A Study on Log Analysis Approaches Using Sandia Dataset

Mir Mehedi A. Pritom*, Chuqin Li†, Bill Chu‡ and Xi Niu§
Department of Software and Information Systems
College of Computing and Informatics
University of North Carolina at Charlotte
Charlotte, NC, USA
Email: *mpritom@uncc.edu, †cli30@uncc.edu, ‡billchu@uncc.edu, §xniu2@uncc.edu

***Abstract*—Modern enterprises collect, process, and analyze security data from various system and network logs. Previous studies show that, handling large security datasets and detecting anomalies from those are key challenges faced by most of todays' enterprises. Unfortunately most security professionals are inexperienced at performing data analysis. In this paper, we study published works analyzing one publicly accessible log dataset (Sandia Dataset) published by Los Alamos National Laboratory. We evaluate their data analysis methodology as well as results and found significant flaws in most analysis methodologies.**

I. Introduction

A lot of research has been published on methods to analyze security log data. Cisco's survey on 200 IT and cybersecurity professionals with knowledge and/or responsibility of network security and analytics at their respective organizations shows around 72% of organizations think network security monitoring has become more difficult and challenging over the past two years because of increase in network traffic, malware volume, and new evasion techniques to evade existing defense systems. There are also challenges like, network blind spots, communication gaps between security and operations teams in large organizations, and timely data collection issues as mentioned by large corporations [1].

Zuech et al. [2] addressed big data technologies as a solution to manage these big heterogeneous security data. They specifically discussed Data Fusion, using Hadoop technologies in Heterogeneous Intrusion Detection Architectures, and current state of Security Information and Event Management (SIEM) systems to minimize security data analysis challenges. Their work also showed how correlating security events across heterogeneous data sources can enhance cyber threat analysis and cyber intelligence.

An empirical study on log analysis [3] discussed the utility of log analysis in various fields. Though there are lots of challenges working with logs because of heterogeneous formats [4], [5], logs are used to understand system behaviors [6], [7], detect faults and flaws in system [8], mark system vulnerabilities [9], understand relationships among network entities [10], modeling and predicting user behaviors in a system [11], [12], finding insider threats [13], and drive simulation of systems [14]. There are various techniques available for log analysis such as, event and host clustering [15], [10], anomaly detection [16], [17], [18], [19], [20], visualization [21], dependency mining [22], and data extraction [23]. According to a SANS Institute survey on threat hunting, existing security data analytics techniques include basic search, filtering, behavior based analysis, visualization and reporting, statistical modeling, machine learning, and Bayesian inference from the logs [24].

Teaching and research on log analysis in universities is challenging due to shortage of real-world log datasets for research studies as most of the existing datasets available (e.g., DARPA, KDD99, University of New Mexico Dataset, IES, etc.) were generated many years ago [2], [25], [26]. However, according to Zuech et al. [2], it is evident that most of the researches are still using those decade old faulty data sets for these kind of log analysis studies. The reason is simply lack of available quality datasets.

In 2015, Los Alamos National Laboratory published multiple log datasets named "Comprehensive, Multi-Source Cyber-Security Events" which is also known as **Sandia Dataset** [27]. This dataset includes separate files for authentication, process logs on various computers, DNS, netflow logs along with validated anomalies detected by their *red Team* [28]. In this paper, we perform a meta study of eight published papers reporting analysis of the Sandia dataset. Our goal is to critically analyze their research methods given known flaws in the Sandia dataset [28]. This paper can be used also for teaching log analysis at universities.

The remainder of this paper is organized as follows. Section II provides an overview of the Sandia Dataset [27] along with previous studies with the dataset published by Los Alamos National Laboratory [28]. In Section III, we present the detailed analysis for each work. Finally, section IV concludes this work with discussion on current research methodologies for log analysis.

II. Overview of Previous Works

A. Dataset Overview

The publicly available dataset Comprehensive, Multi-Source Cyber-Security Events is released by Los Alamos National Laboratory [27]. This data includes 5 different data tables: authentication, process, DNS, flows (also known as *netflow*) and red team. A brief description of the overall dataset is given here [27]. However, here in this section we will provide some

978-1-5090-2991-4/17/$31.00 ©2017 IEEE

TABLE I
BASIC STATISTICS FOR DIFFERENT DATASET

Datasets	No. of events recorded	No. of source computers	No. of destination computers
Authentication	1,051,430,459	16,230	15,895
Process	426,045,096	11,960	N/A
Flows	129,977,412	11,154	8,711
DNS	40,821,591	15,013	13,776
Red team	749	4	301

basic statistics of individual data tables along with high level analysis to visualize some of the existing flaws in the dataset. Figure 1 shows a timeline analysis of different data table while Table I depicts a summary statistics of five datasets on number of events, number of unique source and destination computers.

