Chapter 40
Trends in Open Source Software for Data Protection and Encryption Technologies

Lucía Gómez Teijeiro and Thomas Maillart

40.1 Introduction

Software editors and practitioners have increasingly developed and used open-source software tools to implement their cybersecurity strategies. By its unique intellectual property regime, open-source software fosters transparency and sharing values, which have been recognized as important to finding and fixing vulnerabilities and quickly avoiding threats. By selecting 41 technologies related to the one presented in the book, we show that open-source software for cybersecurity is a rapidly growing complex ecosystem of 3456 GitHub repositories with 5000+ users. While some repositories are prominent, many have evolved under the radar, serving niche or emergent needs. Here, we provide the first account of trends in open-source software for cybersecurity and develop a non-parametric forecasting approach to provide an outlook of its development towards 2025.

40.2 Open Source Software and Cybersecurity

Following Eric Raymond's adage, "Given enough eyeballs, all bugs are shallow" [1], key promises of open source software (OSS) have been transparency, task self-selection, and peer-review [2]. In times of increasing economic, social, and political challenges in cyberspace, securing full access to software code has become a critical aspect of digital sovereignty [3]. Organizations face numerous dangers using software they do not control, such as forced technology obsolescence, product

L. G. Teijeiro (✉) · T. Maillart
University of Geneva, Geneva, Switzerland
e-mail: lucia.gomez@unige.ch; thomas.maillart@unige.ch

V. Mulder et al. (eds.), *Trends in Data Protection and Encryption Technologies*,
https://doi.org/10.1007/978-3-031-33386-6_40

discontinuity, and cybersecurity risks. For organizations with short business cycles, such risks are limited compared to the opportunity to use somewhat highly efficient closed-source solutions. However, for critical infrastructures built over decades or more, the risk of not having control over software or hardware code is serious. For instance, the European Organization for Nuclear Research (CERN) has been at the forefront of open-source software and open hardware strategy developments precisely because their research infrastructures take more time to build and operate than the expected lifespan of most technology providers [4].

OSS development, as a community of collective action [5], carries numerous benefits associated with the power of collective intelligence [6, 7]. Those benefits are highly desirable in many cyber-security applications (e.g., hunting vulnerabilities through bug bounty programs) [8]. Moreover, given its short reaction overhead, collective action appears to be a rational response to increasingly time-critical cybersecurity challenges [9].

With an increasing need for transparency and the pressure to ensure continuously reliable systems, OSS for cybersecurity is expected to keep developing as a complement and an alternative to closed source.

40.3 GitHub: A Social Coding Paradigm in Software and Hardware Development

GitHub was established in 2008 [10] as a *social coding* platform based on *git* technology, a distributed software version control system initiated by Linus Torvalds to efficiently track changes in software source code in the decentralized setting compatible with Linux Kernel development [11]. Nowadays, GitHub has become the primary online platform for collaborative OSS development. Here, we studied GitHub repositories associated with data protection and encryption.[1] We found that the number of created repositories increases exponentially (c.f., Fig. 40.1).

The exponential growth of the repository creation rate is expected for data protection and encryption, given that it is a relatively new GitHub platform. In addition, as more OSS code accumulates, the marginal cost of repository creation decreases. Indeed, previous software artifacts can be reused as a complex adaptive network of package dependencies [12], git forks, or simply through code copy-paste.

[1] We investigated 9003 GitHub repositories created since 2008 relating to the 41 data protection and encryption technologies. We collected descriptive data for each repository (description, keywords, README.md) and creation date.

