Harvard Data Science Review • Issue 5.1, Winter 2023

Statistical Cybersecurity: A Brief Discussion on Challenges, Data Structures, and Future Directions

Francesco Sanna Passino[1] Niall M. Adams[1] Edward A.K. Cohen[1] Marina Evangelou[1] Nicholas A. Heard[1]

[1]Department of Mathematics, Imperial College London, London, England, United Kingdom

DOI: https://doi.org/10.1162/99608f92.240383c7

Introduction

We congratulate the authors of Hero et al. (2023) for a very interesting and comprehensive review of the current challenges in the field of statistics and data science for cybersecurity applications. In this commentary, we would like to contribute by expanding upon some of the points raised by the authors in their Section 3, "Data-Driven Cybersecurity for Enterprise Systems," describing some of the challenges faced by statisticians working in cybersecurity, in particular regarding data structures, and emphasising directions for future work and research in statistical modeling for cybersecurity.

As Hero et al. (2023) point out, statistical modeling currently represents the main tool for *anomaly-based detection*, which looks for deviations from a model of the normal behavior of the network (see, for example, Chandola et al., 2009). As discussed in the article, statistical models have the main advantage of being able to assign anomaly scores to previously unseen events, by *borrowing strength* between different users, hosts, and processes. In this way, previously unobserved attacks, or *zero-day exploits*, can still potentially be identified. This remarkable feature of statistical models has been demonstrated extensively in the literature, with the objective, for example, to discover compromised credentials and lateral movement within a computer network (Neil et al., 2013).

Statistical models for cybersecurity also present domain-specific challenges that might not be commonly observed in other applications. For example, as the authors point out in their list of challenges in data-centric cybersecurity, training of statistical models is often difficult because of the *dearth of labels* problem. Labels about attacks and intrusions are often available only on a small and controlled subset of the data, whereas the remaining activity is its large majority, assumed to be benign. For example, in user-authentication and network flow data, only events from red-team exercises might be labeled and considered as malicious. From a statistical perspective, caution is needed when evaluating models only using a small subset of known labels: we might be at risk of *overfitting* models on the available labels, potentially ignoring undetected attacks in the unlabelled set of observations. Therefore, constant feedback and communication between cyber analysts and data scientists is required for calibrating models and understanding their output. Furthermore, automated threat detection should ideally be based on highly interpretable models, which can be appropriately modified and updated to react and adjust to novel attack patterns.

Keeping these challenges in mind, this commentary discusses three common data structures that are observed in cybersecurity and related statistical methodologies. This is followed by a discussion on future possible research directions in data fusion and streaming data analysis, which are needed to combine evidence from different data sources and update models when new data are available.

Three Examples of Data Structures in Cybersecurity

Data collected on Internet of Things (IoT) devices come in heterogeneous forms. Three of the most common data structures observed in cybersecurity applications are graphs, point processes, and textual data. Each of these data types requires different statistical techniques for analysis. It is important to remark that data in cybersecurity applications do not come exclusively in these forms. Also, different data structures and corresponding models in cybersecurity are often deeply intertwined by complex dependencies. Therefore, different data sources might be appropriately *combined* via *meta-analysis* and *data fusion* techniques, in order to capture the full complexity of a computer network.

Graphs. Enterprise computer networks can be mathematically interpreted as *complex dynamic graphs*. For example, in network flow data, hosts and their IP addresses correspond to the nodes, and edges are drawn if data transfers between them are observed. In cybersecurity, nodes and edges have additional information that could be incorporated into statistical models. In the network flow data example, nodes might have associated geolocations, and edges could occur on different ports. Under the representation of an enterprise computer network as a graph, there are essentially three main types of models that can be built, at different levels of resolution: *global*, *node-based*, and *edge-based* models.

Global models describe the connectivity patterns observed over the entire network by modeling, for example, the graph adjacency matrix or a transformation thereof. Latent position models (LPMs; Hoff et al., 2002) have emerged as suitable techniques to find low-dimensional hidden structures from adjacency matrices. In cybersecurity, two special cases of LPMs are particularly useful: generalized random dot product graphs (GRDPG; Rubin-Delanchy et al., 2022), which can be estimated quickly via spectral decompositions, and hierarchical Poisson matrix factorization models (HPMF; Gopalan et al., 2015), whose likelihood function only depends on the number of observed links, which is particularly advantageous for sparse large graphs observed in cybersecurity.

