils, “How Unique and Traceable Are Usernames?,” in *Privacy Enhancing Technologies*, 2011, pp. 1–17.

[7] R. Zafarani and H. Liu, “Connecting Corresponding Identities across Communities,” in *Third International AAAI Conference on Weblogs and Social Media*, 2009.

[8] P. Saint-Andre and A. Melnikov, “Preparation, Enforcement, and Comparison of Internationalized Strings Representing Usernames and Passwords,” RFC Editor, RFC7613, Aug. 2015.

[9] K. M. Cornetto and K. L. Nowak, “Utilizing Usernames for Sex Categorization in Computer-Mediated Communication: Examining Perceptions and Accuracy,” *Cyberpsychol. Behav.*, vol. 9, no. 4, pp. 377–387, Aug. 2006.

[10] IEEE Standards Association, “Guidelines for 48-Bit Global Identifier (EUI-48).” IEEE Standards Association, 12-Jan-2015.

[11] IEEE Standards Association, “Guidelines for 64-Bit Global Identifier (EUI-64).” IEEE Standards Association, 12-Jan-2015.

[12] IEEE Standards Association, “Guidelines for Use of Organizationally Unique Identfiier (OUI) and Company ID (CID).” IEEE Standards Association, 12-Jan-2015.

[13] B. Aboba, E. Guttman, and S. Cheshire, “Dynamic Configuration of IPv4 Link-Local Addresses.” IETF Network Working Group, May-2005.

[14] T. Narten, R. Draves, and S. Krishnan, “Privacy Extensions for Stateless Address Autoconfiguration in IPv6.” IEEE Network Working Group, Sep-2007.

[15] R. Droms, “Dynamic Host Configuration Protocol.” IETF Network Working Group, Mar-1997.

[16] J. Bound, B. Volz, C. E. Perkins, T. Lemon, and M. Carney, “Dynamic Host Configuration Protocol for IPv6 (DHCPv6).” IETF Network Working Group, Jul-2003.

[17] Federal Committee on Statistical Methodology, “Statistical Policy Working Paper 22 (Second version ,2005),” Statistical and Science Policy, Office of Information and Rgulatory Affairs, Office of Management and Budget, 2005.

[18] B. Zhou, J. Pei, and W. Luk, “A brief survey on anonymization techniques for privacy preserving publishing of social network data,” *ACM Sigkdd Explor. Newsl.*, vol. 10, no. 2, pp. 12–22, 2008.

[19] C.-Y. Chow and M. F. Mokbel, “Trajectory Privacy in Location-based Services and Data Publication,” *SIGKDD Explor Newsl*, vol. 13, no. 1, pp. 19–29, Aug. 2011.

[20] R. Chen, B. C. M. Fung, N. Mohammed, B. C. Desai, and K. Wang, “Privacy-preserving trajectory data publishing by local suppression,” *Inf. Sci.*, vol. 231, pp. 83–97, May 2013.

[21] K. Al-Hussaeni, B. C. M. Fung, and W. K. Cheung, “Privacy-preserving trajectory stream publishing,” *Data Knowl. Eng.*, vol. 94, Part A, pp. 89–109, Nov. 2014.

[22] H. Zakerzadeh, C. Aggarwal, and K. Barker, “Privacy-Preserving Big Data Publishing,” presented at the 27th International Conference on Scientific and Statistical Database Management, San Diego, CA, USA, 01-Jul-2015.

[23] J. I. Lane, “The Trade-off Dilemma,” in *Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies*, P. Doyle and J. Lane, Eds. 2001.

[24] C. Dwork, “Differential Privacy,” in *Proceedings of the 33rd International Conference on Automata, Languages and Programming - Volume Part II*, Berlin, Heidelberg, 2006, pp. 1–12.

[25] E. Fayyoumi and B. J. Oommen, “A survey on statistical disclosure control and micro-aggregation techniques for secure statistical databases,” *Softw. Pract. Exp.*, vol. 40, no. 12, pp. 1161–1188, Nov. 2010.

[26] J. M. Abowd and S. D. Woodcock, “Disclosure limitation in longitudinal linked data,” *Confidentiality Discl. Data Access Theory Pract. Appl. Stat. Agencies*, vol. 215277, 2001.

[27] N. Patki, R. Wedge, and K. Veeramachaneni, “The Synthetic Data Vault,” in *Data Science and Advanced Analytics (DSAA), 2016 IEEE International Conference on*, 2016, pp. 399–410.

[28] J. M. de Fuentes, L. González-Manzano, J. Tapiador, and P. Peris-Lopez, “PRACIS: Privacy-preserving and aggregatable cybersecurity information sharing,” *Comput. Secur.*, vol. 69, pp. 127–141, Aug. 2017.

[29] J. Sedayao, R. Bhardwaj, and N. Gorade, “Making Big Data, Privacy, and Anonymization Work Together in the Enterprise: Experiences and Issues,” in *2014 IEEE International Congress on Big Data*, 2014, pp. 601–607.