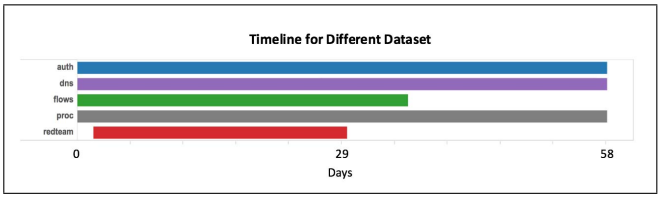

Fig. 1. Timeline of different dataset

The Los Alamos National Laboratory has collected the data from internal network and described their experiences [28]. However, the dataset is not integrated properly over the 58 days period as there are misconfiguration and data collection issues in the network [28], [27]. For the DNS table, only one out of three DNS servers was working properly till day 27 for collecting DNS query records, and the invalid configuration were corrected on day 27 as they mentioned [28]. Figure 2 shows a clear picture of how distribution of DNS queries changes significantly from day 28 after the reconfiguration. Besides, at day 29, a misconfiguration of the internal network routers completely stops the collection of network flow data as reported by [28]. But from our analysis it seems we get netflow data till day 35 and then it stops completely as shown in figure 1 (in green). Additionally, red team data is only available from day 3 to day 29. We also know that collecting data from real time servers and systems is very challenging and data integrity issues may be unavoidable as data collection may never be perfect in real world [29]. Thus, researchers who use these datasets must be vigilant enough to identify flaws and use only valid data in research to get viable results.

B. Summary of Previous Results

Table II shows a summary of our analysis of all the related papers those have analyzed different tables of Sandia Dataset using different techniques to address various research problems.

III. DETAILED CRITICAL ANALYSIS OF PREVIOUS WORKS

In this section we provide a critical survey of eight papers reporting work related to the Sandia Dataset involving different approaches to solve various problems such as, anomaly detection, behavior analysis, and finding relationships among network entities. Rest of this section introduces each paper with research goals, techniques applied, dataset used, methodologies, results, future works along with our analysis.

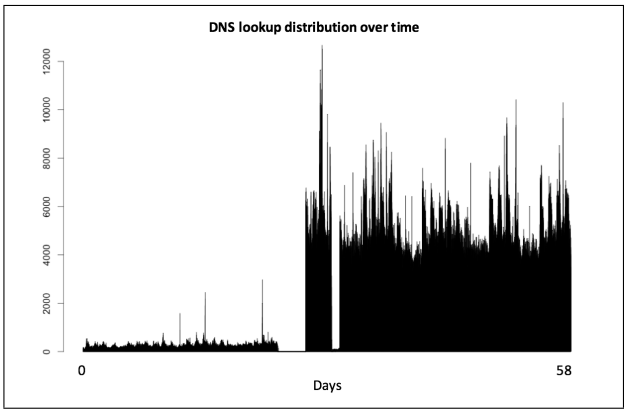

Fig. 2. DNS query distribution over 58 days in LANL's internal network

A. Topic Modeling of Authentication Events in an Enterprise Computer Network [30]

Goal: Detecting misuse of legitimate credentials by monitoring network traffic patterns using topic modeling.

Techniques: In this work they have used Latent Dirichlet allocation (LDA) analysis [32] which is one of the key topic modeling techniques in text mining. They used LDA modeling on computer network traffic data for finding the number of users present in the network.

Datasets: They have used authentication table from Sandia dataset [27].

