Link Prediction for Cybersecurity Companies and Technologies: Towards a Survivability Score

Santiago Anton Moreno[1(✉)], Anita Mezzetti[1], and William Lacube[2]

[1] EPFL, 1015 Lausanne, Switzerland
santiago.antonmoreno@epfl.ch
[2] Cyber Defence Campus, Armasuisse Science and Technology, 1015 Lausanne, Switzerland

Abstract. On the cybersecurity market, novel entities – technologies and companies – arise and disappear swiftly. In such a fast-paced context, assessing the survivability of those entities is crucial when it comes to make investment decisions for ensuring the security of critical infrastructures. In this paper, we present a framework for capturing the dynamic relationship between entities of the Swiss cybersecurity landscape. By using open data, we first model our dataset as a bipartite graph in which nodes are represented by technologies and companies involved in cybersecurity. Next, we use patents and job openings data to link the two entities. By extracting time series of such graphs, and by using link-prediction methods, we forecast the (dis)appearance of links. We apply several unsupervised learning similarity-based algorithms, a supervised learning method and finally we select the best model. Our preliminary results show good performance and promising validation of our survivability index. We suggest that our framework is useful for critical infrastructure operators, as a survivability index of entities can be extracted by using the outputs of our models.

Keywords: Technology monitoring · Network science · Link prediction · Time series · Supervised learning · Critical-infrastructure protection

1 Introduction

The fast-paced development of technologies reshapes the security of information systems [8]. Examples of technologies that redefine cyberdefense are numerous: e.g., quantum computing threatening cryptography protocols, adversarial machine learning, novel communication protocols, behaviour-based authentication of IDS, distributed ledgers. In such a complex technology-development context, both threats and opportunities emerge for actors of the cyberspace [3], including operators of critical infrastructures (CIs). Consequently, a race for a technological advantage takes place between attackers and defenders [7].

D. Percia David et al. (Eds.): CRITIS 2021, LNCS 13139, pp. 228–233, 2021.
https://doi.org/10.1007/978-3-030-93200-8_13

The assessment of the cybersecurity technological landscape has become a central activity when it comes to develop cyberdefense strategies [2], especially in the context of the recent supply chain attacks against CIs[1]. Such an assessment helps defenders to grab an edge in this technological race by developing threat-intelligence tools to reduce the information asymmetry between attackers and defenders [12]. In particular, it enables to foster cyberdefense by identifying the survival probabilities of entities – i.e., technologies and companies – involved in the cybersecurity technological landscape and, then investing in the most relevant ones. Especially, these aspects constitute strategic tools for procurement, one of the greatest challenges faced by governments and CI operators [5].

In this work, we aim to contribute to the technological landscape assessment effort by presenting a framework for capturing the relationship between entities of the Swiss cybersecurity technological landscape. By using a dataset coming from the *Technology & Market Monitoring* (TMM) platform, we first model the data as a bipartite graph in which nodes (i.e., entities) are represented by technologies and companies involved in cybersecurity. We then use patents and job openings to link entities. By extracting time series of such a graph, and by using link-prediction methods, we forecast the (des)appearance of links between entities. We apply several similarity-based algorithms and a supervised learning machine-learning model that uses outputs from the former. Next we select the best model based on perfomance measures. We suggest that our framework is useful for decision-makers involved in the security of CIs, as a survivability score of entities can be extracted by either using the similarity metrics or probability calculations from the supervised learning model.

The remainder of this paper is structured as follows: Sect. 2 present the related work; Sect. 3 presents the data and methods; Sect. 4 shows the preliminary results; Sect. 5 sets the agenda for future works and discusses limitations; while Sect. 6 concludes.

2 Related Work

Percolation theory has been previously used as a network-centrality measure – i.e., for determining the degree of influence of a node within a given network –, as well as for investigating the effects of a node disappearance on the overall network structure (e.g., [10]). In network science, such a percolation phenomenon can be investigated through link-prediction methods (e.g., [9]). By accounting for network structures and other available variables, link-prediction methods extract metrics accounting for the likelihood of edges (dis)appearances through time (e.g., [9]). In this respect, Kim et al. used link prediction to forecast technology convergence [4]. Additionally, Benchettara et al. [1] adapted link-prediction

[1] The 2020 Global Supply Chain Cyberattack is believed to have resulted through a supply chain attack targeting the IT infrastructure company SolarWinds, which counts many critical infrastructures among its clients. In order to fight against this type of attack, our framework may offer the possibility to identify less-secure elements in the supply chain.

metrics for bipartite networks. Moreover, Silva et al. [11] and Tylenda et al. [13] explored time dependant metrics to use with time series within link-prediction analysis. Finally, supervised learning has been applied to link-prediction investigations: Mohammad et al. [1] applied supervised learning to a co-authoring network for several classification algorithms.

