# Related Work

An insight from data privacy research early in the project led to refactoring, specifically cyber event log records can not be fully anonymized and provide adequate input into the cybersecurity investigation process (DeYoung, 2018; Pang et al., 2006; Qardaji & Li, 2012; Rasic, 2020). The term “anonymized”, as related to data sharing, means original data subjects can not be reassociated with the shared anonymized data. Orthogonally, deidentification does not contain directly identifiable data but acknowledges reassociation through links with external datasets is always a possibility(Garfinkel, 2015; GDPR, 2018; Narayanan et al., 2016). Both terms are used interchangeably throughout the literature, therefore used as search keywords but considered as deidentification during analysis.

## Laboratory Environment Development

CyBOK knowledge area eight(Rashid et al., 2021), *Security Operations and Incident Management* (SOIM), covers many topics including log source descriptions, detection models, threat intelligence and incident response. Thirty-six pages allocated to this knowledge area briefly explain key concepts and items to consider if developing a security operations center but provide little information on how various functions interact within SOIM. In fairness to the author, entire books are written about each concept introduced in the SOIM knowledge area, therefore the CyBOK guide is an overview rather than design for security monitoring. MITRE’s World Class Strategies of CSOC 2nd edition (Knerler et al., 2022) provides excellent foundational guidance to determine functional and governance requirements for an effective SOIM capability but limiting technical guidance to architectural recommendations for defensible cyber event log collection within an organization’s live network, excessive for a lab design.

Building and maintaining organization-wide infrastructure for cyber event log collection, detection and analysis often requires several full-time employees (Knerler et al., 2022; Murdoch, 2018). Therefore, it is not surprising that replicating the computer systems and activities observed in a large organization for testing purposes, often called a cyber range, still represents considerable effort and most are purpose-built with limited scope. (Chouliaras et al., 2021) analyze twenty-five cyber ranges and training or research work they were designed to support, indicating most ranges were associated with military or academic organizations rather than opensource offerings suitable for extension into this project.

Some cyber security testing books provide guidance on lab building (Bodungen et al., 2016; Smith, 2021) to practice attack techniques but *The practice of Network Security Monitoring* (Bejtlich, 2013) is the rare publication providing design guidance event monitoring and collection in a lab environment. Though the lab instructions are technically dated, Bejtlich’s methodology for testing, attacking, and monitoring within the remainder of the book stands the test of time. Like Bejtlich, *The Blue Team Handbook, SOC, SIEM and Threat Hunting Use Cases* (Murdoch, 2018) focuses on detection of cyber attacks and implementation of detection solutions, although the author assumes investigations take place in an existing live environment and does not delve into lab development.

Research did not identify any regulatory standards for cyber ranges but simulating computer systems, network ranges and user activity in a representative environment for the threat scenario was a consistent theme (Furfaro et al., 2017; Somarakis et al., 2020; Urias et al., 2017). To support future research or validation testing, the project’s lab design is detailed in appendix A.

## Privacy Research

The CyBOK privacy and online rights knowledge area (Rashid et al., 2021) considers three privacy design paradigms:

* Confidentiality provided through various technical approaches,
* Control via data handling agreement,
* Transparency in the form of user awareness regarding data collection.

Like SOIM, the privacy knowledge area skims a wide surface area but does include obfuscation-based inference control as a technical confidentiality protection approach. Obfuscation measures appear to be the best path toward this project’s data confidentiality and utility goals, an approach taken by many previous projects sharing network trace data for third party assessment (Mivule & Anderson, 2015; Mohammady et al., 2018; Pang et al., 2006; Xu et al., 2001) and more recent research protecting personally identifiable information (PII) in other types of cyber event logs (DeYoung, 2018; Menges et al., 2021; Rasic, 2020).

(Fung et al., 2010), released a survey of obfuscation-based research focused on data subject privacy preservation while retaining a high degree of utility within published datasets in addition to expanding on many techniques discussed in CyBOK. The survey explains privacy preserving data publishing (PPDP) approaches and effectiveness measurement in-depth. Like many other researchers, Fung et al use examples of medical, census or financial record to illustrate deidentification concepts. Unfortunately, when applying such techniques to cybersecurity event logs the resulting dataset is not suitable for intrusion analysis (DeYoung, 2018; Menges et al., 2021; Rasic, 2020).