Methodology: They considered User-Computer (here 'User' is *source computer* and 'Computer' is *destination computer*) connections in authentication table as bipartite graph. They defined each unique 'destination computer' as a separate **word** which is the basic unit of their vocabulary. A day in authentication table is considered as a **document** which is a collection of words. Finally, they have a **corpus** of documents after certain number of days in authentication records. However, one of their key assumption is that they observed all weekend and night time activities as automated traffic. By this assumption, they identified 3 suspect users (e.g., considered as human) with no activity in weekends while they were also inactive in some weekdays. These 3 users sample generated 27 days of data where they have authentication activities with 42 different destination computers as the authors mentioned. These 27 days served as the documents for LDA while the *vocabulary size* is 42. Then they applied LDA model for 3 topics (each 'User' is a topic) to see if they can differentiate different user's activity with this technique.

Results: The authors concluded their result as LDA model successfully clustered documents for the three users into three separate groups of nine documents. They also stated that the

TABLE II
SUMMARY ANALYSIS OF PAPERS WORKED WITH SANDIA DATASET

Paper	Goal	Dataset Used	Techniques	Their Research Findings	Our Analysis
[30]	Detecting misuse by legitimate credentials	Authentication	LDA model	Positive	Input data validation is missing, unrealistic assumptions, may not be applicable in real-world, unvalidated results
[16]	Anomaly detection	Authentication, netflow, red team	Regression based model	Positive	Input data validation is missing, unrealistic assumptions
[10]	Similarity among network entities	Netflow	Clustering analysis	Non-conclusive	May not be scalable in real-world, unvalidated results
[17]	Anomaly detection	Authentication, red team	Probability based model	Positive	Research methodological flaws, input data validation is missing
[18]	Anomaly detection	Netflow, red team	Correlation based statistical model	Positive	Input data validation is missing
[21]	Anomaly detection	Authentication, red team	Visualization	Positive	Input data validation is missing, important results information are missing
[20]	Anomaly detection	Authentication, process, red team	Probabilistic recommender system	Qualitative results	Input data validation is missing, questions in research methodology
[31]	Unusual connectivity in networks	Netflow	Spectral analysis	Positive	Input data validation is missing, unvalidated results

method was able to discover individual users separately as topics from a topic modeling aspect.

Future Directions: In future the author wants to extend the analysis for larger number of users but they mentioned that complexity of the system would increase significantly with the increase of users.

Critical Analysis: They discussed weekdays and weekends activities from authentication events but they have not mentioned how did they differentiate those activities. Given that, the publishers of Sandia Dataset [27] already mentioned "specific timeframe used is not disclosed for security purposes" it is not clear how can we distinguish weekdays activities from weekend activities without differentiating them in the beginning. Again, they assumed all night and weekend activities as automated traffic which may not be true as legitimate users as well as bad guys may use the system in weekends and at night times which will go undetected by their approach. Again, they assumed a correct number of topic is already known which is unrealistic in real world scenario.

B. Activity-based temporal anomaly detection in enterprise-cyber security [16]

Goal: They proposed an anomaly detection technique to distinguish 'abnormal' behaviors from 'normal' ones.

Techniques: The authors proposed a regression based anomaly detection method as a complement to existing signature based techniques and rank anomalies for intensive analysis.

Datasets: They used authentication, netflow and red team tables from the dataset [27]. For authentication data they have considered user-user and user-computer interactions while for flow data they considered each record as a computer-computer interaction.

Methodology: At first they collected seven days of data and created continuous bins for every 5 minutes of data. They took into considerations of weekdays and weekends in their calculations. They measured how anomalous a particular bin is, in terms of deviation from the 'normal' behavior known from historical data. They used first 7 days as training data for the regression based prediction model to compute a anomaly score for each bin in the next three days (used as test data).

Results: The author reported that the bin with second most anomalous score seems to be associated with red team activity in the dataset. The author stated these results as a successful regression based anomaly detection mechanism. They also indicated three more bins linked with red team activity in the top ten list.

Future Directions: The author reported the improvements to the statistical models and creating streaming version for real-time anomaly detections as their future works.

Critical Analysis: They did not consider the whole dataset of 58 days. They only took a subset of first 10 days (7 days training, 3 days as test data) of the dataset without validating. They also did not mention how they differentiated weekdays data from weekends as the data is initially anonymized by LANL [27]. Again, to fit models, they excluded users with only single event logs which may not be true in real world scenarios and help bad guys to evade their detection system.

C. Uncertainty aware clustering for behaviour in enterprise networks [10]

Goal: The author proposed a novel technique for exploring similarity relationships among network entities.