Node-based and *edge-based* models are constructed at the level of an individual host or host pair (see, for example, Turcotte et al., 2014). These models can capture complex relationships at a finer level of granularity, that would be missed from a global graph analysis. Data on nodes and edges could usually be interpreted as realizations of point processes, the second data type discussed in this commentary.

Point processes. When analyzing individual hosts, data often consist of realizations of events over time, with additional features characterizing the event type. In the statistics literature, such data structures are named *marked point processes*. A particular class of point processes is represented by mutually exciting processes, which include popular models such as Hawkes processes, commonly used in numerous applications including empirical high-frequency trading in finance and geological sciences. Mutually exciting processes have a particularly appealing construction for modeling computer networks for intrusion detection, as these are designed to learn complex dependencies between observations both within and between different processes

running simultaneously on a network or on a host. Promising work evaluating the performance of such processes for the purposes of anomaly detection has been developed (Shchur et al., 2021), including applications to network security (Shlomovich, Cohen, & Adams, 2022; Zheng et al., 2021.

One of the main difficulties for successfully fitting point process models on cybersecurity data is that a deep understanding of the data collection process is required. For example, in network flow data, some events occur periodically in bursts of activity caused by the division in packets: these events are appropriately modeled via point processes with periodic intensity functions. On the other hand, events triggered by a human user are better modeled via self-exciting processes (Price-Williams & Heard, 2020), and would need to be modeled separately from automatically generated and periodic events. Therefore, even on individual nodes or edges, it is necessary to *combine* different modeling approaches in order to produce reliable anomaly scores.

Another major issue is that point processes, while generally modeled as continuous time objects, rarely reveal themselves as such, with events often being binned into a time series of counts. The extent to which data are binned is typically dictated by the data storage or recording devices and results in a loss of temporal resolution and information. New inference methodology that can counter these binning operations has recently been developed for the cybersecurity setting (Shlomovich, Cohen, Adams, & Patel, 2022), and models that directly handle the count data themselves is an active area of research (Fokianos et al., 2020). However, more broadly, the effects on inference and prediction tasks resulting from applying binning transformations to the data is yet to be fully understood.

The final difficulty we highlight here is that events might have associated *marks*: for example, a user logging on to a host might also run some commands on the machine, expressed in the form of *text*, the third data source discussed in this commentary. The difficulty of incorporating these marks in point process models often means they are ignored, while in fact they are information-rich and there to be exploited, if used correctly.

Textual data. In the introduction to this commentary, it was noted that statistical models in cybersecurity often suffer from the *dearth of labels* problem. An exception to this rule is session data collected on honeypots, which could be used as examples of attacks to an enterprise computer network, which makes them particularly useful for developing intrusion detection systems (Highnam et al., 2021). Session data correspond to sequences of commands, which can be interpreted as *text data*. Cyber analysts are often interested in *classifying* such sessions into groups with similar behavior, and identifying sessions corresponding to novel attacks. A popular statistical model for textual data is latent Dirichlet allocation (LDA), where each word is drawn from a mixture of distribution corresponding to latent unknown *topics*. Unfortunately, vanilla LDA is not suitable for honeypot data, since it presents unidentifiability and convergence issues that make its interpretation complex for cyber analysts and threat experts. Therefore, LDA and other more general language models require careful adaptation for use in cybersecurity applications. For example, extensions of LDA that aid identifiability and interpretation have been employed in the literature to classify session data collected on honeypots Sanna Passino et al., 2023;

, helping analysts at Microsoft to discover a variant of the MIRAI malware that attempts to take over existing coin miner infrastructure (Bevington, 2021).

More generally, methods for text analysis have great potential for tasks such as malware identification and classification (see, for example, Gibert et al., 2020). Recent advancements in language modeling in the field of deep learning could provide significant benefits to cybersecurity applications. Such models are not only useful for the *analysis* of session data, but also for their *preprocessing* and *parsing*. For example, software libraries based on deep neural networks and transformers have been developed for flexibly parsing computer logs, adding the flexibility that architectures based on regular expressions might lack.

Future Directions: Streaming Methods and Data Fusion

In the previous section, we noted three common data types in cybersecurity and some of the related statistical challenges. As discussed in Hero et al. (2023), one of the main difficulties is the *adversarial nature* of the problem. Therefore, all the statistical methodologies developed for cybersecurity applications must be resilient to new threats, and adapted when necessary. Also, because of the multiple data sources available, evidence from different models must be appropriately *combined*. In statistics, these challenges could be addressed via *streaming data analysis* and *data fusion* methods.