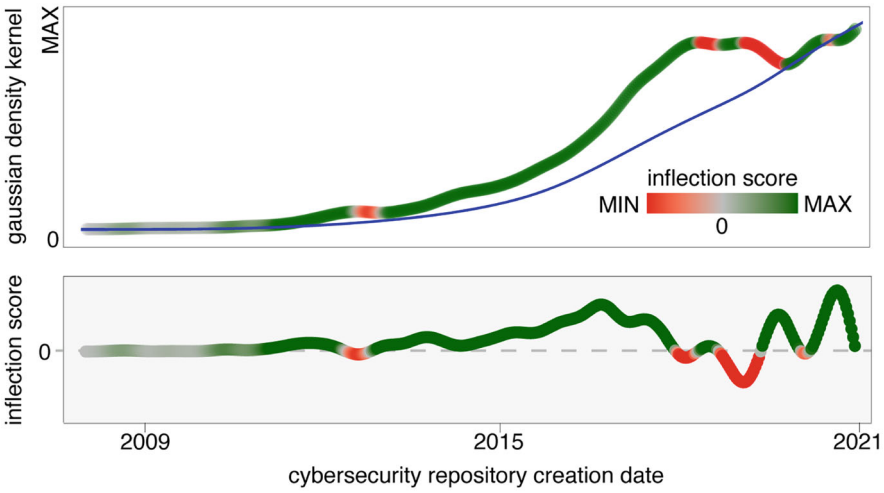

Fig. 40.1 (upper panel) Evolution of repository creations with a color-coded continuous measure of inflection. Repository creation is best fitted by an exponential model (blue curve) with rate $k = 1/\tau = 0.88$ ($p < 0.001$ and $R^2 = 0.88$). (lower panel) inflection score captures the velocity (i.e., the derivative) of repository creations

40.4 Clustering the Complexity of OSS Cybersecurity Ecosystems

When considering OSS ecosystems in data protection and encryption, a significant challenge is to make sense of a complex landscape of repositories covering overlapping topics. Indeed, frameworks used or developed in GitHub repositories are likely to cover several technologies, some more pervasive than others. Figure 40.2 shows how technologies, as queried on GitHub search engine, intersect with clusters of repositories build using non-supervised machine learning on (1) repository descriptions, (2) keywords, and (3) README files.[2] Some technology categories robustly match specific clusters (e.g., digital signatures, symmetric cryptography, blockchain, Web3), while others spread across several clusters (e.g., 0,1,2) thus being less specific.

[2] Text was processed for term frequency-inverse document frequency (tf-idf) word embedding and reduced into a 2D Uniform Manifold Approximation and Projection for Dimension Reduction (UMAP). Communities were detected using Louvain clustering.