Techniques: To reduce the behaviors of an entity to a small set of statistical parameters the author used statistical models. Then, they applied cluster analysis on challenging heterogeneous network dataset.

Datasets: The author used the netflow dataset [27] in this work. However, they only considered 43 unique edges (pair of *source* and *destination* computers) from the first 21 days of netflow dataset.

Methodology: Their statistical modeling approach to describe an edge will reduce different length time series to a constant length which helped to apply standard clustering analysis of network entity behaviors. They calculated a score for each edge with a Vector Auto-Regression (VAR) model [33] based on duration, byte-load and inter-arrival time of edges. Then they acquired a similarity matrix using the similarity measure for all pairs of edges and tried to cluster edges with ward's method of hierarchical clustering [34]. They also considered the uncertainty corresponding to VAR coefficient estimates in the clustering process by a dissimilarity measure technique defined in [35].

Results: The dataset they worked with contained 52 unique devices (42 source and 11 destination computers). The results of the experiment showed that each set of edges have diverse amounts of traffic with a mean, minimum and maximum of 249, 41, and 636 events respectively. However, in the end the author was non-conclusive about finding malicious activities with these clustering techniques as there are various explanations for normal operation.

Future Directions: As part of future study the author indicates that this clustering method can be applied to devices or other entities, for example, collections of nodes or edges to see the relationship changes instead of considering only edges.

Critical Analysis: The author mentioned about considering a subset(first 21 days) of whole dataset. The reason behind this as they mentioned is significant failure of data collection in later days [27], [28]. However, they only picked a subsample of 43 edges to minimize the complexity of simulations which may create questions for their methods implication on real-world heterogeneous enterprise network data. Finally, their result is non-conclusive and author shows no evaluation to validate their results.

D. Network-wide anomaly detection via the Dirichlet process [17]

Goal: The author proposed a probability-based technique for detecting abnormal connectivity behavior within a computer network.

Techniques: They used statistical probabilistic model for anomaly detection and rank malicious source computers for further investigation.

Datasets: They have used the authentication and red team dataset published by Los Alamos National Laboratory [27].

Methodology: Their hypothesis is that there are many destination computers with small number of source computers connecting to them and unusual connections to those machines are apparently seen as potential threats. Their approach generates directed graph using authentication events. Then used Dirichlet process [36] for modeling P-values of nodes. They used some scoring technique to score anomalousness of nodes and edges. Finally, they used Hadoop mapreduce [37] for minimizing the complexity to handle 58 days authentication events and rank over 16,000 *source computers*.

Results: The result shows that their approach ranked the computer C1769, a red team malicious computer, ranked fifth out of 16,230 with a p-value of approximately 3×10^{-4}. Additionally, the author reported that another known red team source computer, C18025, ranked 95^{th} with p-value 0.006. They pointed these 2 detection as validation of their algorithm.

Critical Analysis: They considered whole 58 days dataset of authentication events. As we can see from figure 1 that the timeline of authentication data is complete. Hence it will not affect their results. However, they have not discussed any input data validation strategies in their research methodology.

E. Correlation-based Streaming Anomaly Detection in Cyber-Security [18]

Goal: Author proposed a procedure to detect anomalies on individual edges of the network graph.

Techniques: They tried to generate graphs from network flow events and then apply statistical model.

Datasets: They have used flow and red team table from the dataset [27]. However, they only considered the first 3 days snapshot into their calculations.

Methodology: They sketched a two stage procedure, named SCAD(Streaming Correlation based Anomaly Detection) to correlate each pair of edges. A computer network is modeled as a directed graph where each edge directed from source to destination computers. Then, they calculated the p-values for each transformed correlation estimate and used those score for calculating anomalousness of an event at a given time.

Results: The author reported that they could not detect any red team labeled anomalies using their method in their 3 days window.

Future Directions: The author mentioned that similar approach can be applied to other dataset such as, authentication dataset.

Critical Analysis: They did not discuss data validation process in their approach as we already know netflow dataset has missing records [28]. Again, in data preparation part, they have mentioned that they considered first 3 days of netflow data and stated these 3 days contain "over 20% of the Netflow events that occur during the 57 day period of the original data set". However, at day 29, a misconfiguration of the internal network routers completely stops the collection of network flow data as discussed in [28]. Moreover, from figure 1, we can see no flow events recorded after day 35 in the dataset. So, this means the author did not validated the timeline of flow dataset which might have affected their research results.