Streaming data analysis. Streaming data analysis methods are aimed at processing data that are generated as continuous flows, using a fixed and limited amount of memory and computational power. In statistical applications, this usually translates into models with *sequential updates* for estimates of the model parameters, which could be adapted when new observations are available. A flexible framework for streaming data analysis methods is to incorporate a *forgetting factor* parameter within the model, used to express how far into the past data is 'remembered' and consequently how quickly the parameter estimates adapt to changes in the state of nature. Forgetting factor approaches have been successfully applied on a variety of streaming anomaly detection tasks in cybersecurity (see, for example, Riddle-Workman et al., 2018).

Anomaly detection models for cybersecurity applications should also be able to score events involving previously unseen entities. For example, in text analysis, models should be capable of handling *previously unobserved tokens* appearing in the vocabulary, which might correspond to new examples of malware, or new variants of bots. Also, new attack types might arise, and algorithms should be able to identify whether new observations are not a good match with previously identified intents. Within this context, Bayesian nonparametric methods could be used as a principled way to model infinite dimensional discrete distributions, providing statistical tools to score events involving previously unseen entities (see, for example, Sanna Passino et al., 2023; Zheng et al., 2021).

Data fusion and meta-analysis. As discussed in Hero et al. (2023), and in other parts of this commentary, data collected in cybersecurity applications tend to be highly complex. Therefore, it is often beneficial to *break down* the full model for a given data structure into simpler components, at a finer level of granularity. Under this

framework, it is crucial to understand *how to combine evidence* obtained from such models. In statistics, this task is usually known as *meta-analysis*, a procedure aimed at merging the results of multiple independent studies, obtaining an overall *global* effect (see, for example, Hedges & Olkin, 2014).

Also, statistical models for cybersecurity could benefit from *fusing* information collected from different sources. A specific challenge observed with cybersecurity data is that data observed from different sources cannot be considered *independent*, and often present *correlation*. For example, consider the case of the unified host and network data set released by the Los Alamos National Laboratory (Turcotte et al., 2018): a model for the network connectivity observed from a given host should be combined with models for the processes that are run on the machine, giving a full picture of the activity of the node by fusing two different data sources (for an example with multi-type clustering, see Riddle-Workman et al., 2021). We believe that efforts in this direction are needed from the community of researchers in statistical cybersecurity, with the objective of improving existing anomaly detection systems.

Disclosure Statement

Francesco Sanna Passino, Niall M. Adams, Edward A.K. Cohen, Marina Evangelou, and Nicholas A. Heard have no financial or non-financial disclosures to share for this article.

References

Bevington, R. (2021, September 17). *Unusual MIRAI variant looks for mining infrastructure*. Microsoft Sentinel Blog. https://techcommunity.microsoft.com/t5/microsoft-sentinel-blog/unusual-mirai-variant-looks-for-mining-infrastructure/ba-p/2756669

Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. *ACM Computing Surveys*, *41*(3), Article 15. https://doi.org/10.1145/1541880.1541882

Fokianos, K., Støve, B., Tjøstheim, D., & Doukhan, P. (2020). Multivariate count autoregression. *Bernoulli*, *26*(1), 471–499. https://doi.org/10.3150/19-BEJ1132

Gibert, D., Mateu, C., & Planes, J. (2020). The rise of machine learning for detection and classification of malware: Research developments, trends and challenges. *Journal of Network and Computer Applications*, *153*, Article 102526. https://doi.org/10.1016/j.jnca.2019.102526

Gopalan, P., Hofman, J. M., & Blei, D. M. (2015). Scalable recommendation with hierarchical Poisson factorization. In M. Meila & T. Heskes (Eds.), *UAI'15: Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence* (pp. 326–335). ACM. https://dlnext.acm.org/doi/10.5555/3020847.3020882

Hedges, L., & Olkin, I. (2014). *Statistical methods for meta-analysis*. Elsevier Science.

Hero, A. O., Kar, S., Moura, J., Neil, J., Poor, H. V., Turcotte, M., & Xi, B. (2023). Statistics and data science for cybersecurity. *Harvard Data Science Review*, *5*(1). https://doi.org/10.1162/99608f92.a42024d0

Highnam, K., Arulkumaran, K., Hanif, Z., & Jennings, N. R. (2021, July 23). *BETH dataset: Real cybersecurity data for anomaly detection research* [Poster presentation]. ICML Workshop on Uncertainty and Robustness in Deep Learning, Los Angeles, CA. http://www.gatsby.ucl.ac.uk/~balaji/udl2021/accepted-papers/UDL2021-paper-033.pdf

