TALLINN UNIVERSITY OF TECHNOLOGY
School of Information Technologies

Roman Müller 177361IVCM

ANALYSIS OF THE ESTONIAN X-TEE NETWORK BASED ON CENTRALIZED LOG DATA

Master Thesis

Supervisor: Sven Nõmm

Jaan Priisalu

Tallinn 2019

TALLINNA TEHNIKAÜLIKOOL
Infotehnoloogia teaduskond

Roman Müller 177361IVCM

EESTI X-TEE VÕRGU ANALÜÜS TSENTRALISEERITUD LOGIANDMETE PÕHJAL

Magistritöö

Juhendaja: Sven Nõmm

Jaan Priisalu

Tallinn 2019

Author’s declaration of originality

I hereby certify that I am the sole author of this thesis. All the used materials, references to the literature and the work of others have been referred to. This thesis has not been presented for examination anywhere else.

Author: Roman Müller

13.05.2019

Abstract

Estonia uses the X-Road system for her internal communication between the state, insurances, banks and other companies. An extensive amount of open data about those communications is published online. This research will analyse this open data using process discovery, timing and social network analysis in order to gain insight into the structure of the Estonian bureaucratic network and identify possible vulnerabilities in this network.

Three different process discovery algorithms will be evaluated and two of them implemented and used to search for process-based dependencies. Processes are valuable targets for an attacker as he or she could focus on a small number of nodes to disrupt large parts of the network.

Furthermore, the timing of the communication events in the network is analysed to determine if the activities in the network are mostly performed by automated services or a human workforce. Based on the results of this different attacks are possible, as humans and machines are vulnerable in different ways.

Finally, the network graph of X-Road is created and divided into meaningful clusters. The clusters and important nodes are analysed and ranked using different metrics such as prestige and in-betweenness to determine which nodes are most important for the network.

This thesis is written in English and is 59 pages long, including 6 chapters, 20 figures and 14 tables.

1 Summary

In this research I used several different methods to analyse the X-Road logging data in order to unearth dependencies, weaknesses and serve as a basis for attacks on the X-Road network.

In chapter 3, I used two different process discovery techniques. The first algorithm failed to find processes that made up a significant part of the overall network. The duration-based algorithm was a bit more successful but still failed to discover a significant web of processes that could lead to a cascading failure scenario in the network.

Overall, a significant part of the X-Road actors communication relies on human workforce. The ability to determine the automation grade of a node allows an attacker to specifically draft an attack plan for each node. The information when the workers are the busiest can be used to improve the chances of a social engineering attack, as stressed workers are more vulnerable.

The capacity analysis identified that one of the most important nodes (by several metrics), the Information Technology Center of the Ministry of the Interior (Siseministeeriumi infotehnoloogia- ja arenduskeskus) might have a capacity problem in one of its subsystems. Overall most of the high performing nodes don´t show a strong correlation between the duration of a requests and the amount of requests indicating that the network is quite robust.

Chapter 5 subdivided the graph into clusters and analysed the largest clusters and most important nodes. An attacker that wants to gain an overview on how the Estonian bureaucracy is structured can gain this insight using the described methods.

The results of this document are based on the non-public data provided by the Estonian State Information Authority, which differs from publicly available data by having more precise timestamps. The precise timestamps are heavily used in the process discovery algorithm to establish the sequence of events. It is also used to estimate the automation factor of a node by checking if the event occurred during work hours and in the elasticity

analysis to estimate the capacity of a node. Would the same analysis be performed with the publicly available unprecise data, the results are likely to also be less precise. However, this loss of precision could be partly mitigated by ingesting more data over a longer period.

Overall, my research was unable to discover a concrete vulnerability, unexpected dependency or any indication that a node would be a very easy and valuable target. However, it still provides a good starting point for an attacker to identify important systems and dependencies. The possibilities of this must be weighed against the benefits of publishing this data.

Further research is recommended into timing analysis, such as if it is possible to determine the amount of people or servers providing a service by looking at the patterns of communication. The capacity analysis could be improved by choosing a custom binning interval for each node that corresponds to that node's average duration instead of the average duration of the whole dataset. Furthermore, research into the combination of X-Road data with other open data sources is needed. The Estonian government provides a list of these sources on its open data portal[1].