In earlier research (Samarati & Sweeney, 1998) recognized reidentification using indirect or quasi-identifiers such as zip code, gender or date of birth remained possible after direct identifiers like name or social insurance number were removed or masked. The authors proposed the concept of K-Anonymity, I.E., reducing precision in certain quasi-identifiers or suppressing unique records until a person can be associated with no less than “*K*” number of people. Generalization measures like changing birth date to year of birth do little to distort other valuable data in a medical record but will not work for cyber event record analysis. Analysts often search event data for patterns such as connections between specific computer hosts or multiple authentication failures for a specific account to identify security issues(Ho et al., 2021; Khan & Parkinson, 2019; Vaarandi et al., 2016), relying on complete records with precise quasi-identifiers, something k-anonymity endeavours to prevent. Further to this point, actual intrusion event volume is typically orders of magnitude lower than benign traffic (Sommer & Paxson, 2010), removing outliers to meet k-anonymity requirements would remove the very evidence analysts should be training to find.

Synthetic contextual data or “dummy addition” approaches to obscure valid records (Fung et al., 2010; Rashid et al., 2021) confound event frequency-based detections and creation of convincingly realistic fake data is not trivial (Chow & Golle, 2009; Toemmel, 2021). Finally, obfuscation via perturbation techniques like random noise or data shuffling (Yancey et al., 2002) prevents accurate clustering analysis that relies on precise and consistent values within event records such as username, event outcome or IP address (Khan & Parkinson, 2019; Wurzenberger et al., 2020).

Protecting IP address privacy within network traces has been ongoing research for decades(Meiss et al., 2007; Xu et al., 2001) with most original solutions focused on replacing IP addresses while retaining routing characteristics within the records. Pang et al., (2006) determined more nuanced fields like TCP timestamps could also be used for reidentifying internal hosts and services, recommending risk analysis on information disclosed rather than dismiss the deidentification solution outright.

(Mohammady et al., 2018) propose a solution for anonymizing network trace data prior to third party analysis, resolving an information disclosure vulnerability in the original CPAN prefix preserving solution(Xu et al., 2001) by generating multiple datasets of which only the data owner can reidentify the original, not considered for the project due to complexity.

More recent synthesis of deidentification concepts and cyber event data can been found in academic contributions using pseudonymization techniques (Varanda et al., 2021a; Zimmer et al., 2020) that include consistently applying pseudonyms across different data sets, a requirement for correlation analysis. (Varanda et al., 2021b) demonstrated the use of custom functions within a log processing pipeline to deidentify data inline, part of the approach taken by this project. (Zimmer et al., 2020) created a framework called PEEPLL to further address potential reidentification attacks using random pseudonym generation while enabling data sharing between different parties. Both pseudonymization solutions alter sensitive field data formats, limiting use with other platforms.

In addition to maintaining global consistency to support investigation, an additional pseudonymization requirement is often retrieving original data values once a problem has been identified (GDPR, 2018; Rasic, 2020; Varanda et al., 2021a), necessitating long term storage of pseudonym mapping in both solutions above. (Chadwick et al., 2020) point out the limited value of third-party analysis of pseudonym data if the true intrusion source cannot be reidentified for response measures. Qardaji & Li, (2012) proposed applying new IP prefix-preserving pseudonyms at periodic intervals, acknowledging that the prefixes must remain consistent within the analysis period. The bucketing algorithm proposed by Qardaji & Li would not support analysis of events occurring over longer time periods since IP information would differ for each flow group.

## Data Source Challenges

A Security Event Log Deidentification (SELD) project success requires replacing sensitive portions of cybersecurity event log records with data providing equivalent detection or analysis capability without confidentiality impacts. Three technical challenges that must be overcome to meet these requirements are:

* Consistent pseudonym application for all log types to facilitate correlation,
* Pseudonymized data resistance to unauthorized reidentification once shared,
* Deidentified fields values must retain original formats to enable ingestion into analysis platforms.

Pseudonymized log fields require each log entry with a specific field value like an IP address or host name be replaced with the same pseudonym to avoid distorting the dataset. Early research primarily focused on network traces(Pang et al., 2006; Xu et al., 2001), but cyber event log collection now includes additional sources like application servers, authentication services and operating systems further complicating global pseudonym consistency. A precise field element like a username must be replaced with the same pseudonym to correlate event sequences across different log types. Varanda et al., (2021) proposed inline processing but transforms field data structure. Only one other inline processing solution, applying transformation functions based on field data type during data ingest was identified in the literature. A modular approach for processing multiple log types, FLAIM, an NCSA research project (Slagell et al., 2006), appears long retired but provided conceptual validation for the SELD approach.