Hoff, P. D., Raftery, A. E., & Handcock, M. S. (2002). Latent space approaches to social network analysis. *Journal of the American Statistical Association*, *97*(460), 1090–1098. https://doi.org/10.1198/016214502388618906

Neil, J., Hash, C., Brugh, A., Fisk, M., & Storlie, C. B. (2013). Scan statistics for the online detection of locally anomalous subgraphs. *Technometrics*, *55*(4), 403–414. https://doi.org/10.1080/00401706.2013.822830

Price-Williams, M., & Heard, N. A. (2020). Nonparametric self-exciting models for computer network traffic. *Statistics and Computing*, *30*(2), 209–220. https://doi.org/10.1007/s11222-019-09875-z

Riddle-Workman, E., Evangelou, M., & Adams, N. M. (2018). Adaptive anomaly detection on network data streams. In *2018 IEEE International Conference on Intelligence and Security Informatics (ISI)* (pp. 19–24). IEEE. https://doi.org/10.1109/ISI.2018.8587401

Riddle-Workman, E., Evangelou, M., & Adams, N. M. (2021). Multi-type relational clustering for enterprise cyber-security networks. *Pattern Recognition Letters*, *149*, 172–178. https://doi.org/10.1016/j.patrec.2021.05.021

Rubin-Delanchy, P., Cape, J., Tang, M., & Priebe, C. E. (2022). A statistical interpretation of spectral embedding: The generalised random dot product graph. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *84*(4), 1446–1473. https://doi.org/10.1111/rssb.12509

Sanna Passino, F., Mantziou, A., Ghani, D., Thiede, P., Bevington, R., & Heard, N. A. (2023). *Unsupervised attack pattern detection in honeypot data using Bayesian topic modelling*. ArXiv. https://doi.org/10.48550/arXiv.2301.02505

Shchur, O., Turkmen, A. C., Januschowski, T., Gasthaus, J., & Günnemann, S. (2021). Detecting anomalous event sequences with temporal point processes. *Advances in Neural Information Processing Systems*, *34*. https://proceedings.neurips.cc/paper/2021/file/6faa8040da20ef399b63a72d0e4ab575-Paper.pdf

Shlomovich, L., Cohen, E. A. K., & Adams, N. (2022). A parameter estimation method for multi-variate binned Hawkes processes. *Statistics and Computing*, *32*(6), Article 98. https://doi.org/10.1007/s11222-022-10121-2

Shlomovich, L., Cohen, E. A. K., Adams, N., & Patel, L. (2022). Parameter estimation of binned Hawkes processes. *Journal of Computational and Graphical Statistics*, *31*(4), 990–1000. https://doi.org/10.1080/10618600.2022.2050247

Turcotte, M. J. M., Heard, N., & Neil, J. (2014). Detecting localised anomalous behaviour in a computer network. In H. Blockeel, M. van Leeuwen, & V. Vinciotti (Eds.), *Advances in Intelligent Data Analysis XIII* (pp. 321–332). Springer. https://doi.org/10.1007/978-3-319-12571-8_28

Turcotte, M. J. M., Kent, A. D., & Hash, C. (2018). Unified host and network data set. In N. Heard, N. Adams, P. Rubin-Delanchy, & M. Turcotte (Eds.), *Data science for cyber-security* (pp. 1–22). World Scientific. https://doi.org/10.1142/9781786345646_001

Zheng, P., Yuan, S., & Wu, X. (2021). Using Dirichlet marked Hawkes processes for insider threat detection. *Digital Threats: Research and Practice*, *3*(1), 1–19. https://doi.org/10.1145/3457908

References

- Bevington, R. (2021, September 17). *Unusual MIRAI variant looks for mining infrastructure*. Microsoft Sentinel Blog. https://techcommunity.microsoft.com/t5/microsoft-sentinel-blog/unusual-mirai-variant-looks-for-mining-infrastructure/ba-p/2756669 ↩
- Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. *ACM Computing Surveys*, *41*(3), Article 15. https://doi.org/10.1145/1541880.1541882 ↩
- Fokianos, K., Støve, B., Tjøstheim, D., & Doukhan, P. (2020). Multivariate count autoregression. *Bernoulli*, *26*(1), 471–499. https://doi.org/10.3150/19-BEJ1132 ↩
- Gibert, D., Mateu, C., & Planes, J. (2020). The rise of machine learning for detection and classification of malware: Research developments, trends and challenges. *Journal of Network and Computer Applications*,

↩

- Gopalan, P., Hofman, J. M., & Blei, D. M. (2015). Scalable recommendation with hierarchical Poisson factorization. In M. Meila & T. Heskes (Eds.), *UAI'15: Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence* (pp. 326–335). ACM. https://dlnext.acm.org/doi/10.5555/3020847.3020882

↩

- Hedges, L., & Olkin, I. (2014). *Statistical methods for meta-analysis*. Elsevier Science.