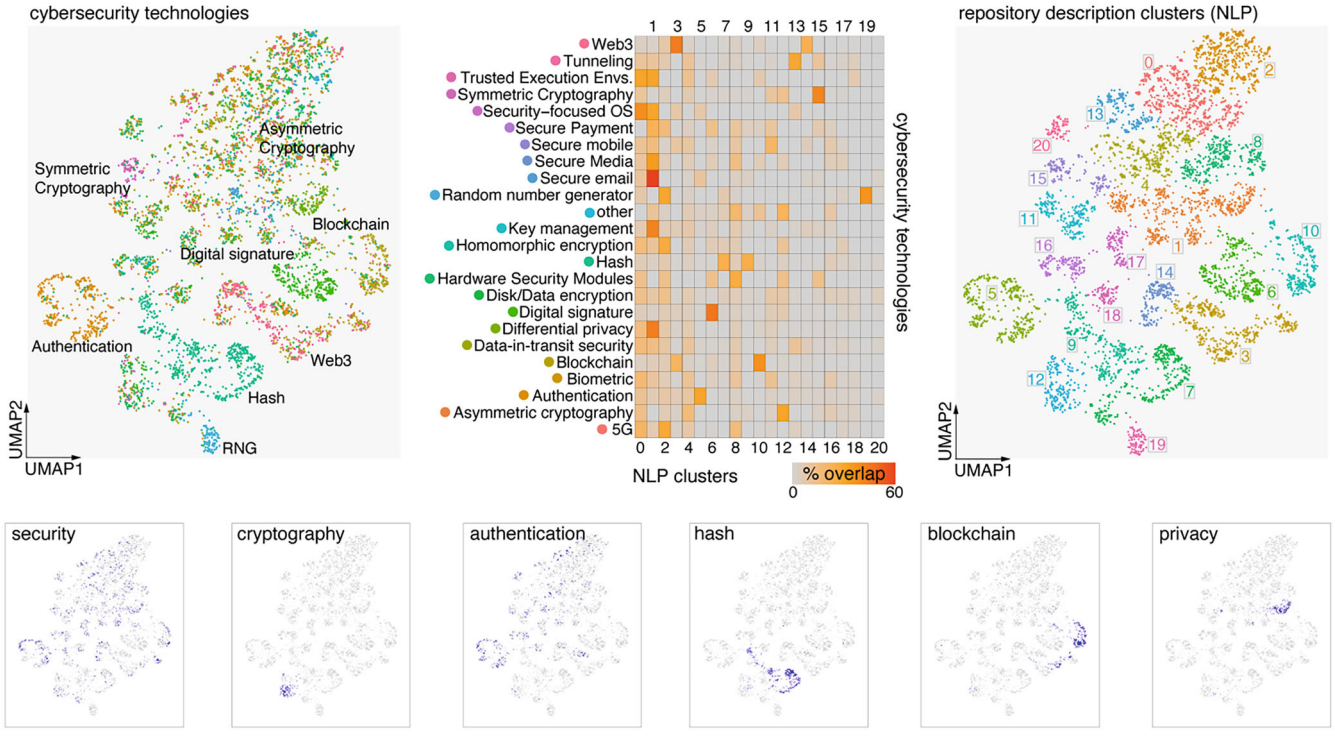

Fig. 40.2 Natural language processing (NLP) embedding and clustering on repository description feature across the data protection and encryption technologies covered in this book. As queried on the GitHub topic search, those categories differ significantly from description clusters generated using NLP. Some categories robustly match specific clusters (e.g., RNG and blockchain), while others spread across several clusters (e.g., 0,1,2), thus being less specific

40.5 Outlook Towards 2025

Monitoring OSS repositories for data protection and encryption technologies is like investigating a hidden giant finally emerging to the light of day: the number of repositories being created has been growing exponentially until now. Some became successful commercial products (e.g., Threema in Switzerland), others became central components of Web security architectures (e.g., OpenSSL), while many are still addressing niche needs. Notably, some of these niches will eventually turn mainstream. Therefore, detecting and monitoring current and future repositories that count, respectively will count, for cybersecurity is critical to identify and harness development opportunities for data protection and encryption technologies, digital sovereignty, and sound business.

Combining long-term exponential growth rates, inflection dynamics, and growth density for each data protection and encryption category, we forecasted their development until 2025. Figure 40.3 shows that forecast until 2025, combined with their historic growth dynamics.[3]

40.5.1 Consequences for Switzerland

Improving OSS monitoring for data protection and encryption is critical for Switzerland. As a small country with limited ability to see domestic tech giants emerge and yet a reputation of safety and reliability, Switzerland's researchers and entrepreneurs have an edge in leveraging OSS ecosystems. One example is Threema, which has built an authoritative secure messaging OSS app. In addition, having full access to software code is crucial for the accountability of solutions provided by the industry and hence, for the cybersecurity of critical infrastructures. Finally, understanding and forecasting future trends in OSS cybersecurity ecosystems is key to assessing and anticipating the evolution of critical data protection and encryption technologies.

[3] Specifically, we fitted and cross-penalized three TES models over creation date dynamics: density kernel, exponential cumulative distribution, and inflection score.

