Harvard Data Science Review • Issue 5.1, Winter 2023

Challenges for Anomaly Detection in Large-Scale Cyber-Physical Systems

George Michailidis[1]

[1]**Department of Statistics, University of California Los Angeles, Los Angeles, California, United States of America**

Published on: Mar 20, 2023

DOI: https://doi.org/10.1162/99608f92.7b8b6a89

The authors of the Hero et al. (2023) article should be congratulated for their nice overview of problems and challenges arising from cybersecurity threats in large enterprise systems, and the role of statistical and data science methods to address them. The broad methodological thrusts discussed include distributed statistical inference, data fusion, anomaly detection, and adversarial machine learning. In the sequel, a number of issues related to the challenging anomaly detection problem and the associated one of change point detection are briefly discussed.

The scope of anomaly and change point detection in cyber-physical systems is more general than simply security considerations. For example, modern computer and communications networks and enterprise systems have become ubiquitous in the lives of individuals, as well as the function of organizations and governments. They support many services, from mature ones such as file sharing, email, web browsing, and cloud computing to fast-evolving ones such as remote education and telemedicine. However, to achieve their potential, certain requirements on quality-of-service need to be met by different stakeholders, including network providers, designers of enterprise systems, and applications developers. Further, quality-of-service can degrade for various reasons, including intrusions by unauthorized users, broad coordinated attacks (e.g., distributed denial of service), or even underprovision of resources by network providers or exceedingly bandwidth-hungry applications. Further, as pointed out in Hero et al. (2023), the ever-increasing scale and complexity of network and enterprise systems contribute to the challenges.

Anomaly detection broadly refers to the problem of finding patterns in the data that do not conform to expected behavior Chandola et al., 2009. Anomalies can correspond to outliers, that is, isolated observations taking very large/low values, burst of outliers, sequences of observations that deviate from normally established patterns, and other unusual events. Note that novelty detection (Pimentel et al., 2014) is a topic related to anomaly detection, and aims at detecting emergent novel patterns in the data. On many occasions, change point detection techniques are incorporated in a novelty detection pipeline. The change point detection problem is concerned with identifying changes in the distribution of the data, either in an offline or an online manner. In the offline version of the problem, the entire sequence of univariate/multivariate time observations is available at analysis time and the goal becomes to identify if there exist any change points, and in their presence identify their locations in the sequence, assuming some model/mechanism for their temporal evolution. The offline setting is most useful for performing root cause analysis and gaining insights into factors that contributed to the distribution shift and also for annotating the data sequences that can subsequently be used for training supervised learning models. On the other hand, the online version assumes that new observations are obtained over time and the goal is to identify a change in the distribution that currently governs their behavior as soon as possible, possibly followed by some mitigation action. This setting is most useful for monitoring purposes and for activating mitigation policies; examples of mitigation strategies for distributed denial of service attacks in software defined networks are reviewed in Valdovinos et al. (2021).

There is a large body of literature on techniques aiming to identify anomalies (see, e.g., surveys by Ahmed et al., 2016; Chalapathy & Chawla, 2019; and Chandola et al., 2009) or change points (see, e.g., surveys by Truong et al., 2020, and Aminikhanghahi & Cook, 2017). A quick inspection of the literature reveals that most methods are tailored to univariate streams of data. In the presence of multiple streams, a popular strategy is to leverage univariate techniques for each stream and then use different rules to aggregate the results. The advantage of this strategy is technical simplicity and ease of implementation in a distributed manner, since each 'detector' needs to communicate to a fusion center only its decision. On the other hand, multivariate techniques are more capable of detecting *coordinated* anomalies that may lie below the threshold of individual detectors. However, such techniques exhibit higher communication and computational costs. Distributed algorithms that minimize communication costs become particularly useful in this setting.

Despite the extensive literature on anomaly detection techniques, enterprise systems and cyberthreats pose novel challenges; selected ones are outlined next. Anomaly detection techniques require identification of normal behavior; however, the latter may keep evolving and thus the current normal 'regime' requires identification that can be accomplished by employing online change point detection techniques. Note that an extensive body of work on this topic has been developed in the statistical process control literature—see, for example, Stoumbos et al. (2000), Bersimis et al. (2007), and references therein. Often, anomalies are the result of malicious activity, and therefore sophisticated adversaries aim to make the anomalous observations appear rather normal, thus making the task of defining normal behavior more difficult. Many anomaly detection techniques require labeled data, but in fast-evolving dynamic environments, the latter becomes costly to acquire. Finally, in different domains, the definition of an anomaly may differ. For example, for certain applications, only significant increases in response delays can lead to a degradation of quality-of-service and the associated quality of experience for end users. On the other hand, for a well-provisioned communications network, even a slight increase in the packet loss rate may be consequential for certain services.