Once deidentified, SELD requires cyber event logs to retain their original format to allow ingesting into an existing security investigation platform and identify intrusion events within that data despite both direct and indirect identifiers being modified. Deidentification solutions that replace sensitive fields with hashed or encrypted values (Menges et al., 2021; Varanda et al., 2021a) limit analysis to a solution specific platform. Data must also appear realistic for effective anomaly detection whether viewed within an analyst investigation tool like a SIEM or in artificial intelligence experiment. (Teoh et al., 2018) used professional security analyst reviews of complete log data to label an AI training set. Recent cyber workforce development research advocates modular approaches to match the training exercises with the environment to be protected and realistic event data, both benign and attack(Somarakis et al., 2020; Urias et al., 2017) .

Time stamp values are another pseudonymization opportunity and a constraint to be observed. Deidentified or not, cybersecurity event logs must maintain the relative time-based relationships between events and specific values within each record for context needed to differentiate benign records from intrusion indicators(Dwyer & Marius Truta, n.d.; Ho et al., 2021; Khan & Parkinson, 2019; Roes, 2017). Timestamp data becomes a sensitive quasi-identifier because external sources such like news, breach reports or visual reconnaissance could provide reidentification linkage or behavioural profiling opportunities (Douriez et al., 2016; Mansour et al., 2021; Rodríguez-Hoyos et al., 2019; Tudor et al., 2015). An acknowledged gap in the current SELD solution, time stamp deidentification would require much more experimentation and validation than the project time frame permitted.

## Managing Privacy and Utility Trade-offs

Cybersecurity event logs commonly take the form of attribute tuples, have predefined structure and data formats, and may contain sensitive data regarding system’s security posture (Khan & Parkinson, 2019). Modern log volume often exceeds millions, and occasionally billions of events per day (IBM, 2021; Microsoft, 2022). This rich data mining environment facilitates system resilience and cybersecurity assurance enabling security professionals to combine their domain expertise with big data techniques (DeYoung, 2018). Unfortunately, larger collections of semi-structured data accessed through analytic engines add reidentification risk due to the volume of indirect identifiers and unique patterns created within this (DeYoung, 2018; Montjoye, De et al., 2013; Narayanan & Felten, E., 2019) data. The “honest but curious” (HBC) external data-handler is an adversarial consideration in privacy related research papers, including maintaining organizational privacy with an outsourced managed security service provider (MSSP) (Rasic, 2020; Zhang et al., 2006). How data can be reanalyzed or combined with other sources in the future to derive new insights is unknowable to the organization sharing the deidentified data (Douriez et al., 2016; Garfinkel, 2015; Narayanan et al., 2016; Sweeney, 1997) therefore risk assessment should precede data sharing.

Traditional prevention measures like high degrees of generalization and aggregation or removing individual data identifiers to create statistical analysis datasets with reduced privacy concern (Golle, 2006) are ineffective approaches for sharing investigable cybersecurity event data. (Chadwick et al., 2020) move the cybersecurity event record privacy conversation toward a continuum decision based on the detail level needed by the data processors and their stakeholders. Acceptable privacy levels are achieved through data sharing agreements and software controls, but the authors acknowledge the solution’s complexity and inability to share between different trust levels. Ohm (2010) summarized this tension a decade earlier as

*"Utility and privacy are, at bottom, two goals at war with one another. In order to be useful, anonymized data must be imperfectly anonymous".*

The seminal work “Broken Promises of Privacy” (Ohm, 2010) analyses data anonymization issues from a legal policy perspective, often referencing the earlier, highly influential “Weaving Technology and Policy Together to Maintain Confidentiality” (Sweeney, 1997). NIST IR 8053 (Garfinkel, 2015) revisits both papers and several other sources further defining deidentification techniques from a procedural practice perspective. Although Garfinkel and Ohm cite the AOL 2006 reidentification report (Barbaro & Zeller, 2006) their protection focus is PII rather than attributing cyber events to a specific person, which until recently was not considered PII(Hintze, 2018).

Garfinkel (2015), points out the European privacy committee Article 29 Working Party do not consider pseudonymized data anonymized but do consider it a security control (Emam, El & Alvarez, 2015), a similar position later reflected in GDPR article 32 (GDPR, 2018). Consequently, organizations contemplating cyber event data sharing, are left with the burden of assessing reidentification risk themselves and knowingly or not, already participate.Technical support issues with IT systems often require sharing sensitive system data and many organizations employ third party data analysis for some cybersecurity services (Gartner, 2022), deidentification is not the norm in either scenario.