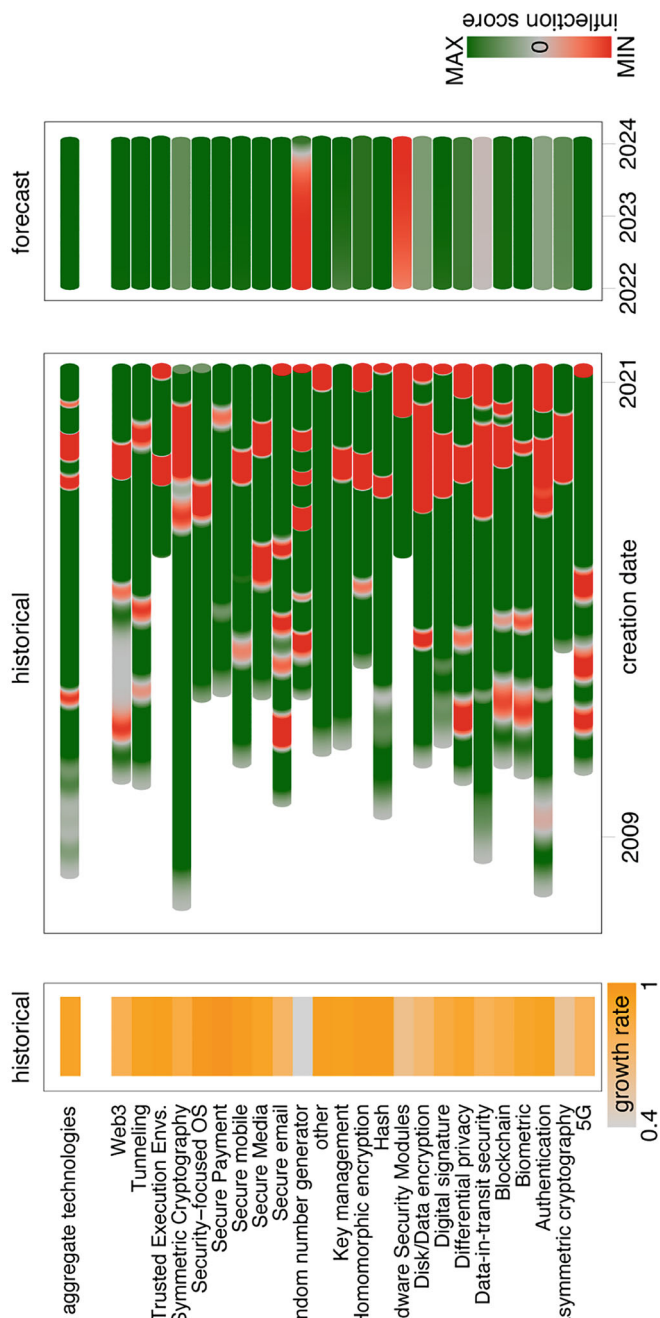

Fig. 40.3 (left panel) Exponential growth rate of repository creation per technology. (middle panel) Evolution of inflection velocity on repository creations over the history of categories. (right panel) Inflection velocity forecast until the end of 2024

References

1. Eric Raymond. The cathedral and the bazaar. *Knowledge, Technology & Policy*, 12(3):23–49, September 1999.
2. Yochai Benkler. Coase's Penguin, or, Linux and "The Nature of the Firm". *The Yale Law Journal*, 112(3):369+, December 2002.
3. Julia Pohle and Thorsten Thiel. Digital sovereignty. *Internet Policy Review*, 9(4), December 2020.
4. Pietari Matti Veikko Kauttu. Open hardware as an experimental commercialization strategy: challenges and potentialities | CERN IdeaSquare Journal of Experimental Innovation. July 2019.
5. Elinor Ostrom. *Governing the Commons: The Evolution of Institutions for Collective Action (Political Economy of Institutions and Decisions)*. Cambridge University Press, November 1990. Published: Paperback.
6. Didier Sornette, Thomas Maillart, and Giacomo Ghezzi. How Much Is the Whole Really More than the Sum of Its Parts? 1 1 = 2.5: Superlinear Productivity in Collective Group Actions. *PLoS ONE*, 9(8):e103023, August 2014.
7. Thomas Maillart and Didier Sornette. Aristotle vs. Ringelmann: On superlinear production in open source software. *Physica A: Statistical Mechanics and its Applications*, 523:964–972, June 2019.
8. Thomas Maillart, Mingyi Zhao, Jens Grossklags, and John Chuang. Given enough eyeballs, all bugs are shallow? Revisiting Eric Raymond with bug bounty programs. *Journal of Cybersecurity*, 3(2):81–90, June 2017.
9. Sébastien Gillard, Dimitri Percia David, Alain Mermoud, and Thomas Maillart. Efficient Collective Action for Tackling Time-Critical Cybersecurity Threats, October 2022. arXiv:2206.15055 [physics].
10. GitHub, October 2022. Page Version ID: 1115484009.
11. Git, October 2022. Page Version ID: 1116314204.
12. T. Maillart, D. Sornette, S. Spaeth, and G. von Krogh. Empirical Tests of Zipf's Law Mechanism in Open Source Linux Distribution. *Physical Review Letters*, 101(21):218701+, 2008.