Data-driven risk quantification for proactive security

Savino Dambra

THESE DE DOCTORAT DE
SORBONNE UNIVERSITE
préparée à EURECOM

École doctorale EDITE de Paris n° ED130
Spécialité: «Informatique, Télécommunications et Électronique»

Sujet de la thèse:

Data-driven risk quantification for proactive security

Thèse présentée et soutenue à Biot, le 14/12/2021, par

Savino Dambra

Rapporteurs	Prof. Juan Caballero	IMDEA Software Institute
	Prof. Tudor A. Dumitraş	University of Maryland
Examinateurs	Prof. Giovanni Lagorio	Università degli studi di Genova
	Dr. Merve Şhain	SAP Security Research
Directeur de thèse	Prof. Davide Balzarotti	EURECOM
Co-directeur de thèse	Dr. Leyla Bilge	Norton Research Group

Résumé

Les données sont l'une des ressources les plus précieuses au monde, à tel point qu'elles sont devenues le nouveau pétrole. Les stratégies fondées sur les données, qui reposent sur des mesures objectives, des faits et des informations tirées de l'analyse des données, sont devenues la norme dans la société et le monde industriel d'aujourd'hui, car elles aident les entreprises à comprendre les besoins actuels du marché, à ajuster les processus de production et à améliorer leur efficacité globale.

La cybersécurité ne fait pas exception. L'analyse des données de télémétrie peut aider les organisations à détecter les cybermenaces, les intrusions non autorisées, les infections par des logiciels malveillants et elle est fondamentale pour une réaction rapide et exhaustive aux cyberincidents. De plus, les données collectées peuvent fournir des informations exploitables et des indicateurs objectifs qui peuvent aider les organisations à prédire leurs risques de cybersécurité et à éviter les événements indésirables en adoptant des mesures proactives. Néanmoins, la faisabilité et l'efficacité des mesures proactives dépendent d'une cascade de défis: comment quantifier les cyber-risques d'une entité, quels indicateurs peuvent être utilisés pour les prédire, et de quelles sources de données peuvent-ils être extraits?

Dans cette thèse, nous énumérons les défis actifs auxquels les praticiens et les chercheurs sont confrontés lorsqu'ils tentent de quantifier les cyber-risques et de les contextualiser dans le domaine émergent de la cyber-assurance. Nous passons ensuite en revue les études antérieures qui se sont penchées sur ce sujet et proposons plusieurs directions de recherche et problèmes non résolus qui nécessitent l'expertise des experts en sécurité des systèmes.

Nous sélectionnons ensuite et commençons à explorer certaines des questions soulevées.Nous évaluons l'incidence de différentes mesures et postures de sécurité sur les risques d'infection par des logiciels malveillants et nous évaluons la pertinence de neuf indicateurs extraits des machines pour étudier la nature systématique de ces risques.

Enfin, nous fournissons des preuves de l'importance de la collecte de sources de données et d'une approche holistique pour la mesure des risques. Nous examinons le 'web tracking' et démontrons à quel point les risques liés à la vie privée sont sous-estimés lorsque l'on tente de les quantifier en excluant la perspective des utilisateurs.

Abstract

Data is one of the most valuable resources in the world to such an extent as to become the new oil. Data-driven strategies based on objective metrics, facts, and insights derived from data analysis have become the mainstream in nowadays society and industrial world as they help businesses to understand the current market needs, adjust production processes and improve their overall efficiency.

Cyber security is no exception. The analysis of telemetry data can help organizations to detect cyber threats, unauthorized intrusions, malware infections and it is fundamental for a quick and exhaustive reaction to cyber incidents. More than that, collected data can provide actionable information and objective indicators that can help organizations to predict their cybersecurity risks and avoid adverse events by adopting proactive measures. Nevertheless, the feasibility and efficacy of proactive measures depend upon a cascade of challenges: how can one quantify the cyber risks of a given entity, what reliable indicators can be used to predict them, and from which data sources can they be extracted?

At first, in this thesis we enumerate active challenges that practitioners and researchers face when attempting to quantify cyber-risks and contextualize them in the emerging domain of cyber insurance. We then go over prior studies that looked at this topic and propose several research directions and unsolved problems that require the domain expertise of system security experts.

We then select and start exploring some of the questions raised by our analysis. We evaluate the incidence that different security measures and security postures have on malware-infection risks and assess the goodness of nine host-extracted indicators when investigating the systematic nature of those risks.

We finally provide evidence about the importance that data-source selection together with a holistic approach have on risk measurements. By looking at web tracking we demonstrate how underestimated privacy risks are when attempting to quantify them by excluding the users' perspective.

Contents

List of Figures

List of Tables

Glossary

AI Artificial Intelligence

AIC Akaike Information Criterion

AV Antivirus

ANOVA Analysis of Variance

BGP Border Gateway Protocol

CCPA California Consumer Privacy Act

CDP Chrome debugging protocol

CISO Chief Information Security Officer

DNS Domain Name System

DoS Denial of Service

EIOPA European Insurance and Occupational Pensions Authority

FMEA Failure Mode and Effect Analysis

GDPR General Data Protection Regulation

GLM Generalized Linear Model

HAZOP Hazard and Operability Analysis

HTTP Hypertext Transfer Protocol

ICT Information and Communication Technologies

IT Information Technology

IoT Internet of Things

IoU Intersection over Union

IP Internet Protocol

ISO International Organization for Standardization

ISP Internet Service Provider

PPI Pay-Per-Install

PCS Platform Control System

PUA Potential Unwanted Application

PUP Potential Unwanted Program

SoK Systematization of Knowledge

SRG Software Reliability Growth

TPR True Positive Rate

VPN Virtual Private Network

Chapter 1

Introduction

The modern society is highly dependent on Information and Communication Technologies (ICT). It is practically impossible to imagine what life would be like without technology, as almost everything we do involves the use of an electronic device and requires an internet access. As the use of technology increases and becomes multi-fold in every sector, so does the volume of raw information that it produces, and that is collected and stored: according to a recent report [97], we create 2.5 billion gigabytes of data every single day, with 90% of that data in the whole world being created in the last two years alone. Not surprisingly, data is one of the most valuable resource in the world —to such an extent as to become the new oil [100]— and the reason for that is simple: transforming raw data into meaningful information can yield to a plethora of valuable insights for businesses in order to monitor their performance, meet goals and reach objectives. Data-driven strategies based on objective metrics, facts and insights gleaned from data to drive strategic business decisions have become the mainstream in nowadays society and industrial world. For example, data analytics have paramount importance to understand the current market landscape, change production processes and trigger the creation of new products that match the market needs; in the same way, the knowledge gained beforehand of what customers want makes marketing campaigns easier and better oriented, as well as the customization of advertisements to address a specific segment of an entire customer portfolio; data analytics can also pave the way for other potential profitable opportunities or for solving problems and improving the overall operational efficiency.

Cyber security is no exception. The massive amount of data generated through the everyday use of software, collected by monitoring tools on enterprise networks, produced on the web by users from all over the world, and further scrutinised by data analysts can serve the purpose of detecting cyber-security threats and raising the alarm when a security incident occurs. In this respect, data analysis can help organizations in detecting potential frauds, unauthorized intrusions with unusual network traffic patterns, malware infections, hardware failures, data leaks and security breaches. An efficient data monitoring system is fundamental for an early threat identification, damage prevention, business interruption limitation, and for a quick *reaction* to adverse events.

In addition, data can provide actionable information and objective indicators that can help businesses to predict their cyber-security risks, and switch from a reactive to a *proactive* approach. Proactive cyber security is nowadays gaining importance and current risk management methodologies have been complemented by adding incident avoidance and prevention

strategies to enhance existing reactive measures. While in case of incident the latter establish guidelines about determining the level of the threat, assessing the amount of the damage, and installing measures to prevent such an event from reoccurring, proactive measures provide a more holistic approach to securing Information Technology (IT) systems, focusing indeed on prevention and avoidance rather than detection and response. Proactive cyber-security measures include spotting the most vulnerable parts of a system, forecasting those with the highest likelihood of being attacked, protecting the most valuable assets, prioritizing security spendings for hardening areas or people at higher risk, and even transferring part of this risk to third-party entities by underwriting cyber-insurance policies. Nevertheless, the goodness and feasibility of proactive measures depend upon a cascade of challenges: how can one quantify the cyber risks of a given entity? What reliable indicators can be used to predict them? And from which data sources can they be extracted?