Organizations in specific industries may also be participating in government managed/mandated threat monitoring (Canarie, 2021; TSA, 2022) or Information Sharing and Analysis Centers (ISAC). (Chadwick et al., 2020) identify the need for data sharing agreements and differing degrees of deidentification within their framework but for the security leader ultimately accountable for the likelihood and impact of reidentification risks the research is not straightforward. ISO 31000 defines risk as the “effect of uncertainty on objectives”(Tranchard, 2018) and list IT, organizational resilience, crisis management and security among the many risk receptors that management systems must address, therefore contributing to cyber security research may be the best long-term solution.

Policy maker interest in quasi-identifiers leading to the reidentification of individuals seems to coincide with research literature on digital-to-physical reidentification using pseudonymized smart meter and mobile location data (Cleemput et al., 2018; Finster & Baumgart, 2013; Montjoye, De et al., 2013). Despite proving reidentification feasibility, researchers did not report legal or physical consequences, though potentially a future impact as quasi-identifiers gain consideration in privacy regulations.

There is also debate on the practicality of reidentification attacks, proponents draw attention to external data requirements and very low success rates when the deidentified datasets are analyzed within designated boundaries (Cavoukian & Castro, 2014) and the accuracy of identification claims(Barbaro & Zeller, 2006). Opponents point out the fallacy of predicting all data sources, current and future, an attacker could have access to perform reidentification attacks (Narayanan & Felten, E., 2019; Narayanan et al., 2016). Although Naraynan et al dismiss ad hoc deidentification as inadequate privacy protection, pseudonymization within security products is not widely commercially available today. Assessing GDPR compliant data processing viability with an experimental SEIM, (Menges et al., 2021) acknowledge future work needed to address certain incident identification gaps their research uncovered.

Organizations have controls beyond deidentification such as limiting the volume of data shared, vetted recipients versus public release (Narayanan et al., 2016) and periodic pseudonym rotation (Qardaji & Li, 2012). Additionally, reidentification within cybersecurity event logs will typically be limited to identifiers like IP addresses, event time or computer account names, data like passwords are typically not logged. Research has long shown service and system identification is possible without verbatim reidentification (Coull et al., 2007; Khan & Parkinson, 2019; Murakami, 2019). While identification of topology or vulnerable hosts is less than ideal, the adversary must gain access to the environment then bypass other security controls before gaining full access (Kordy et al., 2014). When control measures, potential impacts and benefits are considered, certain cyber event log sharing scenarios may be within an organization’s risk tolerance allowing consideration of this projects deidentification program.

# References

Barbaro, M. & Zeller, T. (2006) A Face Is Exposed for AOL Searcher No. 4417749. *New York Time* [Online]. Available from: <http://www.nytimes.com/2006/08/09/technology/09aol.html?ei=5090&e...> [Accessed 3 December 2022].

Bejtlich, R. (2013) *The Practice of Network Security Monitoring*. 1st ed. San Francisco: No Starch Press.

Bodungen, C., Singer, B., Shbeeb, A., Wilhoit, K. & Hilt, S. (2016) *Hacking Exposed Industrial Control Systems: ICS and SCADA Security Secrets & Solutions*. 1st ed. McGraw-Hill Education 2016.

Canarie (2021) *New Funding to Deliver CanSSOC Cybersecurity Threat Intelligence to Canada’s Research and Education Sector - CANARIE* [Online]. Available from: <https://www.canarie.ca/new-funding-to-deliver-canssoc-cybersecurity-threat-intelligence-to-canadas-research-and-education-sector/> [Accessed 2 December 2022].

Cavoukian, A. & Castro, D. (2014) *Big Data and Innovation, Setting the Record Straight: De-Identification Does Work* [Online]. Available from: <https://www2.itif.org/2014-big-data-deidentification.pdf> [Accessed 2 December 2022].

Chadwick, D. W., Fan, W., Costantino, G., Lemos, R. de, Cerbo, F. Di, Herwono, I., Manea, M., Mori, P., Sajjad, A. & Wang, X. S. (2020) A Cloud-Edge Based Data Security Architecture for Sharing and Analysing Cyber Threat Information. *Future Generation Computer Systems* [Online], 102 January, pp. 710–722. Available from: <https: //www.sciencedirect.com/science/article/pii/S0167739X19300895> [Accessed 18 November 2022].