F. Detecting Malicious Logins in Enterprise Networks Using Visualization [21]

Goal: The author proposed a visualization tool, APT-Hunter, by which analysts can observe malicious activities and detect malicious logins.

Techniques: APT-Hunter helps security analysts to apply their knowledge to discover malicious logins inside an enterprise network.

Datasets: They used authentication table for login events and red team table to evaluate their results. They reported that the dataset contains 1.05 billion logins, with about

12,000 users and 18,000 computers on 58 consecutive days of anonymized data. They also partitioned the login events into days and considered only those login events that happen more than 10% of the days to form the model.

Methodology: Advanced Persistent Threats (APT) attacks mostly evade a network by stealing legitimate credentials which is hard to detect by available signature-based anti-virus technology. APT-Hunter is designed to detect this kind of attacks. the tool is basically composed of two components: *login processor & aggregator* and *pattern matcher*. APT-Hunter aims to detect the anomalous login of four types: source change, destination change, user change, and source & destination & user change. By targeting any specific known malicious source computer, an analyst is able to interpret all events from that source computer as malicious using the *pattern matcher* feature.

Results: The author reported that after introducing APT-Hunter and providing a 30 minutes training on the usage of its features, two participants were able to detect 349 out of 749 read team flagged malicious logins using APT-Hunter. The average number of false positives rate observed is 0.005% of all the logins in the dataset.

Future Directions: In future experiments APT-Hunter should involve more participants as the author mentioned. They also indicated to include more features to detect more variants of attacks.

Critical Analysis: Since authentication table is complete, it is fine for them to use the login events from authentication table. However, they have not discussed any data validation before they use and APT-Hunter itself does not have this functionality either. Again, they have not mentioned how much time those two users needed to find those anomalies.

G. Poisson Factorization for Peer-Based Anomaly Detection [20]

Goal: They proposed a probabilistic recommender system for peer-based anomaly detection.

Techniques: The author used a Poisson factorization model based recommendation system [38] to find anomalies from process dataset and validate results with red team table [27].

Datasets: They have used authentication, process, and red team table of Sandia dataset [27].

Methodology: They basically analyzed two features of user behavior. First, the process invoked by the user and second, the computers on which users authenticated. They generated a matrix from number of times a user invoked a process or authenticate a machine. Then, the data was further modeled by a k-dimensional Poisson factorization model [38]. At last, a recommendation system was built to determine if the observed process-machine pairs are considered normal with respect to the model learned over some training period.

Results: The author reported that precision for users who are least active is worse than users with a higher activity level and the precision performance is better for recommending *process* events than *authentication* events. They also stated that they successfully detected red team labeled events with their approach.

Future Directions: The author stated that similar method could also be applied to netflow data and then all the results for each dataset could be combined to determine the malicious network events.

Critical Analysis: In this work author have not validated their data before using. They also have not mentioned which part of the original data to be considered as test and training dataset for modeling. But they have mentioned that their test data lies in the same time-frame when red team exercise was done. From our analysis of Sandia Dataset, we have found that the red team exercise is done within the first 29 days as shown in Figure 1. which may raise an eyebrow on their research methodologies as test data period must be earlier than the training data period.

H. Disassortativity of Computer Networks [31]

Goal: The author proposed techniques to show how Big data analytics for detecting unusual connectivity in an enterprise network can be improved.

Techniques: They have tried to show LANL netflow events [27] as graphs and used spectral analysis to show that connectivity does not imply as similarity in computer networks which they defined as *disassortativity*.

Datasets: They have used the netflow dataset from [27].

Methodology: At first the authors drew network flow events in graphs to show the connection differences among computers observed over one-minute and five minutes. Each graph is undirected, simple graph and could be denoted as a adjacency matrix. Spectral analysis was then applied on these matrices to draw out strong evidence of *disassortative* behavior.

Results: The author reported that by taking *disassortivity* into consideration, strong improvements have been shown in predicting new connections in a network.

Future Directions: They stated that this technology could be combined with other statistical model to predict anomalous connection in any enterprise network.

Critical Analysis: Here again, the author have not validated the dataset before using. However, they only considered a subset of few minutes from flow data so it did not affect their results. They proposed a complementary method to be used with other anomaly detection procedures. Their main motivation here was to show *disassortative* behaviors of network entities in a large network rather than a regular anomaly detection approach.