However, to the best of our knowledge, link prediction as a method for assessing the dynamics of the cybersecurity technological landscape has not been explored yet. At least, we found no work focusing on predicting the survivability of entities composing the cybersecurity technological landscape. In this work, we present a network-analytics framework that employs link prediction and supervised learning to build a survivability score of entities composing the cybersecurity technological landscape.

3 Data and Methods

3.1 Data

We use the data collected by the TMM platform (ca. 1 TB) to create a bipartite network composed of technologies and companies of the Swiss cybersecurity technological landscape.[2] The TMM platform is an information system developed by armasuisse Science and Technology (S+T). TMM aims to exploit big data and open-source information in an automated way for intelligence purposes. The TMM system crawls and aggregates information from different online resources as patent offices (*Patentsview*), commercial registers (*Zefix*) and websites (*Wikipedia* and *Indeed*) to obtain a list of companies, patents and job openings based in Switzerland. By using the companies list, patents and job openings data, we link companies to technology, creating a bipartite network (2'996 nodes). We use predefined keywords of cybersecurity related technologies to compute word similarity with TMM technologies and select the most relevant. We verify the obtained list afterwards to delete any irrelevant technology and thus we obtain 69 keywords[3] from TMM. Data, available from March 2018 to December 2020 (34 time-series entries), are crawled from these platforms at different rates, and aggregated monthly. In the obtained graphs, we observe that the companies with most links are well established and long-lived tech companies like IBM but we can find all sorts of companies like the Swiss Post, Novartis and Ikea.

3.2 Methods

We define a network $G = (V, E)$, wherein V is any finite set called the vertex set and $E \subseteq V \times V$, called the edge set, corresponds to relation between elements of V. Let $x, y \in V$, such as:

– the neighborhood of x is $\Gamma(x) = \{y \in V s.t. (x, y) \in E\}$;

[2] https://tmm.dslab.ch//home.

[3] keyword list: https://tinyurl.com/jswtsmmn.

- the degree of x is $\delta_x = |\Gamma(x)|$;
- there is a path between x and y if there exists $(x_0, x_1, ..., x_n)$ such that $x_0 = x$ $x_n = y$ and $(x_i, x_{i+1}) \in E \; \forall \, 0 \leq i \leq n-1$;
- a graph G is said to be bipartite if there exists $A, B \subset V$ such that if $(x, y) \in E$ then x and y are not in the same subset A, B.

In traditional link prediction, one computes the metrics for each possible edge in a frozen network. Then, if the metric is higher than a given threshold, the edge will appear in the next step. In our case, we compute specific metrics – listed below from ((1) to (3)) – for all graphs in the time series, except for the last entry. We use time series ARIMA modelling on each metrics to predict them for the final entry and use the last graph as a validation set to compute performance metrics presented here under.

As a next step, we apply a supervised learning framework in order to obtain the best results from the metrics computed. We use a Support Vector Machine (SVM) classifier that classifies each edge to a label 0 or 1, representing the existence or not of that edge in the network. This classifier needs feature for each edge to make a decision, so we use the three similarity metrics described here under as features [11]. We train the classifier on all but the last graph and obtain test performance on it.

Since the networks are sparse, the classification problem is highly imbalanced and thus we use the area under the receiver operating characteristic curve (AUC) as the main performance metric, which is widely used in link prediction frameworks [9]. We select and evaluate eight potential metrics from prior literature and adapt the best three ones to build the predictions:
(1) *Preferential Attachment Index* [9]: The mechanism of preferential attachment has been used to generate evolving scale-free networks, where the probability of a new link forming from x is proportional to δ_x. The corresponding similarity index can be defined as:

$$s_{xy}^{PA} = \delta_x \cdot \delta_y. \tag{1}$$

(2) *Katz Index* [9]: It is a global index based on the ensemble of all paths. It sums the number of paths of a given length between x and y multiplied by a damping coefficient. The mathematical expression reads:

$$s_{xy}^{Katz} = \sum_{l=1}^{\infty} \beta^l \cdot (A^l)_{xy}. \tag{2}$$

Wherein A is the adjacency matrix of the network and β is a free parameter that damps the influence of long paths.
(3) *Hyperbolic Sine Index* [6]: The exponential of the adjacency matrix is used as a metric in unipartite link prediction, but as we work with a bipartite graph, we can take the odd part of the exponential, which is the hyperbolic sine. It can be derived by the following sum:

$$S^{sinh} = sinh(\alpha A) = \sum_{i=0}^{\infty} \frac{\alpha^{1+2i}}{(1+2i)!} A^{1+2i}. \tag{3}$$

4 Preliminary Results

We apply the methodology and algorithms presented above and obtain the receiver operating characteristic curve (ROC) and AUC diagnostics in Fig. 1. As we can see the SVM highly improves the AUC of the link prediction by approximately 4% compared to the best unsupervised method. *Preferential Attachment Index* is worse than a random classifier which could be explained by the fact that link appearance probability in our graphs is poorly related to nodes degrees. Companies may prefer investing in emerging technology which are not linked to many entities, because they may seek exclusivity in the race for technological edge.