Chouliaras, N., Kittes, G., Kantzavelou, I., Maglaras, L., Pantziou, G. & Ferrag, M. A. (2021) Cyber Ranges and TestBeds for Education, Training, and Research. *Applied Sciences* [Online], 11, p. 1809. Available from: <https://doi.org/10.3390/app11041809> [Accessed 5 November 2022].

Chow, R. & Golle, P. (2009) Faking Contextual Data for Fun, Profit, and Privacy [Online]. In: *Proceedings of the ACM Conference on Computer and Communications Security*, *2009*. pp. 105–108. Available from: <https://dl.acm.org/doi/10.1145/1655188.1655204> [Accessed 27 November 2022].

Cleemput, S., Mustafa, M. A., Marin, E. & Preneel, B. (2018) De-Pseudonymization of Smart Metering Data: Analysis and Countermeasures. *2018 Global Internet of Things Summit, GIoTS 2018*, November.

Coull, S. E., Wright, C. V, Monrose, F., Collins, M. P. & Reiter, M. K. (2007) Playing Devil’s Advocate: Inferring Sensitive Information from Anonymized Network Traces. *NDSS* [Online], 7, pp. 35–47. Available from: <https://www.ndss-symposium.org/ndss2007/playing-devils-advocate-inferring-sensitive-information-anonymized-network-traces/> [Accessed 28 November 2022].

DeYoung, M. (2018) *Privacy Preserving Network Security Data Analytics*. Virginia Tech.

Douriez, M., Doraiswamy, H., Freire, J. & Silva, C. T. (2016) Anonymizing NYC Taxi Data: Does It Matter? *Proceedings - 3rd IEEE International Conference on Data Science and Advanced Analytics, DSAA 2016* [Online], December, pp. 140–148. Available from: <https://ieeexplore.ieee.org/document/7796899> [Accessed 28 November 2022].

Dwyer, J. & Marius Truta, T. (n.d.) *Finding Anomalies in Windows Event Logs Using Standard Deviation* [Online]. Available from: <https://ieeexplore.ieee.org/document/6680025> [Accessed 22 April 2023].

Emam, K. El & Alvarez, C. (2015) A Critical Appraisal of the Article 29 Working Party Opinion 05/2014 on Data Anonymization Techniques. *International Data Privacy Law* [Online], 5 (1) February, pp. 73–87. Available from: <https://academic.oup.com/idpl/article/5/1/73/2863828> [Accessed 18 November 2022].

Finster, S. & Baumgart, I. (2013) Pseudonymous Smart Metering without a Trusted Third Party. *Proceedings - 12th IEEE International Conference on Trust, Security and Privacy in Computing and Communications, TrustCom 2013*, pp. 1723–1728.

Fung, B. C. M., Wang, K., Chen, R. & Yu, P. S. (2010) Privacy-Preserving Data Publishing. *ACM Computing Surveys (CSUR)* [Online], 42 (4) June, pp. 1–53. Available from: <https://dl.acm.org/doi/10.1145/1749603.1749605> [Accessed 19 November 2022].

Furfaro, A., Piccolo, A., Sacca, D. & Parise, A. (2017) A Virtual Environment for the Enactment of Realistic Cyber Security Scenarios. In: *Proceedings of 2016 International Conference on Cloud Computing Technologies and Applications, CloudTech 2016*, *February 8, 2017*. Institute of Electrical and Electronics Engineers Inc., pp. 351–358.

Garfinkel, S. L. (2015) *De-Identification of Personal Information* [Online]. Available from: <http://dx.doi.org/10.6028/NIST.IR.8053> [Accessed 10 November 2022].

Gartner (2022) *Top 3 Strategic Priorities for Security and Risk Management Leaders Leadership Vision for 2022*.

GDPR (2018) *Article 32 : Security of Processing* [Online]. GDPR Resources & Information. Available from: <https://www.gdpr.org/regulation/article-32.html> [Accessed 2 December 2022].

Golle, P. (2006) Revisiting the Uniqueness of Simple Demographics in the US Population [Online]. In: *Proceedings of the ACM Conference on Computer and Communications Security*, *2006*. pp. 77–80. Available from: <https://dl.acm.org/doi/10.1145/1179601.1179615> [Accessed 2 December 2022].

Hintze, M. (2018) Viewing the GDPR through a De-Identification Lens: A Tool for Compliance, Clarification, and Consistency. *International Data Privacy Law* [Online], 8 (1) February, pp. 86–101. Available from: <https://academic.oup.com/idpl/article/8/1/86/4763693> [Accessed 10 November 2022].