Analogously, new challenges emerge for both the offline and online settings of the change point detection problem, with selected ones outlined next. A wealth of techniques focus on detecting changes in the mean of univariate or multivariate data streams. However, in many applications, it is more appropriate to focus on higher moments (e.g., variance/covariance) or even on the tail behavior of the distribution. One could argue that nonparametric techniques may be more suitable due to their generality, but they come with their own challenges, including computational ones and possible lack of adequate detection power. Further, new data structures such as network or tensor data require adaptation of existing, or development of new detection algorithms, as well as technical developments for providing theoretical guarantees for their performance (see, e.g., Bhattacharjee et al., 2020; Keshavarz et al., 2020; and the panel discussion in Stevens et al., 2021). Further, current algorithms for both offline and online settings are centralized in nature and require synchronous observations. However, the scale of enterprise systems and power constraints in Internet-of-Things systems require distributed computations and the ability of detection algorithms to accommodate asynchronous observations. On that front, technical developments for data fusion tasks in wireless sensor

networks (see, e.g., Ji & Cai, 2012; Rabbat & Nowak, 2004) are of interest to the problem of change point detection.

In summary, modern large-scale enterprise systems and cyberthreats have created a number of exciting opportunities for new statistical and data science methodology for the anomaly detection problem. However, for the research community to make fast progress, availability of new large-scale, well-documented and curated data sets that reflect current infrastructure and applications developments would be necessary.

Disclosure Statement

The work of GM was supported in part by NSF grant DMS 2210358.

References

Ahmed, M., Mahmood, A. N., & Hu, J. (2016). A survey of network anomaly detection techniques. *Journal of Network and Computer Applications*, *60*, 19–31. https://doi.org/10.1016/j.jnca.2015.11.016

Aminikhanghahi, S., & Cook, D. J. (2017). A survey of methods for time series change point detection. *Knowledge and Information Systems*, *51*(2), 339–367. https://doi.org/10.1007/s10115-016-0987-z

Bersimis, S., Psarakis, S., & Panaretos, J. (2007). Multivariate statistical process control charts: An overview. *Quality and Reliability Engineering International*, *23*(5), 517–543. https://doi.org/10.1002/qre.829

Bhattacharjee, M., Banerjee, M., & Michailidis, G. (2020). Change point estimation in a dynamic stochastic block model. *The Journal of Machine Learning Research*, *21*(1), 4330–4388. https://www.jmlr.org/papers/v21/18-814.html

Chalapathy, R., & Chawla, S. (2019). *Deep learning for anomaly detection: A survey*. ArXiv. https://doi.org/10.48550/arXiv.1901.03407

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

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

Ji, S., & Cai, Z. (2012). Distributed data collection and its capacity in asynchronous wireless sensor networks. In *2012 Proceedings IEEE INFOCOM* (pp. 2113–2121). IEEE. https://doi.org/10.1109/INFCOM.2012.6195594

Keshavarz, H., Michailidis, G., & Atchad´e, Y. (2020). Sequential change-point detection in high-dimensional Gaussian graphical models. *The Journal of Machine Learning Research*, *21*(1), 3125–3181.

https://jmlr.org/papers/volume21/18-410/18-410.pdf

Pimentel, M. A., Clifton, D. A., Clifton, L., & Tarassenko, L. (2014). A review of novelty detection. *Signal Processing*, *99*, 215–249. https://doi.org/10.1016/j.sigpro.2013.12.026

Rabbat, M., & Nowak, R. (2004). Distributed optimization in sensor networks. In K. Ramchandran & J. Sztipanovits (Eds.), *Proceedings of the 3rd International Symposium on Information Processing in Sensor Networks* (pp. 20–27). ACM. https://doi.org/10.1145/984622.984626

Stevens, N. T., Wilson, J. D., Driscoll, A. R., McCulloh, I., Michailidis, G., Paris, C., Parker, P., Paynabar, K., Perry, M. B., Reisi-Gahrooei, M., Sengupta, S., & Sparks, R. (2021). Research in network monitoring: Connections with SPM and new directions. *Quality Engineering*, *33*(4), 736–748. https://doi.org/10.1080/08982112.2021.1974035