The central X-Road servers also collect more confidential information that this analysis did not have access to. Analysing the confidential subject field might give better results for process mining. However, since this field might contain sensitive information it was not part of the analysed dataset.

Analysing the behaviour of the network over time could be an alternative way of discovering dependencies. A prolonged downtime of a service (either scheduled or accidental) is expected to first lead to a decrease and then to an increase in communication as the service is restored and backlog cleared. This would allow to discover dependencies that were not discoverable by process mining or social network analysis. A recent event

[1] https://opendata.riik.ee/

that would be interesting to analyse is the downtime of the online system of the Tax and Customs board in May 2019[1].

[11] https://www.emta.ee/et/katkestus

2 References

[1] M. Janssen, Y. Charalabidis, and A. Zuiderwijk, “Benefits, Adoption Barriers and Myths of Open Data and Open Government,” *Inf. Syst. Manag.*, vol. 29, no. 4, pp. 258–268, Sep. 2012.

[2] Open Government Partnership, “Estonia.” [Online]. Available: https://www.opengovpartnership.org/countries/estonia. [Accessed: 16-Feb-2019].

[3] Estonian Open Government Data Portal, “Why is it important to make data openly available?” [Online]. Available: https://opendata.riik.ee/en/about/. [Accessed: 16-Feb-2019].

[4] D. Hand, “Open data is a force for good, but not without risks | Society | The Guardian,” *The Guardian*, 2012. [Online]. Available: https://www.theguardian.com/society/2012/jul/10/open-data-force-for-good-risks. [Accessed: 18-Feb-2019].

[5] N. Timmins, “Crime maps ‘hit reporting of crime,’” *Financial Times*, 2011. [Online]. Available: https://www.ft.com/content/c6b65e3e-aca2-11e0-a2f3-00144feabdc0. [Accessed: 18-Feb-2019].

[6] J. Kucera and D. Chlapek, “Benefits and Risks of Open Government Data,” *J. Syst. Integr.*, vol. 5, no. 1, pp. 30–41, 2014.

[7] A. Hern, “Fitness tracking app Strava gives away location of secret US army bases,” *The Guardian*, 2018. [Online]. Available: https://www.theguardian.com/world/2018/jan/28/fitness-tracking-app-gives-away-location-of-secret-us-army-bases. [Accessed: 18-Feb-2019].

[8] A. Veldre, “Introduction of X-tee,” *Information System Authority Estonia*, 2016. [Online]. Available: https://www.ria.ee/en/state-information-system/x-tee/introduction-x-tee.html. [Accessed: 08-Feb-2019].

[9] E-Estonia, “Interoperability services: x-road.” [Online]. Available: https://e-estonia.com/solutions/interoperability-services/x-road. [Accessed: 08-Feb-2019].

[10] T. Mölder, A. Luoma, T. Repo, and Y. Kataoka, “X-Road Terms and Abbreviations,” 2019. [Online]. Available: https://github.com/nordic-institute/X-Road/blob/develop/doc/terms_x-road_docs.md. [Accessed: 21-Jan-2019].

[11] W. van der Aalst, T. Weijters, and L. Maruster, “Workflow mining: discovering process models from event logs,” *IEEE Trans. Knowl. Data Eng.*, vol. 16, no. 9, pp. 1128–1142, Sep. 2004.

[12] S. Pourmirza, R. Dijkman, and P. Grefen, “Correlation Miner: Mining Business Process Models and Event Correlations Without Case Identifiers,” *Int. J. Coop. Inf. Syst.*, vol. 26, no. 02, p. 1742002, Jun. 2017.

[13] D. R. Ferreira and D. Gillblad, “Discovering Process Models from Unlabelled Event Logs,” Springer, Berlin, Heidelberg, 2009, pp. 143–158.

[14] F. Fouss, M. Saerens, and M. Shimbo, *Algorithms and Models for Network Data and Link Analysis*. Cambridge: Cambridge University Press, 2016.

[15] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, “Fast unfolding of communities in large networks,” Mar. 2008.