IV. Discussion and Conclusion

According to above mentioned analysis it is quite clear that most of the existing works with Sandia Dataset are questionable due to their methodological flaws given that there are flaws in the dataset. Most of the papers did not validate their input data before using them in analysis. Only one paper mentioned that they have used a subset of the dataset because of the flaws in dataset while others did not even mention easily detectable flaws in the data. Most papers

used statistical approaches to find anomalies from various logs while visualization and text mining techniques were used by one paper each. Two of those eight papers discussed behaviors among network entities but did not try to detect any anomalies from the dataset. It is quite evident that researchers are eager to find anomalies by analyzing these large logs. However, they are not conscious about their research methodology. Most of them did not consider to validate the quality of the data which is obvious from their research methodology. We pointed out that a timeline analysis of multiple inter-related datasets is a basic need for all security data analysis. We believe this meta study highlights the importance of carefully vetting data before analysis as it is a waste of time to analyze flawed data.

REFERENCES

[1] Jon Oltsik. Network Security Monitoring Trends, August 2016.

[2] Richard Zuech, Taghi M Khoshgoftaar, and Randall Wald. Intrusion detection and big heterogeneous data: a survey. *Journal of Big Data*, 2(1):3, 2015.

[3] Sara Alspaugh, Bei Di Chen, Jessica Lin, Archana Ganapathi, Marti A Hearst, and Randy H Katz. Analyzing log analysis: An empirical study of user log mining. In *LISA*, pages 53–68, 2014.

[4] Adam Oliner, Archana Ganapathi, and Wei Xu. Advances and challenges in log analysis. *Queue*, 9(12):30, 2011.

[5] Adam Oliner, Archana Ganapathi, and Wei Xu. Advances and challenges in log analysis. *Communications of the ACM*, 55(2):55–61, 2012.

[6] Yanpei Chen, Kiran Srinivasan, Garth Goodson, and Randy Katz. Design implications for enterprise storage systems via multi-dimensional trace analysis. In *Proceedings of the Twenty-Third ACM Symposium on Operating Systems Principles*, pages 43–56. ACM, 2011.

[7] Yanpei Chen, Sara Alspaugh, and Randy Katz. Interactive analytical processing in big data systems: A cross-industry study of mapreduce workloads. *Proceedings of the VLDB Endowment*, 5(12):1802–1813, 2012.

[8] Rolf Isermann. *Fault-diagnosis systems: an introduction from fault detection to fault tolerance*. Springer Science & Business Media, 2006.

[9] Gideon Cohen, Moshe Meiseles, and Eran Reshef. System and method for risk detection and analysis in a computer network, October 4 2005. US Patent 6,952,779.

[10] Maha Bakoben, Niall Adams, and Anthony Bellotti. Uncertainty aware clustering for behaviour in enterprise networks. In *Data Mining Workshops (ICDMW), 2016 IEEE 16th International Conference on*, pages 269–272. IEEE, 2016.

[11] Melissa JM Turcotte, Nicholas A Heard, and Alexander D Kent. Modelling user behavior in a network using computer event logs. *Dynamic Networks and Cyber-Security*, 1:67, 2016.

[12] Marina Evangelou and Niall M Adams. Predictability of netflow data. In *Intelligence and Security Informatics (ISI), 2016 IEEE Conference on*, pages 67–72. IEEE, 2016.

[13] Justin Myers, Michael R Grimaila, and Robert F Mills. Towards insider threat detection using web server logs. In *Proceedings of the 5th Annual Workshop on Cyber Security and Information Intelligence Research: Cyber Security and Information Intelligence Challenges and Strategies*, page 54. ACM, 2009.

[14] Hårek Haugerud and Sigmund Straumsnes. Simulation of user-driven computer behaviour. In *LISA*, pages 101–108, 2001.

[15] Adetokunbo AO Makanju, A Nur Zincir-Heywood, and Evangelos E Milios. Clustering event logs using iterative partitioning. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1255–1264. ACM, 2009.

[16] Mark Whitehouse, Marina Evangelou, and Niall M Adams. Activity-based temporal anomaly detection in enterprise-cyber security. In *Intelligence and Security Informatics (ISI), 2016 IEEE Conference on*, pages 248–250. IEEE, 2016.

