Forecasting Trends in Data Protection and Encryption Technologies using Wikipedia Pageview Statistics

The rapid and profound growth of digital technologies leaves room for more powerful and frequent cyber-attacks[4]. Thus, the goal of this quantitative study is to identify, analyze, and forecast trends related to data protection and encryption technologies. Based on Wikipedia pageview statistics, we study the public interest[1, 3, 2] in 36 technologies that were previously identified by field experts using the Delphi method. We explore the relationships between those technologies by measuring, analyzing, and classifying the time-varying attention (proxied by pageviews) given to each technology. Next, we predict these time series using different forecasting models. As a baseline for our models, we use open datasets, such as arXiv, Google Trends, and Twitter hashtags.

First, we collect data from Wikipedia's pageview statistics. Wikipedia pageview statistics permit downloading for free the number of pages visited over a given period at the chosen frequency, in CSV, JSON, or PNG format. It provides daily, monthly, and yearly data. To balance the velocity and the quantity of data, we use the monthly frequency for all data sources. We gather data from Wikipedia pageviews for the 36 technologies over 82 months, from July 2015 to April 2022. Therefore, this dataset will serve as a measure of global popularity as well as a proxy for public interest in encryption and data protection technologies. Then, we extract the number of scholarly articles uploaded to the arXiv repository. This repository service created in 1991 offers more than 2 million scientific articles related to approximately 170 technical fields and includes more than 3 Teraoctets of pdf. arXiv is open access, so the scientific articles, as well as the metadata (authors, title, sub-category, and publication date), are freely downloadable. The number of e-print on arXiv is used as a measure to quantify the public interest in each of our 36 technologies. Again, we have chosen to calculate this metric with monthly frequency, on a timeline similar to the Wikipedia pageview statistics. We subsequently extract data from Google Trends with the same approach as arXiv. Google Trends only report normalized data, *i.e.*, each data point is divided by the total number of searches per location and period. This index goes from 0 to 100. We collect data from technologies individually to ensure that each technology is not influenced by the others. Following this, we normalize our time series to obtain an equal scale for all our data.

In Figure 1, we depict the estimates of a Dynamic Type Warping that clusters technologies.[1] Out of the 36 technologies analyzed, we note the following implications:

- 12 technologies display a positive trend (C). We interpret this trend as an increase in public interest in well-known growing technologies, such as "Blockchain", "Homomorphic encryption" and "Zero-knowledge proof".
- 5 technologies display a constant trend (B). We interpret it as a stable public interest in technologies, such as "Electronic voting".

[1] We obtain similar results using a K-means algorithm

- 19 technologies display a negative trend (A). We interpret it as a decrease in public interest in well-established technologies, such as "Hash function", "Email encryption" and "Database encryption".

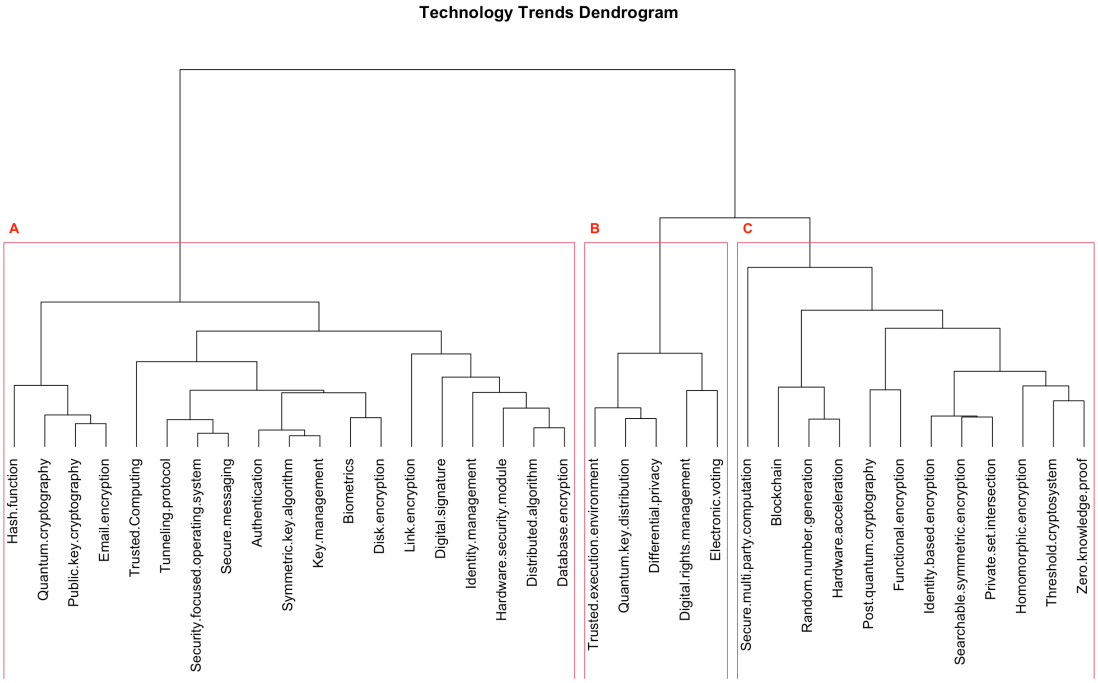

Figure 1: A dendrogram showing the selected 36 technologies classified into three clusters : (A) negative trends, (B) constant trends, and finally (C) positive trends. We use the Dynamic Time Warping (DTW) method to estimate the similarity between two temporal sequences that do not align exactly in time, speed, or length. The observations range from July 2015 to April 2022 and are normalized and decomposed to filter out the noise and potential seasonality and keep only the trend.

Moreover, we find strong and significant correlations between some of our technologies. Generally, we observe positive correlations, for instance between "symmetric key algorithm" and "digital signature", there is a significant correlation of 0.71 at the 5% level. However, a significant negative correlation of 0.53 is also observed at the 5% threshold between "biometric" and "zero-knowledge proof".

Information resources on technology and innovation are often extracted from very specific and scientific data. Here, we study technologies from a wide range of sources, including the general public, which allows us to have a peripheral view of the maturity, the dynamics and current status of each of the selected technologies. Altogether, our findings expand the TechMining literature by providing insight into the public interest in data protection and encryption technologies.

References

[1] Guedes-Santos, J., Correia, R. A., Jepson, P., Ladle, R. J., 2021. Evaluating public interest in protected areas using wikipedia page views. Journal for Nature Conservation 63, 126040.

[2] Kämpf, M., Tessenow, E., Kenett, D. Y., Kantelhardt, J. W., 2015. The detection of emerging trends using wikipedia traffic data and context networks. PloS one 10, e0141892.

[3] Roll, U., Mittermeier, J. C., Diaz, G. I., Novosolov, M., Feldman, A., Itescu, Y., Meiri, S., Grenyer, R., 2016. Using wikipedia page views to explore the cultural importance of global reptiles. Biological conservation 204, 42–50.

[4] Shalf, J., 2020. The future of computing beyond moore's law. Philosophical Transactions of the Royal Society A 378, 20190061.