1.1 Problem statement

Cyber-risk quantification is still an open problem among researchers and practitioners. All the frameworks and methodologies that are currently in use are limited to listing the best security practices, providing a score based on potential hazardous events that can materialize with a certain likelihood, and simulating adverse scenarios to assess subsequent damages. Unfortunately, these solutions only have a *qualitative* foundation and base their analysis, assessments, and consequently their results on metrics based on experts knowledge and their previous experience. On top of that, very few and scattered studies exist in the systems security community that tackle *quantitative* cyber-risk estimation: as a result, we still lack rigorous and reproducible ways to understand which security risks can even be predicted in the first place, which features are most useful for such prediction and from where they can be extracted. A first goal of this thesis is to raise awareness in our community towards these problems and present the main challenges of quantitative risk estimation and proactive security. Although nowadays those are very debated topics in the security field, their practical applications are very limited and the insights gathered from prior research struggle to be included in security products. For this reason, we decided to present those thematics by contextualizing them in an emerging area where those are crucial: cyber insurance. Indeed, when underwriting cyber policies, insurance carriers have to necessarily transform information about the prospective clients in a score that reflects their risk classes with the purpose

of computing premiums. In this respect, cyber-risk quantification and the capability to predict future events have vital importance for the success or failure of the whole cyber-insurance market itself. By currently relying on qualitative methodologies, risk assessment techniques in this in domain are desperately in need of data-driven solutions which measure the risk exposure by leveraging objective indicators and overcoming estimations based on experts opinions.

The identification of good indicators becomes then paramount for reliably predicting cyber-risks. Despite a few prior studies exist in this direction [208, 70, 69, 235, 61, 259, 79, 216, 68], there are still several research areas that have not been explored, aim at predicting different cyber events, focus on specific aspects or narrow their scope to particular environments. Indeed, the effectiveness of methodologies, indicators and information sources strictly depends upon the problem under analysis: for example, predicting the risk of experiencing malware infections requires data sources and a series of indicators that may differ from those needed for data-breach incidents. The second goal of this thesis is to look at the malware infection risks of machines used in two distinct environments (i.e., those used at home and in enterprises) and evaluate the systemic and systematic nature of those risks. Our goal is to understand the incidence that different security postures and cyber hygienes have on malware infections together with evaluating the predictive power of nine host-based attributes extracted from real-world telemetry.

Finally, cyber-risk quantification and the research of powerful indicators are two tasks that cannot be carried if not supported by reliable data sources. More than that, obtained results can also be misleading if not considering a holistic perspective in which insights are gleaned by approaching risk analysis from different points of view that are subsequently aggregated. Therefore the last objective of this thesis aims at showing how risk measurements from a novel perspective and by using distinct data sources can provide new insights and also challenge long-held beliefs. We operate in the domain of web-tracking and our goal in this case is to demonstrate how underestimated privacy risks are when their quantification occurs by excluding the users' perspective.

1.2 Contributions and thesis outline

This thesis is centered around the importance of data-driven measurements for the purpose of quantifying cyber-security risks. The manuscript is organized into 6 chapters.

Following this introduction (Chapter 1), in Chapter 2 we provide the reader with the necessary background knowledge to understand the main contributions of the thesis. The chapter starts from the description of a classic insurance process for the purpose of identifying its main phases and actors, and then clarifies the differences and peculiarities when the same scheme is applied to the cyber domain. We also discuss previous studies that attempted to correlate indicators extracted from telemetry data to the risk of cyber incidents and malicious software encounters. We then cover previous works that quantified the extent of web tracking by using different types of telemetry.

In Chapter 3 we report our Systematization of Knowledge (SoK) centered around the system aspects of cyber insurance. In this contribution, we conduct a comprehensive literature review with the aim of categorizing previous research on the topic, presenting existing challenges to cyber-security researchers, and listing future research directions and unsolved problems which require the domain expertise of system security experts. Chapter 3 is based on the publication *SoK: Cyber Insurance - Technical Challenges and a System Security Roadmap* [94] presented at the *IEEE Symposium on Security & Privacy (S&P) 2020*.

In Chapter 4, we start exploring some of the open challenges presented in Chapter 3. In particular, we conduct an exploratory study on risk indicators of malware encounters by analyzing real-world telemetry and by differentiating between home machines and those used in enterprise environments. We assess whether the different security measures adopted in either parties have a role in reducing malware infections. We then continue with verifying if machine attributes such as the days and hours of activity together with the volume of generated files, the number of installed software vendors, the recidivist infected state, and the geographical location can provide valuable insights as risk indicators of malware infections. We also discover that a portion of this risk is systematic and does not depend on the adopted security measures. Chapter 4 is based on the paper *A Comparison of Systemic and Systematic Risks of Malware Encounters in Consumer and Enterprise Environments* currently under submission at the *ACM Transactions on Privacy and Security (TOPS)*.

In the subsequent Chapter 5, we focus on the importance that the selection of data sources together with a holistic approach have on risk measurements. We analyze the practice of web-tracking and look at it from the users' perspective, whereas previous studies only looked at it from the trackers' one. We perform a correlation analysis to understand what increases users' web-privacy risk, estimate how long it takes to encounter a signifi-

cant fraction of the trackers, and argument on the interesting relationship among privacy and security risks on the web. We also quantify what percentage of the user's browsing history is known to trackers and investigate how much this knowledge could be extended through real or hypothetical collaborations among different tracking companies. Our point of view allows us to conclude that privacy risks due to web tracking are higher than what estimated in the past from other perspectives. Chapter 5 is based on the publication *When Sally Met Trackers: Web Tracking From the Users' Perspective* that will be presented at the *31st USENIX Security Symposium (2022)*.

In the last Chapter 6, we conclude this thesis by summarizing the valuable findings and insights gleaned from our work and outlining future research directions to explore.

1.3 Ethical considerations

Two out of the three contributions presented in this thesis (Chapters 4 and 5) analyze datasets derived from human subjects. We analyzed the datasets by collaborating with the researchers of NortonLifeLock [46], a popular security company. The datasets contain the telemetry collected by the company's Antivirus (AV) sensor installed on Windows machines. The data only includes users who voluntarily installed the product, accepted the company's privacy policy [176], and opt-in to share their data. As privacy advocates, we treat each piece of information —from its collection, throughout its analysis, to its storing— in a way that preserves the customers' privacy and identity. The customer identifier is anonymized on the client-side and sent in this form to a central system: in our works, we observe users only through numeric anonymized identifiers, that do not contain any detail or endpoint attribute able to trace back to their origin. We never deanonymize users by singularly inspecting their telemetry and we only look at aggregated data.

Chapter 2

Background

This chapter discusses background information to allow the reader to contextualize the thesis. We start by giving an overview of a classic insurance process scheme and then discuss differences and peculiarities of cyber insurances (Section 2.1). Afterwards, we focus on prior studies that investigated the risk of malware infections and evaluated the goodness of indicators for predicting cyber-security events (Section 2.2). We finally discuss past research that attempted to quantify web-tracking risks and the extent to which this practice impacts users (Section 2.3).

2.1 From insurance to cyber insurance

Insurance is a risk management method whose main purpose is to convert the risk of harmful events into an expenditure. As depicted in the diagram reported in Figure 2.1, the insurance process includes several interacting components and generally involves two players: a first supply-side entity who provides insurance, named insurer or insurance company, and a second demand-side entity who buys the insurance, known as insured or policyholder. The two parties interact in two different phases, respectively identified as underwriting (or policy stipulation) and claiming for compensation. During the drafting of a policy, an insurance carrier needs to acquire useful information about the prospective client with the purpose of identifying his risk class. Afterwards, the two parties need to clearly define the conditions, circumstances, and nature of the events that are covered by the policy. Coverage can encompass both first- and third-party losses: while the former is purchased to cover the policyholder against damages or losses suffered by the insured to his person or property (e.g., health, disability insurance), the latter is intended to protect the policyholder against liability for damages or losses caused by the insured to other people or their property (e.g., bystanders hit by insured's car in an accident, stranger's properties damaged by a fire that comes out of insured's house). At this point, the insurer quantifies the material damage that the insured — or third subjects if considered — would be subjected to if these occurrences were to happen. Finally, the insurance company takes on the liability and management of such situations cashing a premium payout from the insured.

The management of client portfolios is another crucial task insurance companies need to consider during the underwriting phase. The goal is typically to maintain a pool of policies, each of them having an independent probability of claim. This diversification averts catastrophic scenarios in which a single incident impacts a large fraction of the clients: in such cases, a significant number of claims would be submitted at the same time and

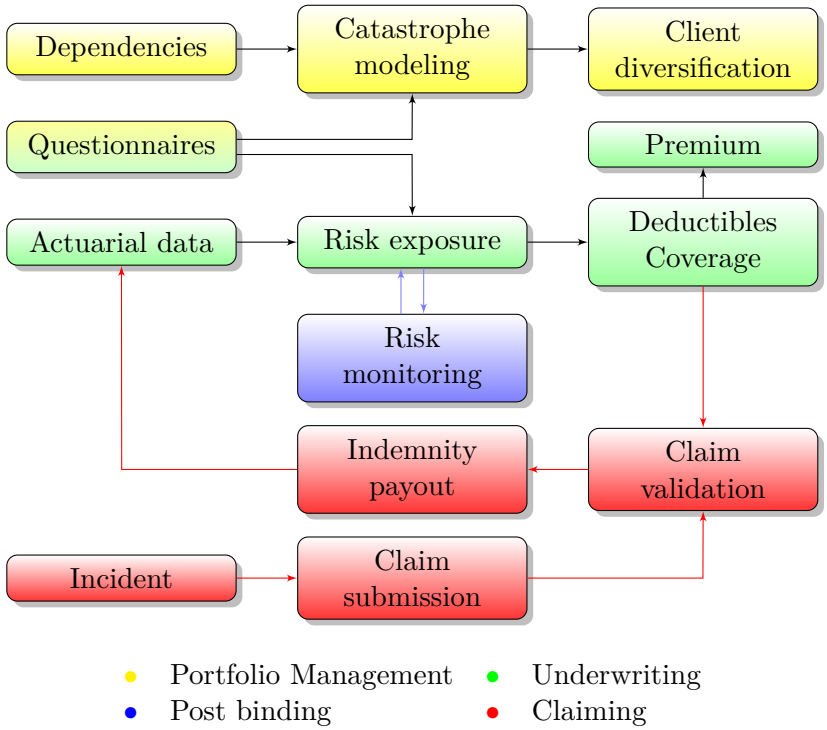


Figure 2.1: Classic insurance process workflow extended in a cyber scenario

the insurance would suffer a huge blow in covering losses. For instance, it may not be a good strategy for an insurance company to insure against fire hazards all apartments located in the same building.