Ho, G., Dhiman, M., Akhawe, D., Paxson, V., Savage, S., Voelker, G. M. & Wagner, D. (2021) Hopper: Modeling and Detecting Lateral Movement [Online]. USENIX. Available from: <https://www.usenix.org/conference/usenixsecurity21/presentation/ho>.

IBM (2021) *WinCollect: Let’s Talk About Log Source Event Rates & Tuning Profiles (Updated)* [Online]. IBM Support. Available from: <https://www.ibm.com/support/pages/wincollect-lets-talk-about-log-source-event-rates-tuning-profiles-updated> [Accessed 21 November 2022].

Khan, S. & Parkinson, S. (2019) Discovering and Utilising Expert Knowledge from Security Event Logs. *Journal of Information Security and Applications*, 48 October, p. 102375.

Knerler, K., Parker, I. & Zimmerman, C. (2022) *11 Strategies of a World-Class Cybersecurity Operations Center | MITRE* [Online]. Available from: <https://www.mitre.org/news-insights/publication/11-strategies-world-class-cybersecurity-operations-center> [Accessed 6 November 2022].

Kordy, B., Piètre-Cambacédès, L. & Schweitzer, P. (2014) DAG-Based Attack and Defense Modeling: Don’t Miss the Forest for the Attack Trees. *Computer Science Review*, 13–14 (C) November, pp. 1–38.

Mansour, H. O., Siraj, M. M., Ghaleb, F. A., Saeed, F., Alkhammash, E. H. & Maarof, M. A. (2021) Quasi-Identifier Recognition Algorithm for Privacy Preservation of Cloud Data Based on Risk Reidentification. *Wireless Communications and Mobile Computing*, 2021.

Meiss, M., Menczer, F. & Vespignani, A. (2007) A Framework for Analysis of Anonymized Network Flow Data [Online]. In: *NSF Symposium on Next Generation of Data Mining and Cyber-Enabled Discovery for Innovation*, *2007*. Available from: <https://redirect.cs.umbc.edu/~hillol/NGDM07/abstracts/poster/MMeiss.pdf> [Accessed 31 October 2022].

Menges, F., Latzo, T., Vielberth, M., Sobola, S., Pöhls, H. C., Taubmann, B., Köstler, J., Puchta, A., Freiling, F., Reiser, H. P. & Pernul, G. (2021) Towards GDPR-Compliant Data Processing in Modern SIEM Systems. *Computers & Security*, 103 April, p. 102165.

Microsoft (2022) *Deploy a Log Forwarder to Ingest Syslog and CEF Logs to Microsoft Sentinel | Microsoft Learn* [Online]. Available from: <https://learn.microsoft.com/en-us/azure/sentinel/connect-log-forwarder?tabs=rsyslog> [Accessed 21 November 2022].

Mivule, K. & Anderson, B. (2015) A Study of Usability-Aware Network Trace Anonymization. *Proceedings of the 2015 Science and Information Conference, SAI 2015*, September, pp. 1293–1304.

Mohammady, M., Wang, L., Hong, Y., Louafi, H., Pourzandi, M. & Debbabi, M. (2018) Preserving Both Privacy and Util-Ity in Network Trace Anonymization [Online]. In: *CCS’18: Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*, *October 2018*. ACM, pp. 459–474. Available from: <https://doi.org/10.1145/3243734.3243809> [Accessed 5 November 2022].

Montjoye, Y. A. De, Hidalgo, C. A., Verleysen, M. & Blondel, V. D. (2013) Unique in the Crowd: The Privacy Bounds of Human Mobility. *Scientific Reports* [Online], 3 (1), pp. 1–5. Available from: <www.nature.com/scientificreports> [Accessed 4 November 2022].

Murakami, T. (2019) A Succinct Model for Re-Identification of Mobility Traces Based on Small Training Data. *Proceedings of 2018 International Symposium on Information Theory and Its Applications, ISITA 2018*, March, pp. 164–168.

Murdoch, D. (2018) *Blue Team Handbook: SOC, SIEM, and Threat Hunting Use Cases*. 1st ed. Don Murdoch.

Narayanan, A. & Felten, E. (2019) *No Silver Bullet: De-Identification Still Doesn’t Work* [Online]. Available from: <http://www.randomwalker.info/publications/no-silver-bullet-de-identification.pdf> [Accessed 25 November 2022].