Stoumbos, Z. G., Reynolds, M. R., Jr., Ryan, T. P., & Woodall, W. H. (2000). The state of statistical process control as we proceed into the 21st century. *Journal of the American Statistical Association*, *95*(451), 992–998. https://doi.org/10.1080/01621459.2000.10474292

Truong, C., Oudre, L., & Vayatis, N. (2020). Selective review of offline change point detection methods. *Signal Processing*, *167*, Article 107299. https://doi.org/10.1016/j.sigpro.2019.107299

Valdovinos, I. A., P´erez-D´ıaz, J. A., Choo, K.-K. R., & Botero, J. F. (2021). Emerging DDoS attack detection and mitigation strategies in software-defined networks: Taxonomy, challenges and future directions. *Journal of Network and Computer Applications*, *187*, Article 103093. https://doi.org/10.1016/j.jnca.2021.103093

References

- Ahmed, M., Mahmood, A. N., & Hu, J. (2016). A survey of network anomaly detection techniques. *Journal of Network and Computer Applications*, *60*, 19–31. https://doi.org/10.1016/j.jnca.2015.11.016 ↩
- Aminikhanghahi, S., & Cook, D. J. (2017). A survey of methods for time series change point detection. *Knowledge and Information Systems*, *51*(2), 339–367. https://doi.org/10.1007/s10115-016-0987-z ↩
- Bersimis, S., Psarakis, S., & Panaretos, J. (2007). Multivariate statistical process control charts: An overview. *Quality and Reliability Engineering International*, *23*(5), 517–543. https://doi.org/10.1002/qre.829

↩

- Bhattacharjee, M., Banerjee, M., & Michailidis, G. (2020). Change point estimation in a dynamic stochastic block model. *The Journal of Machine Learning Research*, *21*(1), 4330–4388. https://www.jmlr.org/papers/v21/18-814.html

↩

- Chalapathy, R., & Chawla, S. (2019). *Deep learning for anomaly detection: A survey*. ArXiv. https://doi.org/10.48550/arXiv.1901.03407

↩

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

↩

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

↩

- Ji, S., & Cai, Z. (2012). Distributed data collection and its capacity in asynchronous wireless sensor networks. In *2012 Proceedings IEEE INFOCOM* (pp. 2113–2121). IEEE. https://doi.org/10.1109/INFCOM.2012.6195594

↩

- Keshavarz, H., Michailidis, G., & Atchad´e, Y. (2020). Sequential change-point detection in high-dimensional Gaussian graphical models. *The Journal of Machine Learning Research*, *21*(1), 3125–3181. https://jmlr.org/papers/volume21/18-410/18-410.pdf

↩

- Pimentel, M. A., Clifton, D. A., Clifton, L., & Tarassenko, L. (2014). A review of novelty detection. *Signal Processing*, *99*, 215–249. https://doi.org/10.1016/j.sigpro.2013.12.026

↩

- Rabbat, M., & Nowak, R. (2004). Distributed optimization in sensor networks. In K. Ramchandran & J. Sztipanovits (Eds.), *Proceedings of the 3rd International Symposium on Information Processing in Sensor Networks* (pp. 20–27). ACM. https://doi.org/10.1145/984622.984626

↩

- Stevens, N. T., Wilson, J. D., Driscoll, A. R., McCulloh, I., Michailidis, G., Paris, C., Parker, P., Paynabar, K., Perry, M. B., Reisi-Gahrooei, M., Sengupta, S., & Sparks, R. (2021). Research in network monitoring:

↩

- Stoumbos, Z. G., Reynolds, M. R., Jr., Ryan, T. P., & Woodall, W. H. (2000). The state of statistical process control as we proceed into the 21st century. *Journal of the American Statistical Association*, *95*(451), 992–998. https://doi.org/10.1080/01621459.2000.10474292

↩

- Truong, C., Oudre, L., & Vayatis, N. (2020). Selective review of offline change point detection methods. *Signal Processing*, *167*, Article 107299. https://doi.org/10.1016/j.sigpro.2019.107299

↩

- Valdovinos, I. A., P´erez-D´ıaz, J. A., Choo, K.-K. R., & Botero, J. F. (2021). Emerging DDoS attack detection and mitigation strategies in software-defined networks: Taxonomy, challenges and future directions. *Journal of Network and Computer Applications*, *187*, Article 103093. https://doi.org/10.1016/j.jnca.2021.103093

↩