Finally, when experiencing losses due to an incident which is potentially covered by the insurance policy, the victim submits a claim to the insurer who makes sure of its validity, assesses the impact of the event and compensates the claimer with an indemnity determined according to the terms of the policy. The contract can also include a deductible, i.e., an amount for which the insured is liable on each loss.

In order to make this entire process possible, the insurer must carefully set its tariffs to ensure that the premiums collected are enough to cover future claims, in addition to yield profit for the insurance firm itself. Unfortunately, this is anything but easy. Indeed, when it comes to selling a finished product or service, a firm can easily determine its price knowing which costs have been incurred for its realization. On the contrary, an insurer who places its product on the market does not know in advance the amount of money required for claim compensations because of their

inherent uncertain nature. In this respect, actuarial techniques allow to estimate these disbursements and overcome the cost uncertainty related to this inverted production cycle. A key element for this estimation relies on statistical methods that study how claims for covered events have evolved over the previous years to forecast their future evolution. Thus, the raw information required to build a classic insurance product consists of a large set of historical records containing claims and compensations for events which have similar characteristics to the ones being insured. Insurance firms usually do not rely only on their own data sources but also take advantage of the market statistics that aggregate historical data of other companies in the same domain. This statistical information, which normally goes under the name of *actuarial data*, is what allows an insurance company to estimate the risk of a certain event or client, given a number of relevant contextual information (acquired during the underwriting phase). This includes, for instance, the driver's age and neighborhood for a car insurance or the age of the building in a house insurance.

Extending Insurances to the Cyber Domain

With the help of Figure 2.1, we now look more closely at how the previous process is applied to the cyber domain by discussing the differences and the main challenges that affect each insurance phase.

Underwriting – As we discussed above, the policy underwriting requires the insurer to collect information from the client that can be useful for the purpose of risk assessment. Following a traditional model, also in the cyber domain this is still performed by a mix of self-assessment questionnaires, checklists, business documentation, meetings, and interviews [243, 226, 52, 50, 130, 86], whose objective is to identify the adopted software and technologies, the deployed security measures, the presence of sensitive data and how it is stored and processed, and any other information that can affect the global security posture of the company under investigation [199, 109]. A deeper analysis can be carried out to tailor the product to the specific customer based on its characteristics and requirements: a monitoring software equipment together with an overhaul of preexisting security logs and telemetry serve this purpose. Finally, some deficiencies and precautions are often advised to the client in order to comply with the best-known security practices [16].

Assessing the cyber risk of organizations or individuals is an overly challenging problem due to a number of reasons including the existence of asymmetric information, the dynamic nature of the cyber ecosystem, and the indirect risk that might be propagated from the relations with the

third parties. Although with the traditional meticulous risk assessment methodologies the underwriters could draw an approximate picture of the customer's risk exposure, they might not be aware of the residual risks that might be known to the counterparts. The possession of a greater material knowledge by one of two parties involved in an economic transaction creates the problem of carrying asymmetric information and this represents a major issue in cyber insurance [58, 196, 60, 217, 65]. A risk assessment that is made by analyzing asymmetric information can lead to adverse selection [48, 116]. For example, unfair risk scores might be assigned to a company whose private and inaccessible information may reveal a severe exposure to risk compared to another with a better security hygiene.

The existence of asymmetric information also impacts negatively the customer side as insurance firms may raise premium prices due to incomplete knowledge and risk overestimation, leading to an expensive, niche, and not-appealing product [238, 160]. High premiums are also the result of insufficient criteria to reduce them: even if a company holds security certifications and profusely invests in self-protection, the effectiveness of these actions against the wide variety of cyber attacks is not clear, making, in turn, difficult to assess to what extent they are useful to reduce the overall risk [238]. A timid step in this direction is the one of some carriers that reduce premiums or deductibles if the client uses risk assessment tools, security technologies, and breach response services of specific vendors [253].

The interdependent nature of the cyber ecosystem makes the risk estimation even more complicated. Nowadays, when cloud computing and outsourcing are two mainstream phenomena, cyber risk is intertwined among all entities that depend on one another [196, 64, 65]. Companies may indirectly get damaged because they use external services that are targeted by a cyber attack: an example is the recent Denial of Service (DoS) attack against DynDNS – which impacted more than sixty of its customers [250]. Thus, a firm's measures and expenditures in self-protection may not proportionally increase its security level when making use of services from third parties that do not invest as well [154, 65]. In the pre-binding phase, risk exposure must be then identified from a holistic standpoint, preferring a due diligence approach to a simple checklist and including in the review all internal and external threat vectors that could potentially compromise pre-insured's security [75].

Actuarial and Pricing – The actuarial approach based on statistical models described above does not fit the cyber domain where historical data of claims and compensations are still scant [196, 118, 112, 48, 238, 60]. Enterprises experiencing a cyber incident have a strong incentive not to publicly

disclose it as this would tarnish their image. As a result, the few available databases [9, 22, 19] contain records which are often vague, missing details, and biased towards large and serious incidents, whose disclosure is unavoidable due to their resonance or due to mandatory-notification laws [20, 18]. The infeasibility of the actuarial approach alone for an accurate risk estimation is corroborated by its ever-evolving components: cyber threats and attack methods swiftly evolve alike defense methods and strategies do [154, 196, 109].

Portfolio Management – As briefly discussed before, a fundamental requirement of traditional insurance schemes is that the insurer should strive to obtain a portfolio of policies with an independent probability of claim submission. This diversification can reduce the likelihood that a single incident could harm a considerable portion of clients – a *catastrophic event* that can have severe consequences and cause the bankruptcy of the insurer [236, 76, 77]. Unfortunately, it is harder to obtain such diversified portfolio in the cyber domain, due to the monoculture of software and hardware products [110, 63, 225, 213]. Although deploying different configurations is possible, recent events have shown that the business continuity of a large set of possible clients – independently of their size, sector, and assets to protect – is simultaneously undermined when a piece of a broadly-used software or hardware is found to suffer from a severe vulnerability [119, 256, 171, 93, 71, 252, 134, 149].

In other domains, a common way insurers protect themselves against catastrophic events such as wildfire and hurricanes is by purchasing policies from other insurance companies. Sadly, the current lack of re-insurers in the cyber domain further exacerbates this problem [108, 196, 53, 170].

Post-Binding Phase – Due to the complications in both the policy underwriting and claiming phases, an additional post-binding phase is introduced, which does not exist in other forms of insurance [75]. In fact, in traditional insurances, the relationship between the firm providing coverage and the policyholder ends once the contract has been signed and the two parties interact again only in case of a claim submission. On the contrary, a cyber insurance may require periodic risk assessment after the underwriting is completed, to allow the insurer and the policyholder to collect updated information related to new threats and evolved risks. Indeed, many cyber-insurance policies already bring supplemental value through the inclusion of risk mitigation, tracking and loss-prevention tools [54]. Clients, in particular small organizations that lack experience, can benefit from this continuous interaction to better ponder their measures towards higher-priority situations [160]. The post-binding phase also helps to prevent the well-known issue of moral

hazard [154, 56, 48, 237] — a form of post-underwriting opportunism by the policyholder, who undertakes incautious actions knowing that, in case of incidents, there exists a counterpart who will bear the brunt and will not be able to verify the presence of negligent and fraudulent actions. In this regard, insurers have to conduct continuous risk assessments to resize the set of inaccessible information of the insured and mitigate its unfair behaviors.
Claim Submission and Validation – Cyber-insurance policies usually cover the costs of incident response and forensic investigations, including the identification of stolen or compromised data and the extent to which third parties have to be informed according to the current regulations. Despite this, a precise quantification of the involved and compromised assets is complicated by their intangible nature [196, 48]. In addition, since jurisdictions may apply different notification laws, each case must be accurately evaluated according to the localization of the indirectly-damaged third party.

The insurer as well compensates for economic losses related to the event. In particular, cyber insurance may refund losses due to business interruption caused by an attack, as well as cyber extortion and stolen assets. This approach is insufficient in the cyber scenario where the above primary losses are often followed by secondary ones that result from a loss of reputation whenever the incident is publicly disclosed [58].