↩

- Hero, A. O., Kar, S., Moura, J., Neil, J., Poor, H. V., Turcotte, M., & Xi, B. (2023). Statistics and data science for cybersecurity. *Harvard Data Science Review*, *5*(1). https://doi.org/10.1162/99608f92.a42024d0

↩

- Highnam, K., Arulkumaran, K., Hanif, Z., & Jennings, N. R. (2021, July 23). *BETH dataset: Real cybersecurity data for anomaly detection research* [Poster presentation]. ICML Workshop on Uncertainty and Robustness in Deep Learning, Los Angeles, CA. http://www.gatsby.ucl.ac.uk/~balaji/udl2021/accepted-papers/UDL2021-paper-033.pdf

↩

- Hoff, P. D., Raftery, A. E., & Handcock, M. S. (2002). Latent space approaches to social network analysis. *Journal of the American Statistical Association*, *97*(460), 1090–1098. https://doi.org/10.1198/016214502388618906

↩

- Neil, J., Hash, C., Brugh, A., Fisk, M., & Storlie, C. B. (2013). Scan statistics for the online detection of locally anomalous subgraphs. *Technometrics*, *55*(4), 403–414. https://doi.org/10.1080/00401706.2013.822830

↩

- Price-Williams, M., & Heard, N. A. (2020). Nonparametric self-exciting models for computer network traffic. *Statistics and Computing*, *30*(2), 209–220. https://doi.org/10.1007/s11222-019-09875-z

↩

- Riddle-Workman, E., Evangelou, M., & Adams, N. M. (2018). Adaptive anomaly detection on network data streams. In *2018 IEEE International Conference on Intelligence and Security Informatics (ISI)* (pp. 19–24). IEEE. https://doi.org/10.1109/ISI.2018.8587401

↩

-

↩

- Rubin-Delanchy, P., Cape, J., Tang, M., & Priebe, C. E. (2022). A statistical interpretation of spectral embedding: The generalised random dot product graph. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *84*(4), 1446–1473. https://doi.org/10.1111/rssb.12509

↩

- Sanna Passino, F., Mantziou, A., Ghani, D., Thiede, P., Bevington, R., & Heard, N. A. (2023). *Unsupervised attack pattern detection in honeypot data using Bayesian topic modelling*. ArXiv. https://doi.org/10.48550/arXiv.2301.02505

↩

- Shchur, O., Turkmen, A. C., Januschowski, T., Gasthaus, J., & Günnemann, S. (2021). Detecting anomalous event sequences with temporal point processes. *Advances in Neural Information Processing Systems*, *34*. https://proceedings.neurips.cc/paper/2021/file/6faa8040da20ef399b63a72d0e4ab575-Paper.pdf

↩

- Shlomovich, L., Cohen, E. A. K., & Adams, N. (2022). A parameter estimation method for multi-variate binned Hawkes processes. *Statistics and Computing*, *32*(6), Article 98. https://doi.org/10.1007/s11222-022-10121-2

↩

- Shlomovich, L., Cohen, E. A. K., Adams, N., & Patel, L. (2022). Parameter estimation of binned Hawkes processes. *Journal of Computational and Graphical Statistics*, *31*(4), 990–1000. https://doi.org/10.1080/10618600.2022.2050247

↩

- Turcotte, M. J. M., Heard, N., & Neil, J. (2014). Detecting localised anomalous behaviour in a computer network. In H. Blockeel, M. van Leeuwen, & V. Vinciotti (Eds.), *Advances in Intelligent Data Analysis XIII* (pp. 321–332). Springer. https://doi.org/10.1007/978-3-319-12571-8_28

↩

- Turcotte, M. J. M., Kent, A. D., & Hash, C. (2018). Unified host and network data set. In N. Heard, N. Adams, P. Rubin-Delanchy, & M. Turcotte (Eds.), *Data science for cyber-security* (pp. 1–22). World Scientific. https://doi.org/10.1142/9781786345646_001

↩

- Zheng, P., Yuan, S., & Wu, X. (2021). Using Dirichlet marked Hawkes processes for insider threat detection. *Digital Threats: Research and Practice*, *3*(1), 1–19. https://doi.org/10.1145/3457908

↩