[17] Nick Heard and Patrick Rubin-Delanchy. Network-wide anomaly detection via the dirichlet process. In *Intelligence and Security Informatics (ISI), 2016 IEEE Conference on*, pages 220–224. IEEE, 2016.

[18] Jordan Noble and Niall M Adams. Correlation-based streaming anomaly detection in cyber-security. In *Data Mining Workshops (ICDMW), 2016 IEEE 16th International Conference on*, pages 311–318. IEEE, 2016.

[19] Patrick Rubin-Delanchy, Daniel J Lawson, and Nicholas A Heard. Anomaly detection for cyber security applications. *Dynamic Networks and Cyber-Security*, 1:137, 2016.

[20] Melissa Turcotte, Juston Moore, Nick Heard, and Aaron McPhall. Poisson factorization for peer-based anomaly detection. In *Intelligence and Security Informatics (ISI), 2016 IEEE Conference on*, pages 208–210. IEEE, 2016.

[21] Hossein Siadati, Bahador Saket, and Nasir Memon. Detecting malicious logins in enterprise networks using visualization. In *Visualization for Cyber Security (VizSec), 2016 IEEE Symposium on*, pages 1–8. IEEE, 2016.

[22] Jian-Guang Lou, Qiang Fu, Yi Wang, and Jiang Li. Mining dependency in distributed systems through unstructured logs analysis. *ACM SIGOPS Operating Systems Review*, 44(1):91–96, 2010.

[23] Wei Xu, Ling Huang, and Michael I Jordan. Experience mining google's production console logs. In *SLAML*, 2010.

[24] Eric Cole. A SANS Survey: Threat Hunting: Open Season on the Adversary, April 2016.

[25] John McHugh. Testing intrusion detection systems: a critique of the 1998 and 1999 darpa intrusion detection system evaluations as performed by lincoln laboratory. *ACM Transactions on Information and System Security (TISSEC)*, 3(4):262–294, 2000.

[26] Matthew V Mahoney and Philip K Chan. An analysis of the 1999 darpa/lincoln laboratory evaluation data for network anomaly detection. In *International Workshop on Recent Advances in Intrusion Detection*, pages 220–237. Springer, 2003.

[27] Alexander D. Kent. Comprehensive, Multi-Source Cyber-Security Events. http://csr.lanl.gov/data/cyber1/, 2015.

[28] Alexander D. Kent. Cybersecurity Data Sources for Dynamic Network Research. In *Dynamic Networks in Cybersecurity*. Imperial College Press, June 2015.

[29] Burke Johnson and Lisa A Turner. Data collection strategies in mixed methods research. *Handbook of mixed methods in social and behavioral research*, pages 297–319, 2003.

[30] Nick Heard, Konstantina Palla, and Maria Skoularidou. Topic modelling of authentication events in an enterprise computer network. In *Intelligence and Security Informatics (ISI), 2016 IEEE Conference on*, pages 190–192. IEEE, 2016.

[31] Patrick Rubin-Delanchy, Niall M Adams, and Nicholas A Heard. Disassortativity of computer networks. In *Intelligence and Security Informatics (ISI), 2016 IEEE Conference on*, pages 243–247. IEEE, 2016.

[32] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

[33] Helmut Lütkepohl. *New introduction to multiple time series analysis*. Springer Science & Business Media, 2005.

[34] Brian S Everitt, Sabine Landau, Morven Leese, and Daniel Stahl. Hierarchical clustering. *Cluster Analysis, 5th Edition*, pages 71–110, 2011.

[35] Maha Bakoben, Anthony Bellotti, and Niall Adams. Improving clustering performance by incorporating uncertainty. *Pattern Recognition Letters*, 77:28–34, 2016.

[36] Thomas S Ferguson. A bayesian analysis of some nonparametric problems. *The annals of statistics*, pages 209–230, 1973.

[37] Jens Dittrich and Jorge-Arnulfo Quiané-Ruiz. Efficient big data processing in hadoop mapreduce. *Proceedings of the VLDB Endowment*, 5(12):2014–2015, 2012.

[38] Prem Gopalan, Jake M Hofman, and David M Blei. Scalable recommendation with poisson factorization. *arXiv preprint arXiv:1311.1704*, 2013.