Time is also a key component when it comes to claim submission. Some attacks may silently compromise a system and remain undiscovered for a considerable time-frame. The validity of claims in such situations is a more arduous issue to formalize in cyber policies. Furthermore, carriers may require forensic investigations prior to claim submission to verify its validity, resulting in an initial disbursement from the insured and a reputation damage due to the disclosed incident.

2.2 Malware-Infection Risks

In the following sections, we look at previous studies that have explored the malware infection landscape in the context of either home-users or organizations. In the second part, we cover those works that have correlated indicators extracted from telemetry data to the risk of cyber incidents and malicious software encounters.

2.2.1 Infection landscape

Many industrial reports published by security companies [40, 39, 41, 36] provide an annual summary of the malware families observed in the wild, the

number of compromised machines, the losses due to data breaches, and the malicious campaigns that targeted different organizations. However, those documents only focus on statistics, do not carry a detailed analysis, and do not make any distinction between consumer and enterprise environments.

In the scientific community, scattered studies leveraged network telemetry or internal logs provided by ad-hoc software to delineate the status and the evolution of the malware landscape. Kotzias et al. [137] analyzed a 3-year-long collection of internal data from 28K enterprises to shed light on their vulnerability patching behaviors and existing threats. The investigation carried out by the authors shows a higher prevalence of malware with respect to potentially unwanted programs (Potential Unwanted Program (PUP)), the presence of more secure and affected industrial sectors, and the fact that the patching of server applications is much worse than the one on the client-side.

Two studies focused on the trends of malware that spreads through Pay-Per-Install (PPI) Services [74, 136]. Caballero et al. [74] built an infrastructure and deployed it in 15 countries to interact with 4 PPI providers. The authors found that 12 out of 20 of the most prevalent families of malware employ PPI services and that this distribution mechanism is more common in richer countries. The follow-up paper narrowed the analysis down to PUP families that spread through PPI services, performing a systematic study of their prevalence using AV telemetry. The results indicate that PUPs are installed on 54% of the considered machines and that up to 25% of them are distributed by a limited number of publishers.

2.2.2 Risk indicators

In recent years, an increasing number of studies have tried to identify risk indicators i.e., measurable features collected from external sources or internal telemetry, that can be correlated with the risk of suffering from cyber incidents. Some of them also applied the features they identified to train prediction algorithms and assess the prevalence of those risks in the future.

Yen et al. [259] used internal telemetry logs of a large organization to spot risk indicators that are correlated to malware encounters. The authors showed that user's demographic features, as age and job title, together with network-related features, such as the frequent use of untrusted internet connections and longer browsing sessions, are effective at predicting which users are more at risk of malware infections. RiskTeller [61] is a prediction tool that leveraged internal data of 18 enterprises to predict which of their machines will be at risk of being infected by a broad spectrum of malware classes. Its classification accuracy reaches 95%, showing that the identified

features are strongly correlated with the likelihood of malware encounters. Liu et al. [151] studied the extent to which cyber security incidents can be predicted by using observed malicious activities associated with network entities, such as spamming, phishing, and scanning. The study shows that the resulting classifier is able to produce fairly accurate predictions over a forecasting window of 2-3 months. The same authors also attempted to predict the likelihood of organizations to suffer a cyber incident by using an algorithm that only uses externally observable features [150]. The authors trained a classifier by combining signs of network mismanagement, such as misconfigured Domain Name System (DNS) or Border Gateway Protocol (BGP), with malicious activity time series, such as spam, phishing, and scanning activity sourced from these organizations. Despite 10% of false positives, the prediction reaches 90% accuracy, suggesting the possibility of forecasting an organization's breach without internal information. Thonnard et al. [235], discussed organization and individual-level features that are likely to reflect the risk of experiencing targeted attacks. The authors identify enterprise sizes and public profiles of individuals as potential risk factors and show that there exists a degree of correlation with receipt of targeted attacks. In a similar way, Sarabi et al. [208] built a predictor using a set of industry, business and web visibility/population information. The results demonstrate how, and to what extent, these externally-observable features can help forecast an enterprise's relative risk of experiencing different types of cyber incidents.

Fewer prediction studies exist on the consumer side, probably due to the lack of telemetry data for this segment of users. Lévesque et al. [147], performed a 4-month study by collecting real-usage data of 50 subjects and monitoring both user behaviors and possible infections. Using neural networks, the authors developed a predictive model with 80% accuracy at predicting the users' likelihood of being infected. Canali et al., [79] assessed to what extent a user's web browsing behaviors can be used to predict her risk class. The results show how particular types of user actions, such as browsing the web late at night and during weekends, considerably affect the risk exposure. Finally, by leveraging mobile users' browsing patterns and self-reported data, Sharif et al. [216] tried to predict whether users will encounter malicious pages on a long and short term. With an overall accuracy of 87% True Positive Rate (TPR), this work shows how useful on-the-fly predictions can be in protecting users from malware distributed on the web.

2.3 Web-tracking risks

In the previous sections, we discuss prior works that highlight the challenges faced by practitioners and researchers when attempting to quantify cyber risks in the cyber-insurance domain. We also discuss past studies that attempt to predict security risks and adverse events. Since the goal of our last contribution is to also cover privacy risks arising from web tracking, in the following section we provide an overview of this practice and discuss related work that attempts to quantify its extent on the users.

The first tracker, based on a cookie from `digital.net` in `microsoft.com`, was used in 1996 and discovered by an 'archaeological' study conducted by Lerner et al. [145] in 2016 by using the Internet Archives Wayback Machine [124].

The first analysis regarding web tracking was performed in 2009 by Krishnamurthy and Wills [138], where they examined the different technical ways in which third-parties could obtain user-related information. Three years later, the work from Mayer and Mitchell [159], and Roesner et al. [197], helped to lay the foundations for future studies. More recent studies showed that an increasingly larger percentage of the most popular websites include some form of tracking, and that they use a variety of techniques to do it [104, 204, 224, 125].

Olejnik et al. [180] were among the first to use real-user data to study web tracking. The authors discovered that 69% of the users in their dataset had a fingerprint that could differentiate them from the rest based on their web history. This study was recently replicated by Bird et al. [62] with 52K Firefox users, and found an even larger number, with 99% of them showing unique patterns. Falahrastegar et al. [107] also used the web history of real users to check whether user-specific IDs were being sent in requests: authors found this to be very common between certain groups of domains. Vallina et al. [241] performed instead a study based on network traffic of a mobile carrier to check not only the presence, but also the efficiency of the ecosystem based on energy consumption. They found that tracking is very widespread but the delivery strategy is inefficient.

During the last years, the number of works based on real-user data has increased. In 2018, Karaj et al. [131] performed a large-scale study using the information gathered from a browser extension. They calculated some general stats about the different trackers found online, and open-sourced the corresponding global results obtained from the dataset. At the same time, Papadopoulos et al. [185] presented a study focused on mobile devices. By using the data collected from 1,270 users, the authors quantified

the economical cost of showing ads for companies, and the corresponding privacy loss by the users that receive them. The final results indicate that there is a clear imbalance between the two, with the users paying the highest price. The following year, Papadopoulos et al. [184] expanded their idea and analyzed the concept of tracking cookie synchronization by using another dataset of 850 real mobile users. They found that 97% of the users are actually exposed to this type of practices in the first week of browsing. Most recently, the work from Hu et al. [121] leveraged real-world browsing histories to measure the prevalence of different tracking organizations in UK and China. Authors discovered that there is a big difference in the companies involved, with home-grown third-party operators in China, and US players dominating the UK market. Finally, Mishra et al. [164] studied the relevance of the Internet Protocol (IP) information in the web tracking ecosystem, analyzing the information received from 2,230 users. Results indicate that IP-based tracking is still a viable, as 87% of the participant retained the same address for multiple days.

In summary, many papers attempted to analyze web tracking risks using different types of telemetry, but they centered their work on very specific cases such as user identifiers [107, 184] or web history uniqueness [180, 62, 241]. Despite finding many interesting results, these studies lack a global overview of the phenomenon and a holistic view on the impact that this has on users.

Chapter 3

Technical challenges of cyber-risks quantification in the context of cyber-insurance

3.1 Introduction

This chapter aims at providing an extensive discussion of the technical aspects and open challenges of quantitative cyber-risk estimation contextualized in the cyber-insurance domain.

Our study starts by presenting a systematization of the existing literature along four main axes: risk management techniques and frameworks, game theoretical methods, economic theories, and studies that attempted to predict cyber security events. We clearly organize each contribution and point out which part of the cyber-insurance puzzle it tries to address.

In the second half of the chapter, we introduce four main research areas where we believe that expertise in computer security can support the cyber-insurance domain. This includes risk prediction, automated data collection, catastrophe modeling, and computer forensics. Each section identifies the technical challenges and emphasizes a number of concrete future research directions.

We then conclude this chapter by reporting a few studies that followed ours and that contributed in one of the four research areas that we considered in our systematization.

3.1.1 Why this study matters

Despite the considerable and increasing investments in IT security products [26], it is well understood that cyber attacks cannot be prevented by technical solutions alone and the protection against all possible threats is neither possible nor economically feasible. To proactively defend against adverse events and handle the residual risk that cannot be mitigated with existing measures, organizations are rapidly moving by incorporating cyber insurance into their multi-layer security frameworks. Cyber insurance is defined to be the way to transfer the financial risks related to network and computer incidents to a third party [65]. Compared with traditional insurance policies for business interruption and crime, a cyber-insurance policy can also cover, for instance, digital data loss, damage and theft, as well as losses due to network outages, computer failures, and website defacements.