Narayanan, A., Huey, J. & Felten, E. W. (2016) A Precautionary Approach to Big Data Privacy. *Law, Governance and Technology Series* [Online], 24, pp. 357–385. Available from: <https://link.springer.com/chapter/10.1007/978-94-017-7376-8\_13> [Accessed 2 December 2022].

Ohm, P. (2010) BROKEN PROMISES OF PRIVACY: RESPONDING TO THE SURPRISING FAILURE OF ANONYMIZATION. *UCLA Law Review*, 57, pp. 1703–1776.

Pang, R., Allman, M., Paxson, V. & Lee, J. (2006) The Devil and Packet Trace Anonymization. *ACM SIGCOMM Computer Communication Review* [Online], 36 (1), pp. 29–38. Available from: <https://dl.acm.org/doi/10.1145/1111322.1111330> [Accessed 25 November 2022].

Qardaji, W. & Li, N. (2012) Anonymizing Network Traces with Temporal Pseudonym Consistency. *Proceedings - 32nd IEEE International Conference on Distributed Computing Systems Workshops, ICDCSW 2012* [Online], pp. 622–633. Available from: <https://ieeexplore.ieee.org/document/6258214> [Accessed 29 November 2022].

Rashid, A., Chivers, H., Danezis, G., Lupu, E., Martin, A., Rigby, Y. & Hallett, J. (2021) *The Cyber Security Body of Knowledge* [Online]. United Kingdom: National Cyber Security Center. Available from: <https://www.cybok.org/media/downloads/CyBOK\_v1.1.0.pdf> [Accessed 19 August 2022].

Rasic, A. (2020) *Anonymization of Event Logs for Network Security Monitoring*. Montreal: Concordia University.

Rodríguez-Hoyos, A., Estrada-Jiménez, J., Rebollo-Monedero, D., Forné, J., Trapero, R., Atos, B., Romero, A. Á., Diaz, R., Atos, R., Franco, Z., González, J. J. & Canós, J. H. (2019) Anonymizing Cybersecurity Data in Critical Infrastructures: The CIPSEC Approach. *ISCRAM 2019: 16th International Conference on Information Systems for Crisis Response and Management: conference proceedings: Valencia (Spain): May 19-22* [Online], pp. 1198–1208. Available from: <https://upcommons.upc.edu/handle/2117/176093> [Accessed 1 December 2022].

Roes, T. (2017) *Getting Started with Time Series Analysis in Kibana | Elastic Blog* [Online]. Available from: <https://www.elastic.co/blog/timelion-tutorial-from-zero-to-hero> [Accessed 19 August 2022].

Samarati, P. & Sweeney, L. (1998) *Protecting Privacy When Disclosing Information: K-Anonymity and Its Enforcement through Generalization and Suppression* [Online]. Available from: <https://dataprivacylab.org/dataprivacy/projects/kanonymity/paper3.pdf> [Accessed 22 November 2022].

Slagell, A., Lakkaraju, K. & Luo, K. (2006) FLAIM: A Multi-Level Anonymization Framework for Computer and Network Logs. [Online], June. Available from: <http://arxiv.org/abs/cs/0606063> [Accessed 29 November 2022].

Smith, P. (2021) *Pentesting Industrial Control Systems*. 1st ed. Birmingham: Packt Publishing.

Somarakis, I., Smyrlis, M., Fysarakis, K. & Spanoudakis, G. (2020) Model-Driven Cyber Range Training: A Cyber Security Assurance Perspective. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* [Online], 11981 LNCS, pp. 172–184. Available from: <https://link.springer.com/chapter/10.1007/978-3-030-42051-2\_12> [Accessed 27 November 2022].

Sommer, R. & Paxson, V. (2010) Outside the Closed World: On Using Machine Learning For Network Intrusion Detection. *2010 IEEE Symposium on Security and Privacy* [Online]. Available from: <https://ieeexplore.ieee.org/document/5504793> [Accessed 22 April 2023].

Sweeney, L. (1997) Weaving Technology and Policy Together to Maintain Confidentiality. *Journal of Law, Medicine and Ethics* [Online], 25 (2–3) June, pp. 98–110. Available from: <https://journals.sagepub.com/doi/abs/10.1111/j.1748-720x.1997.tb01885.x> [Accessed 4 November 2022].

Teoh, T.-T., Nguwi, Y.-Y., Elovici, Y., Ng, W.-L. & Thiang, S.-Y. (2018) Analyst Intuition Inspired Neural Network Based Cyber Security Anomaly Detection. *International Journal of Innovative Computing*, 14 (1), pp. 379–386.