The AUC obtained for the top 3 methods assures the validity of our metrics as a building block for a survivability index. This would help decision-makers, involved in CI's security, to identify emerging technology and companies. Future algorithms optimization presented in Sect. 5 should increase performances and thus the validity of the survivability index.

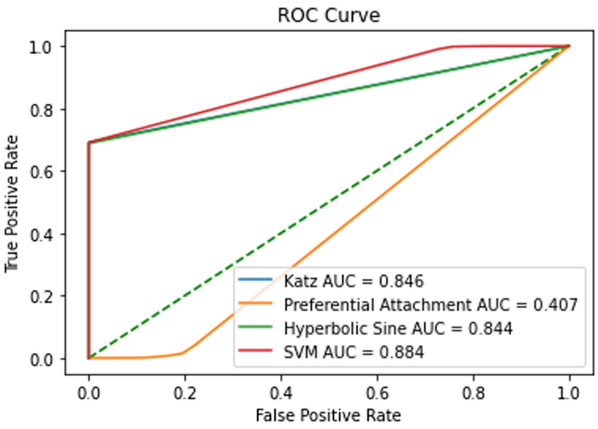

Fig. 1. ROC curves and AUC values for the 4 algorithms considered. Hyperbolic sine and Katz Index ROC curves are indistinguishable from one another. For those index β and α were set to 0.05. The dotted line represents the performance of a random classifier.

5 Further Steps

The next steps are to fine-tune the hyperparameters of our models and apply cross-validation to obtain a more robust performance measure. We want to explore other forecasting methods to have a wider view on the effect it has on performance. We will explore new features like the number of patents or job openings liking a company to a technology. Finally, we will use the best model to compute a survivability index for each entity in the network.

6 Conclusion

To the best of our knowledge, our framework is the first investigation that mimics the creative-destruction and the survival mechanisms of innovations within the cybersecurity technological landscape. By modelling percolation dynamics through link prediction, we path the way for further research aiming to compute a survivability score of different entities (i.e., technologies and companies) represented by nodes of a network (i.e., the graph representation of the Swiss cybersecurity technological landscape). We suggest that our framework is useful for decision-makers involved in the security of critical infrastructures, as a survivability score of entities can be extracted by either using the similarity metrics or probability calculations from the supervised learning method.

References

1. Benchettara, N., Kanawati, R., Rouveirol, C.: Supervised machine learning applied to link prediction in bipartite social networks. In: 2010 International Conference on Advances in Social Networks Analysis and Mining (2010)
2. Fleming, T.C., Qualkenbush, E.L., Chapa, A.M.: The Secret war against the United States: the top threat to national security and the American dream cyber and asymmetrical hybrid warfare an urgent call to action. Cyber Defense Rev. **2**(3) (2017)
3. Jang-Jaccard, J., Nepal, S.: A survey of emerging threats in cybersecurity. J. Comput. Syst. Sci. **80**(5) (2014)
4. Kim, J., Kim, S., Lee, C.: Anticipating technological convergence: link prediction using Wikipedia hyperlinks. Technovation **79**, 25–34 (2019)
5. Keupp, M.M.: Militärökonomie. Springer, Wiesbaden (2019). https://doi.org/10.1007/978-3-658-06147-0
6. Kunegis, J., De Luca, E.W., Albayrak, S.: The link prediction problem in bipartite networks. In: Hüllermeier, E., Kruse, R., Hoffmann, F. (eds.) IPMU 2010. LNCS (LNAI), vol. 6178, pp. 380–389. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-14049-5_39
7. Laube, S., Böhme, R.: Strategic aspects of cyber risk information sharing. ACM Comput. Surv. **50**(5), 1–36 (2017)
8. Lundstrom, M.: Applied physics: enhanced: Moore's law forever? Science **299**(5604), 210–211 (2003)
9. Lü, L., Zhou, T.: Link prediction in complex networks: a survey. Physica A Stat. Mech. Appl. **390**(6), 1150–1170 (2011)
10. Piraveenan, M., Prokopenko, M., Hossain, L.: Percolation centrality: quantifying graph-theoretic impact of nodes during percolation in networks. PLOS One **8**(1), e53095 (2013)
11. da Silva Soares, P.R., Prudêncio, R.B.C.: Time series based link prediction. In: The 2012 International Joint Conference on Neural Networks (IJCNN) (2012)
12. Qamar, S., Anwar, Z., Rahman, M.A., Al-Shaer, E., Chu, B.T.: Data-driven analytics for cyber-threat intelligence and information sharing. Comput. Secur. **67**, 35–58 (2017)
13. Tylenda, T., Angelova, R., Bedathur, S.: Towards time-aware link prediction in evolving social networks. In: Proceedings of the 3rd Workshop on Social Network Mining and Analysis. Association for Computing Machinery (2009)