As evinced by recent market reports, the adoption of cyber insurance has tremendously increased over the last decade, achieving an annual growth rate of over 30% since 2011 [31]. This is also reflected in the growing number of claims submitted for cyber incidents in a wide range of business sectors [34] and that, in few striking cases, have seen insurance companies paying even hundred-million-dollar indemnities [35]. Following this trend, the cyber-insurance market is forecasted to reach 14 billion USD in gross

premiums by 2022 [191] and several indicators confirm this direction. First, cyber crimes have never been so profitable [174] and the growing number of attacks is increasing the awareness of board members about cyber risks and the impossibility of only relying on preventive solutions [228]. This pushes a growing number of companies, among which even more small- and medium-size enterprises, to start considering cybersecurity insurance as a risk mitigation strategy: in fact, data show that 66% of them would need to shut down if hit by a data breach [175]. Another strong driver for the cyber-insurance domain is the introduction of global regulations on personally identifiable information loss, such as General Data Protection Regulation (GDPR) and California Consumer Privacy Act (CCPA). For instance, the need to cover fines and the high cost of handling user notifications are already creating interest in purchasing cyber insurance [249].

This shows that, while researchers and security experts are still debating whether cyber-insurance schemes even make sense and how they could be better implemented, insurance companies are already selling them as part of their portfolio. However, companies are currently struggling against the demand of cyber policies as existing tools and methodologies to quantify risk exposures and pricing are inadequate in the cyber domain. Although past studies have concluded that, without considering catastrophic scenarios, the vast majority of cyber risks are insurable [28, 27, 60], carriers are missing solid methodologies, standards, and tools to carry out their measurements. The result, as we will comprehensively detail later in this chapter, is that purely *qualitative* assessment of such risks leads to inaccurate evaluations, not properly tailored to the customers but mainly based on averages for their industrial sectors [38].

Research-wise, the main aspects, the evolution, and the core challenges of cyber insurance have been studied for more than two decades [154]. Sadly, while researchers have extensively looked at the theoretical aspects of the cyber ecosystem, there exists a very limited number of studies that relied on real data and leveraged the domain expertise of system security experts for cyber-risk quantitication [150, 208, 235, 70, 69, 61, 259, 79, 216, 68].

This chapter aims at covering this gap by providing an extensive discussion of the technical aspects and open challenges in the cyber-insurance domain, emphasizing how security experts can contribute to this rapidly evolving area. We believe the cyber-insurance field raises many technical questions that require the expertise of system security researchers: how can one identify and collect low-level risk indicators and compare them with externally-observable events? Is it possible to automatically extract dependencies among different software and services and capture the risk in-

troduced by the supply-chain of a company? Our main goal is to present a thorough discussion on these problems such that researchers understand that to work properly cyber insurance will require practical solutions that go well beyond its economic and game-theoretical aspects.

3.2 Cyber-Insurance literature systematization

Since its first appearance in the late 90s [154], cyber insurance has been the focus of researchers from different disciplines. For our study, we selected and analyzed 93 works among academic papers, standards, and frameworks. As shown in Figure 3.1, we grouped these works in four main categories and fourteen sub-categories. In particular, we found that previous research has mainly focused on two areas: cyber risk management, which tries to estimate attack probabilities and possible damages, and mathematical modeling and game theory simulations, which aim at deriving interesting properties on the consequences of cyber-insurance adoption. Two additional areas complete the picture: research conducted by the economics community reporting figures from past incidents or discussing the costs of possible scenarios, and research focusing on the prediction of future cybersecurity events.

Since these four macro categories refer to very different research domains, we adopted distinct criteria to select and present the contributions from each of them. *Risk management* is a very wide topic that covers a wide range of domains, ranging from pharmaceutical products to natural disasters. We reported all methodologies and frameworks that are currently used in IT (Section 3.2.1), together with those academic papers presenting risk aggregation techniques. Regarding the contribution from the *economics* community, as an exhaustive discussion would be out of scope for a security conference, we focused on the papers needed to emphasize research problems, existing tools, and on the major findings that can affect the work of security researchers. For this reason, we comprehensively reported all of the attempts made in quantifying economic losses following cyber incidents (Section 3.2.3). The works employing *mathematical modeling and game theory* have already been deeply analyzed by Marotta et al. in [156]. Therefore, in Section 3.2.2, we decided to offer a different systematization that focuses on which property the authors were interested to prove, along with the choice of the simulation parameters —e.g., the market model, the presence of asymmetric information, and the network topology. Finally, since our ultimate goal is a call to actions for security researchers to provide data-driven solutions for the cyber-insurance domain, our study comprehensively presents and compares previous *prediction* attempts in section 3.2.4.

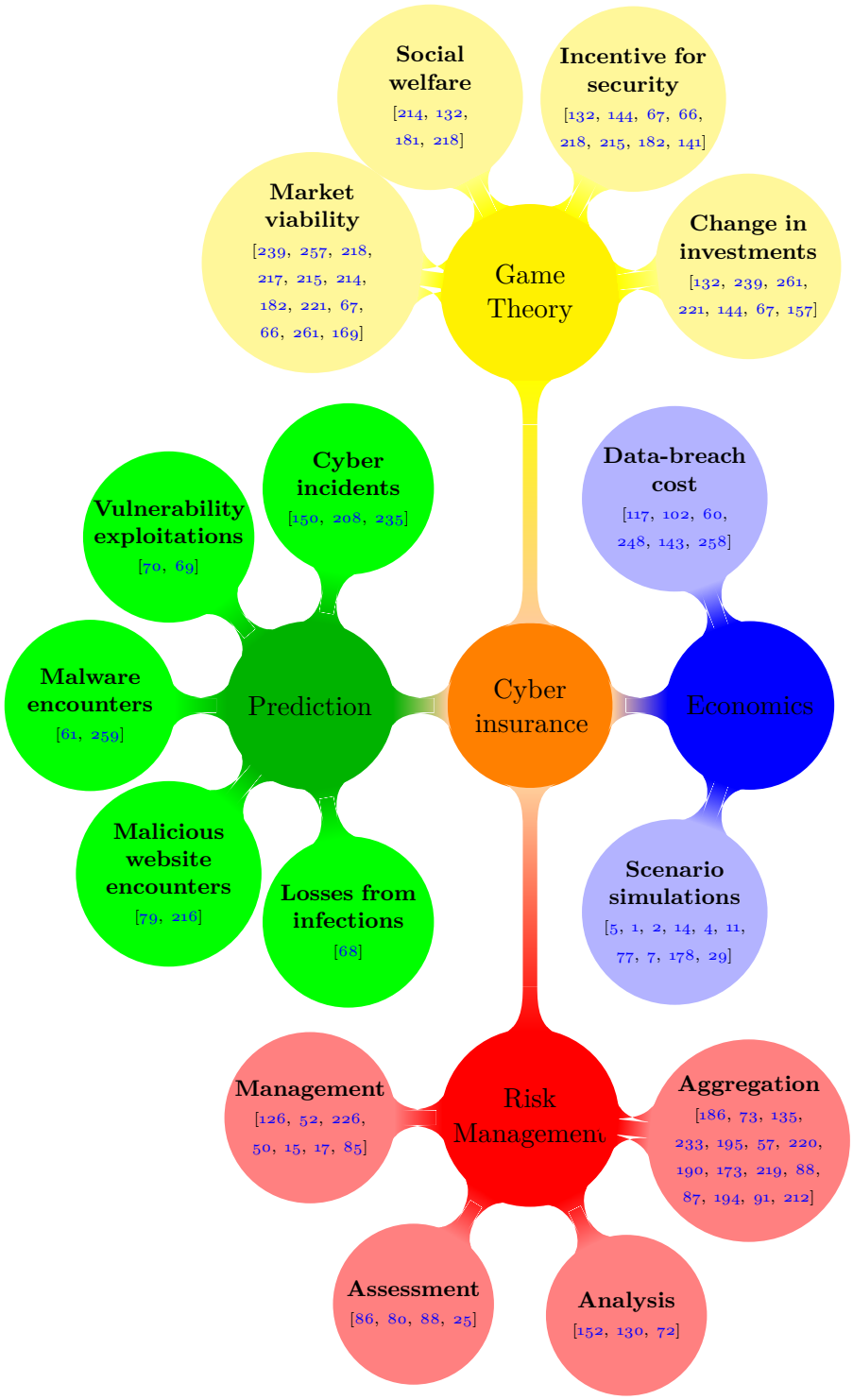


Figure 3.1: Cyber-insurance research areas

3.2.1 Approaches and techniques for cyber risk management

According to the International Organization for Standardization (ISO) standard 31000, a risk management process can be described as a set of tasks whereby it is possible to measure the risk and subsequently develop strategies to monitor and control its evolution [127]. As a result, the first phase of risk management is dedicated to the identification of the valuable assets and of the related threats that represent the main components of risk. Each threat is then analyzed by evaluating its likelihood and possible impact from both a qualitative and quantitative perspective, and results are then aggregated to obtain an overview of the whole risk. These two phases, grouped and referred to as *risk assessment*, are usually followed by a *risk treatment* step, which covers the choice of non-exclusive countermeasures that can be adopted to tackle each of the risk components. Finally, as risks may suddenly change, causing the previous estimations to become incorrect and countermeasure ineffective, a number of *risk monitoring* and reviewing actions are required to continuously update the risk estimation.