Toemmel, C. (2021) Catch Me If You GAN: Using Artificial Intelligence for Fake Log Generation. [Online], December. Available from: <http://arxiv.org/abs/2112.12006>.

Tranchard, S. (2018) *ISO - The New ISO 31000 Keeps Risk Management Simple* [Online]. ISO Standards. Available from: <https://www.iso.org/news/ref2263.html> [Accessed 2 December 2022].

TSA (2022) *Memorandum to Covered Pipeline Owner/Operators* [Online]. Available from: <https://www.tsa.gov/sites/default/files/tsa\_sd\_pipeline-2021-02-july-21\_2022.pdf> [Accessed 2 December 2022].

Tudor, V., Almgren, M. & Papatriantafilou, M. (2015) A Study on Data De-Pseudonymization in the Smart Grid. *Proceedings of the 8th European Workshop on System Security, EuroSec 2015* [Online], April. Available from: <https://dl.acm.org/doi/10.1145/2751323.2751325> [Accessed 1 December 2022].

Urias, V. E., Leeuwen, B. Van, Stout, W. M. S. & Lin, H. W. (2017) Dynamic Cybersecurity Training Environments for an Evolving Cyber Workforce. In: *2017 IEEE International Symposium on Technologies for Homeland Security, HST 2017*, *June 7, 2017*. Institute of Electrical and Electronics Engineers Inc., pp. 1–6.

Vaarandi, R., Kont, M. & Pihelgas, M. (2016) Event Log Analysis with the LogCluster Tool [Online]. In: *Proceedings - IEEE Military Communications Conference MILCOM*, *December 22, 2016*. Institute of Electrical and Electronics Engineers Inc., pp. 982–987. Available from: <https://ieeexplore.ieee.org/document/7795458> [Accessed 21 November 2022].

Varanda, A., Santos, L., Costa, R. L. de C., Oliveira, A. & Rabadão, C. (2021a) Log Pseudonymization: Privacy Maintenance in Practice. *Journal of Information Security and Applications*, 63 December.

Varanda, A., Santos, L., Costa, R. L. de C., Oliveira, A. & Rabadão, C. (2021b) Log Pseudonymization: Privacy Maintenance in Practice. *Journal of Information Security and Applications* [Online], 63 December, p. 103021. Available from: <https://www.sciencedirect.com/science/article/pii/S2214212621001915> [Accessed 15 April 2023].

Wurzenberger, M., Höld, G., Landauer, M., Skopik, F. & Kastner, W. (2020) Creating Character-Based Templates for Log Data to Enable Security Event Classification [Online]. In: *Proceedings of the 15th ACM Asia Conference on Computer and Communications Security*, *2020*. vol. 12. New York, NY, USA: ACM, pp. 141–152. Available from: <https://doi.org/10.1145/3320269.3384722> [Accessed 18 November 2022].

Xu, J., Fan, J., Ammar, M., Moon, S. B. & Atl, S. (2001) On the Design and Performance of Prefix-Preserving IP Traffic Trace Anonymization 2. IMPLICATIONS OF PREFIX-PRESERVING ANONYMIZATION [Online]. In: *IMW ’01: Proceedings of the 1st ACM SIGCOMM Workshop on Internet measurement*, *2001*. pp. 263–266. Available from: <https://dl.acm.org/doi/10.1145/505202.505234> [Accessed 7 November 2022].

Yancey, W. E., Winkler, W. E. & Creecy, R. H. (2002) Disclosure Risk Assessment in Perturbative Microdata Protection. *Inference control in statistical databases* [Online], pp. 135–152. Available from: <http://www.census.gov/DES> [Accessed 4 November 2022].

Zhang, J., Borisov, N. & Yurcik, W. (2006) Outsourcing Security Analysis with Anonymized Logs [Online]. In: *2006 Securecomm and Workshops*, *2006*. pp. 1–9. Available from: <https://ieeexplore.ieee.org/document/4198837> [Accessed 29 November 2022].

Zimmer, E., Burkert, C., Petersen, T. & Federrath, H. (2020) PEEPLL: Privacy-Enhanced Event Pseudonymisation with Limited Linkability [Online]. In: *Proceedings of the ACM Symposium on Applied Computing*, *March 30, 2020*. Association for Computing Machinery, pp. 1308–1311. Available from: <https://doi.org/10.1145/3341105.3375781> [Accessed 4 November 2022].