Risk management is an important process when it comes to information technologies. Therefore, the literature is rich of guidelines, frameworks, and techniques that contextualize it to the digital world. As depicted in Figure 3.1, we grouped under the risk management sub-category the studies that provide a walk-through of the entire procedure [126, 52, 226, 50, 15, 17, 85], defining terms and providing a helpful documentation of how to address issues on risk assessment and treatment, as well as insights on risk monitoring and reviewing. Other works often inherit or revisit a previous risk management methodology and introduce new techniques to implement a specific sub-component. In this respect, we created two different subcategories in which we respectively list the works addressing the whole risk assessment [80, 25, 86, 88] and those narrowing down the discussion on risk analysis[152, 130, 72].

Although widely used standards (such as ISO 27005 [126]) and tools (e.g., NIST SP 800-30 [226], Magerit [52], OCTAVE Allegro [50], Clusif [86] and the one proposed by Microsoft [17]) handle the single stages of the management procedure in a different way, they share a common underlying workflow for assessing individual risks. According to it, the process typically starts by brainstorming which and how cyber-based threats could prevent the company from reaching business goals and team objectives. In this respect, real-life cyber events previously occurred to other companies can be used as source of inspiration. In addition, frameworks often provide guidelines on how to identify this collection, including checklists or questionnaires, and advising to adopt a what-if approach to understand what could

go wrong and what the possible consequences are. The outcome of this process is the creation of a *risk register*, whose structure, together with some examples, is reported in Table 3.1. Once each row has been filled with a description of the threat, including its possible triggers and effects, the impact and likelihood of its materialization are assessed to define the *inherent* risk. Two approaches exist for scoring these factors and the choice of one rather than the other depends on the company itself. Indeed, some tools provide a table of decipherable words with a *qualitative* description, whilst others opt for a *quantitative* numerical sliding scale (e.g., Table I3 and H3 of [25]). It is worth pointing out that the same event could be assigned different values across distinct situations: if an organization's public statement is *"we have built our reputation on our commitment [...] to protect the privacy and confidentiality of personal information"*, the impact of user-data leaks for this company will be higher if compared to another one with different prerogatives. The next step is the identification and mapping of existing mitigations or controls the could reduce the likelihood of each threat: companies often take advantage of existing frameworks that list critical checks and best practices, and indicate the extent to which the control environment reduces the inherent risk. As a result, a value reflecting the *residual* risk is obtained and a three-fold choice opens up: if the value falls within the company's risk appetite limit, no further action is needed in this phase. If not, more controls and mitigations have to be investigated or the residual risk has to be transferred to a third party —e.g. with a cyber-insurance policy.

Finally, in the last sub-category of Figure 3.1, we reported all the methodologies that have been proposed to *aggregate* and propagate individual risks based on tools that capture the relationships among different information components or requirements of an attack. These modeling tools make use of graph theory or model checking to draw conclusions starting from some preconditions. Among them, attack trees are widely used techniques to capture dependencies among threats [186, 73, 135, 233, 195, 57]. Each tree is a leveled diagram made of nodes, leaves and a root; each node represents an attack or a threat which materializes only if all its children are satisfied. The root attack is completed if all nodes are satisfied. Similarly, vulnerabilities or exploits are represented as nodes in attack graphs and conditionally linked to each other according to their preconditions and results. Such composition of vulnerabilities is used to simulate incremental network penetration and attack likelihood propagation with the purpose of measuring the overall security of a system or network [220, 190, 173, 219]. Finally, Hazard and Operability Analysis (HAZOP) [88] and Failure Mode

and Effect Analysis (FMEA) [87], are other two techniques used to break down a complex process into small sections and reason about possible undesired situations, their causes and consequences. Such kind of tools are mostly employed when the use of ICT can introduce a series of hazards in industrial environments [194, 91, 212]. As we will discuss later in this chapter, these methodologies, inherited from other domains, can be unsuitable when employed in cyber scenarios.

Description	Cause	Effect	Inherent Impact	Inherent Likelihood	Inherent Risk	Residual Impact	Residual Likelihood	Residual Risk
Third person gains access to sensitive customer information via stolen credentials	Employee inadvertently inputs access credentials within the source code	1 million customers at risk of identity theft. Company receives significant criticism for its privacy preserving policy	Catastrophic	Possible	High	Catastrophic	Remote	Medium
Sensitive customer data exposed to unauthorised parties	Employee deliberately copied full customers records motivated by personal financial gain	1 million customers at risk of financial theft	Catastrophic	Remote	Medium	Catastrophic	Extremely Remote	Low
Remote code execution on webserver by unauthorised parties	Zero-day vulnerability exploited in third-party library used for customer authentication	1 million customer data at risk of theft. Online platform not available to customers. Business-continuity interruption	Catastrophic	Possible	High	Catastrophic	Possible	High

Table 3.1: Risk register: qualitative assessment examples for inherent and residual risk

3.2.2 Cyber insurance and Game Theory

A large portion of existing contributions employ mathematical modeling and game theory to infer properties and effects of adopting cyber insurance. As comprehensively reported in [156], this approach allows in the first place to create a mathematical model of cyber insurance which takes into account its main actors (insurance carriers, policyholders and regulatory entities), their interdependences (probability of infection and externalities), the network topology (independent nodes, complete graph, random graph, or others) and the market type (competitive, monopolistic, or oligopoly). Once a model has been defined, game theory is used to simulate the behavior of agents: insureds choose their desired level of protection and contract type, insurers instantiate contracts, and regulators come into play by imposing regulation options (mandatory insurance, fines, bonuses, penalties, mandatory investment, etc.). The use of game theory makes it possible to also include in the models the major issue of information asymmetry in its moral hazard and adverse selection forms. This way of tackling cyber insurance is very useful for strategic purposes and allowed researchers, practitioners, and governments to reason about consequences and peculiarities of its employment, and market viability.

Viability of the cyber-insurance market – As already discussed, the starting point of each simulation is the definition of a mathematical model of cyber insurance that considers its main aspects, e.g. market type, type of coverage, existence of asymmetric information, network topology, etc. Therefore, an important finding of each simulation is to verify whether the market defined by such pre-conditions may exist or not, i.e., whether the actors would opt for the insurance case over the non-insurance one. One way to achieve this result is the comparison between the average utility function for agents with ($E[U^I]$) and without ($E[U^N]$) insurance: in the economic theory, this function measures the welfare or satisfaction of an entity from consuming a certain number of goods. Then, if $E[U^I] \geq E[U^N]$ holds, the choice of an insurance policy directly contributes to increase the wealth of an agent [156]. Almost all previous works —among which we find the more realistic settings that include a competitive insurance market, non-zero-profit carriers, the presence of moral hazard and adverse selection, and a partial coverage whose level is defined by the policyholder— fall in this category [239, 257, 218, 217, 215, 214, 182, 221, 67, 66, 261]. Only two studies found that actors who decide not to invest in a cyber policy would benefit from this choice [257, 169]. Yang and Lui [257] concluded that cyber insurance is not a good incentive for all nodes when modeling a competitive market with zero-profit carriers only offering full coverage and

accepting asymmetric information in its moral-hazard form. Naghizadeh and Liu [169] simulated instead a monopolistic profit-neutral insurer, acting as a regulator that imposes fines and grants rebates, and found that this leads to a market failure because of agents would not voluntarily purchase any insurance.

Consequences of cyber-insurance employment – Among the main topics of interest in this area, we find the use of cyber insurance as an incentive for internet security [132, 144, 67, 66, 218, 215, 182, 141], the change in self-protection investments when insurance is available [132, 239, 261, 221, 144, 67, 157] and its contribution for reaching the social welfare [132, 181, 218, 214]. These studies concluded that cyber insurance is not a good incentive for internet security in presence of a competitive or monopolistic market and asymmetric information in its moral hazard form [144, 66, 67, 218, 217]. On the other hand, researchers also concluded that a non-competitive cyber-insurance market can increase internet security if fines are imposed by regulation entities and policy are carefully designed. When analyzing the effect of employing cyber insurance on self-protection, some works show that, if insurance is available, agents prefer not to invest in self-protection, but rather in insurance contracts[113, 129]. In this case, minimal investments imposed by regulators do not change the results. Finally, the usefulness of insurance as a tool to reach the social welfare and the optimal level of self-protection investments has not been yet understood: different studies [215, 179] reached contradictory conclusions on this topic although considering the same preconditions, probably because of adopting different network topologies —which lead to different interactions among actors— throughout their simulations.

3.2.3 The Economics Perspective

Since cyber attacks are often considered inevitable events, cyber experts are increasingly focusing on their economic consequences [1]. In this respect, scenario-based evaluations are a very common approach used to serve two main purposes. For a company, these scenarios provide a useful way to assess the possible consequences of a cyber event [11], to measure the incident response capabilities [4], and to identify the critical systems, people and premises that are needed to continue to serve their customers [2]. On the insurance carriers side, simulations based on scenarios are often used to estimate the financial impact of large-scale attacks or catastrophic events that hit many businesses at once [14]. This simulation practice is rapidly gaining popularity due to the current cyber landscape, in which the costs of recovering from particular types of attacks are way greater than the cost required

to prevent them [5]. Furthermore, tests can help companies to emphasize the presence of valuable data to protect and shed light on interconnected risks that could lead to catastrophic events [1]. Good evidence of this can be found in the decision of the European Insurance and Occupational Pensions Authority (EIOPA) to include, for the first time in 2018, cyber scenarios in the collective insurance stress test used to assess cyber-risk response [6].

The creation process of a scenario-based simulation goes through a multi-stage procedure [77] and it is usually performed by C-Suite executives due to their expertise in business-critical roles and operations [1]. The process starts with the creation of a plausible scenario, defined by a footprint of events to be simulated and a contagion mechanisms among the involved entities [7]. There is a wide range in the type of scenarios that can be used for different applications. For instance, scenarios can be based on historical or synthetic events, they can be generic or specific for a given company or sector, and they can consider single or multiple events [178]. Scenarios allow the simulation of both common digital incidents —like data exfiltration, cyber extortion, denial of service attacks, financial transaction compromise, and cloud service provider failure — as well as rare events — such as cyber-induced fires in buildings or industrial plants, cyber theft of marine cargo, cyber attacks on power grids, or oil rig explosions due to Platform Control System (PCS) compromise [29].

Developing a scenario is a challenging task as it is not easy to fully understand all the systems involved and predict the possible cascading effects that could be triggered [77]. For this reason, developing a *coherent* scenario is a key aspect for successfully achieving the second phase of the simulation that consists of estimating the inducted losses to a business or the impact of claims submitted to an insurance company by taking into account its client portfolio.

The output of the simulation can be further extended beyond a single company by taking into account macroeconomic consequences too [7]. This result can be achieved by selecting a representative subset of the whole population of companies from a wide range of different business sectors and use them to estimate the losses of a given scenario. In turn, this allows for a quantification of the effects on many variables of the global economy [77].

Besides scenario-based simulations, other economic studies attempted to gain insights into cyber risks by leveraging publicly available data. For instance, Eling and Loperfido [102] analyze statistical properties of a data breach information database to show that data breaches significantly differ among each other, hence they must not be put in the same basket but must be mapped to separate risk categories. Using another dataset of publicly

available survey data, Herath et al. attempted instead to build a pricing model for cyber-insurance premiums with the robust copula methodology [117]. Premiums for first-party losses due to virus intrusions are computed with a probabilistic model based on three factors: the occurrence of the events covered by the policy, the time from the issue of the policy to the incident, and the indemnity paid by the insurance in case of the breach occurring. Biener et al. [60] analyzed the world's largest collection of publicly reported operational losses to draw empirical conclusions on whether cyber risks are insurable or not based on Berliner's criteria. Results suggest that cyber risk owns some peculiarities that undermine its insurability, such as its evolving nature, the lack of actuarial data and reinsurance, the severe information asymmetries, the limited coverage and caps, and the high deductibles and premiums for small and medium enterprises.

Wheatley et al. [248] statistically modeled a 15-year cyber-breach dataset to show that the size of an organization is strongly coupled with the frequency and severity of breaches, and the number of information leaked during such events is expected to double within five years from two to four billion items. The handling and response costs of two data breach events are at the center of the study by Layton et al. [143]. Counterintuitively, the authors show that none of the two incidents negatively affected the company stock price and economic growth, secondary and intangible losses have negligible importance with respect to direct losses, and policy and procedure for handling the event have a large effect on the overall cost. On the contrary, in [258] security breaches are found to negatively impacting stock quotation of the victims, especially in the case of e-commerce firms and DoS attacks.

Year Paper	Predicted event	Ground truth		Features	Feature datasets	
2015 [150]	Cyber incidents	Incident reports	Ext	Mismanagement signs Malicious activities	Scanning tools Public scan data	Ext
2015 [208]	Cyber incidents	Incident reports	Ext	Website statistics Industry sector Size Region Popularity	Information services	Ext
2001 [70]	Vulnerability incidents	Incident reports	Ext	Exploit release timing	Vulnerability database	Ext
2010 [69]	Vulnerability exploitation	Vulnerability reports	Ext	Vulnerability features	Vulnerability reports	Ext
2015 [235]	Targeted attacks	Mail scanning service	Int	Industry sector Size Employees features	Industry classification Linkedin Internal telemetry	Int Ext
2017 [61]	Malware encounters	AV Telemetry	Int	Binary file appearance	Internal telemetry	Int
2014 [259]	Malware encounters	AV Telemetry	Int	Demographic VPN logs Network logs	Internal telemetry	Int
2007 [79]	Malicious websites	AV Telemetry	Int	Browsing behaviors	Internal AV service	Int
2018 [216]	Malicious websites	Website Blacklist	Ext	Browsing behaviors Self-reported data	Mobile ISP tracking data User questionnaires	Int
2009 [68]	Losses from malware infection	User questionnaires	Int	Routine Activities Deviant Behavior Guardianship	User questionnaires	Int

Table 3.2: Works on prediction

3.2.4 From risk assessment to risk prediction

So far, a considerable amount of studies, frameworks, and methodologies have focused on assessing the risk of cyber attacks by explicitly defining their underlying causes and triggers. In fact, as we show in Table 3.1, the first column of each row specifies either the particular action, the vulnerability or the exploit that makes the risk materialize. While this *assessment* technique is well established in other domains (e.g., industrial and financial), its effectiveness is still unclear in a cyber scenario. Indeed, if the whole evaluation is based on the *current* knowledge of vulnerabilities present in the system and tools, and on the exploits available to the attackers, it quickly becomes clear that the final measurement has limited lifespan, as new ones are respectively discovered and released on a daily basis. Moreover, when major cyber incidents occur, its root causes and enabling factors are almost always *unknown* to the community, greatly complicating the assessment of the associated risk.

The goal of *prediction* is to overcome this assumption and carry out the risk estimation by leveraging a combination of *risk indicators*, measurable factors that have been empirically proven to reflect the risk across a number of experiments. For instance, back to Table 3.1, lower age, frequent use of untrusted internet connections, and longer browsing sessions at night have been found to be good signs for predicting which users are more at risk of malware infections [259]. And this is done by mentioning none of their actions or incautious behaviors — e.g., the user clicked on a malicious banner or installed malicious software. In a similar way, companies with misconfigured DNS services and expired certificates more frequently show signs of botnet activities, otherwise less likely to be observed in other entities where those are correctly set up [150].

These measurable indicators are merely correlated and not the cause itself of the risk, the same way as the driver age is not the cause of car accidents. But by measuring these signs, experts can make predictions of the likelihood of future events.

Over the past two decades, few scattered studies have focused explicitly on the problem of predicting security-related events. In 2001, Browne et al. proposed a simple formula to predict the amount of security incidents, as a function of time, related to a known vulnerability [70]. Bozorgi [69] used instead publicly available vulnerability databases to predict which, and how soon, a vulnerability is likely to be exploited in the future. In 2005, Schechter [211] looked at the challenges of predicting cyber attacks. He discovers that experts had a much better understanding and success in modeling traditional crimes, such as home burglary [114] while "*attempts to*

bring the quantitative approaches of insurance and risk management to the measurement of [computer] security risk have failed". The author concluded that this is due to the fact that we still lack techniques to measure the security strength of a piece of software (we will get back to this idea of predicting risk through measuring security in Section 3.3.1).

Another traditional way to predict future events is to adapt *Software Reliability Growth (SRG) models* commonly used by the reliability community to describe (typically through a non-homogeneous Poisson process) and predict the evolution of defects in a software artifact. For instance, Condon et al. [89] show that specific classes of computer incidents (such as those that depend on particular vulnerabilities) can be modeled with an SRG, while the total aggregated incident rate can be better approximated by using time series [90].

In 2016, Edwards [101] found that the daily frequency of data breaches can be described by using a negative binomial distribution and used this model to estimate the likelihood of similar incidents in the future. Maillart [153] found instead that the theft of personal information follows a power-tail distribution that is robust independently of the sector and size of the targeted organization.

On a different but related topic, a large corpus of works aimed at predicting the occurrence of new vulnerabilities in software products [262, 210, 222, 246, 115, 51]. However, as we will discuss in Section 3.3.1, it is still unclear how this information can translate to a prediction or the likelihood of being attacked or compromised in the future.

In recent years, prediction techniques have been at the center of few works for the purpose of assessing the risk in different circumstances. In 2009, Bossler et at. [68] investigated the influence of different factors in predicting data losses from malware infections by conducting a survey over 788 college students. More recently, Liu et al. [150], by using a set of external observable features, attempted to predict the likelihood of an organization to suffer a cyber incident in the future. The authors achieved, overall, a 90% accuracy with 10% of false predictions. Cyber incidents are considered also by Thonnard et al. [235], who discussed organization- and individual-level features that are correlated with the risk of experiencing spearphishing attacks. In a similar way, Sarabi et al. [208] build a predictor for cyber incidents using a set of industry, business, and web visibility/population information. RiskTeller [61] is a prediction tool that leverages internal telemetry data to predict which machines are at risk of being infected by a broad spectrum of different malware. Its prediction accuracy reaches 95%, showing that such tool could be used to prioritize security spending towards

machines at higher risk of infection. The same conclusion is reached by Yen et al. [259], who use logs from an antivirus software to infer the risk for hosts in a large enterprise to encounter malware.

On the consumer side, Canali et al. [79], assess to what extent the risk class of a given user can be predicted based only on his web browsing behavior. The authors show how certain types of user actions considerably affect their risk exposure. In a similar way, Sharif et al. [216] use mobile users' browsing patterns complemented with self-reported data to predict whether the users will encounter malicious pages on a long and short period of time. In the latter case, on-the-fly predictions within a browsing session could be useful to proactively prevent malicious-content exposures. All these prediction efforts are summarized in Table 3.2, alongside the type of predicted events, the source of ground truth information, the adopted features, and the data from which they are extracted. The table also shows if the ground truth and the predictive features are extracted from internal sensors (`Int`) or are measured from public external information (`Ext`). We will return on the importance of this aspect in Section 3.3.2.

Finally, few studies have focused on predicting the cost of cyber incidents and data breaches. In this area, Jacobs [128] proposed a regression model based solely on the number of user records compromised. Romanosky [198] introduced more variables (including the revenue and company type) and found that a 10% increase in firm revenues is correlated with a 1.3% increase in the cost of an incident. The author also noted that the price is ultimately related to the size of the company and the size of the breach, and not to the malicious nature of the incident or its outcome.

3.2.5 Discussion

Nowadays, cyber risk management methodologies, results of game theoretical studies, and scenario-based simulations are key components for the development of the cyber-insurance market. In the first case, companies and individuals that want to adopt cyber insurance can take advantage from the existence of these frameworks and guidelines, despite the fact that they were not designed with the insurance market as ultimate goal: indeed, risk management plays a very important role to estimate attack probabilities and possible damages, allowing, in turn, individuals and companies to reason on their needs for a cyber policy. Insurance carriers as well use these tools during contract underwriting for assigning a value to a certain entity's risk and compute premiums for cyber-insurance policies [199].

Unfortunately, all available solutions discussed above have a *qualitative* foundation and base their analysis, assessments, and consequently their

results on metrics based on experts knowledge and previous experience, missing a feedback from real-world experiments and measurable quantities. Existing methodologies rely on checklists, worksheets, knowledge basis, catalogues, tables, and what-if reasoning for identifying threats and hazards. The value of this type of analysis largely depends upon the quality of the used documents and the experience of the experts who brainstorm about undesired events and their effects. In the same way, the use of tools to capture dependencies among threats such as fault trees or the outcome of HAZOP and FMEA studies also assumes that who carries the analysis has detailed knowledge about the areas, operations, and processes that may be exposed to hazardous events and conditions.

The absence of objective measures and the qualitative nature of these methodologies make it also harder to obtain an actual value for the likelihood of a given threat in a cyber scenario: threat probability is, in fact, a key component for assessing risks and, although simulations can approximate the frequency of popular attacks found in the wild, the limitations discussed in the actuarial paragraph of Section 2.1 exacerbate the quantification of such quantity.

Finally, since a sheer number of risk assessment methodologies exists, it is still unclear which one fits best the cyber domain and provides the most precise way to compute the likelihood of cyber incidents. This aspect is further exacerbated by the ever-growing adoption of Internet of Things (IoT) devices, for which new risk metrics and specific risk evaluation methods are still missing [193]. Very similar considerations apply to simulations based on scenarios, as their creation, refinement, and precision to capture the intricate relationships among different entities depends completely on qualitative opinions of expert users and C-suite members.

As risk management methodologies and scenario-based tests, game theory applied to cyber insurance can provide important practical insights. Nevertheless, all conclusions obtained from these studies are purely based on mathematical modeling, with all the limitations that this implies. First of all, the finiteness of modeling can lead to a huge difference between the actors considered and their actual number. Moreover, when using game theory to simulate the behaviors of clients and carriers, players can undertake a limited set of actions and interact with each other only in pre-defined ways, defined by assumptions respectively on the market type and network topology. Unfortunately, there is no measurement or comparison with real-world data that confirms the validity of models and veracity of game theory results.

3.3 Challenges and Open Research Directions

As we described in the previous section, research in the cyber-insurance domain has mostly focused on theoretical studies (from a mathematical viewpoint) and on the analysis of the costs/benefits tradeoff (from an economics viewpoint). At the same time, the system security community has instead been largely ignoring this emerging area. This could be simply the consequence of the lack of interesting problems that require novel and practical solutions, or it could be due to the lack of awareness from our community towards these problems. As we believe the latter to be true, we now focus on some of the areas where researchers' experience with system and network security can play a fundamental role to help the development of the cyber-insurance domain. The contribution of system security researchers can help the development of quantitative, data-driven methodologies, and it can bring automation and support tools to replace questionnaires and qualitative estimations.

In particular, we selected four classes of problems, one for each of the insurance phases: actuarial, underwriting, portfolio management, and claim validation. For each class, we underline the limitations in the current approaches, discuss the challenges of proposing new solutions, and outline a number of open research directions for researchers in the security field. To ease their identification, we tried to mark the main open problems we discuss in the text as $\langle \mathbf{R}n \rangle$.

3.3.1 Area 1: Risk Prediction

"They could tell you exactly the chance of an office building burning down in Midtown Manhattan, but there isn't anyone on this planet who could tell you the probability of a large U.S. retailer being hacked tomorrow"

– Graeme Newman, Director at CFC Underwriting [3]

Cyber-insurance providers employ underwriting tools to collect the information required to differentiate the risk across all the applicants [199]. Today, underwriting questionnaires ask a number of questions which insurance companies believe to be relevant to classify the risk of a potential customer. However, as we discussed in Section 3.2.4, researchers still have to identify reproducible ways to estimate risk based on a number of observable features that had been proven to be meaningful predictors across a number of experiments. The experiments conducted to date were often inconclusive and difficult to compare as they were all conducted on different datasets and none of them was ever repeated or validated by other studies.

As a result, as a community we still lack an understanding of *which security events can even be predicted in the first place, and which features are most useful for such prediction.* This opens several research directions to explore different methodologies to capture and aggregate risk factors.

Measure the security posture of the target. One of the first ideas that comes to mind to understand the risk of cyber incidents is to look at the overall security of a given target. In fact, the security posture of an organization may provide good insights on the level of risk – *if* we assume that a better security hygiene can lower the risk of future attacks. Indeed, at least intuitively, the higher is the security of a system, the lower should be the probability of a security incident affecting that system. If we accept this assumption, risk prediction can be re-formulated as a problem of *measuring security.*

While the fact that security countermeasures could result in a reduced amount of computer abuse was first assessed in 1990 by the seminal work of Straub et al. [227], the link between security posture and cyber risk is not so straightforward and it is still poorly understood today. Security measures can certainly raise the bar for the attackers, but risk also depends on the number of attacks a target may receive—which could be higher for large and popular organizations. Moreover, relevant targets may attract more sophisticated and motivated adversaries, which can make prediction more complicated. But even if we accept this premise to be correct, there are still two serious obstacles to this approach.

First, despite almost four decades of attempts, it is still unclear whether a way to *quantify* security even exists [244]. For instance, in 2009 Verende et al.[244] surveyed many techniques taken from the economics, the computer science, and the reliability community, but still found unclear the validity of the existing results. Second, even if we had a scale to precisely measure security, it is still unknown what is the exact relationship between the level of security and the probability of incidents ⟨**R1**⟩. Simply saying that more security equals less risk is too vague to be practical. Does doubling security reduces the risk by half or by a factor of four? Does the curve reach a plateau, after which adding more security does not provide a tangible reduction in terms of risk?

Measure the behavior of the target. The fact that the *behavior* of the target can considerably affect its overall risk is another aspect which is often taken for granted. The idea is that, regardless of its security posture, the risk of being compromised of a given entity increases simply because of the actions it performs. For instance, if a user spends a considerable amount of her time browsing dubious and less reputable web sites, it seems

reasonable that she would incur higher chances of being infected by malware than a user who only browses corporate and popular sites. Unfortunately, even if this may seem a logical conclusion, researchers have struggled to measure this simple relationship ⟨**R2**⟩. For instance, in 2013 Levesque et al. [139] found that the number of illegal and questionable websites visited by a user is less related to the risk of malware infection than the number of sport or computer sites. Similarly, Bossler et al. [68] found that the time spent performing illegitimate computer activities was not a good predictor of malware infections. Strangely, the authors found that even higher computer skills and the adoption of careful password management failed to reduce this risk.

Many independent studies [79, 139, 216] found instead evidence that the *volume* of performed actions (e.g., the number of software installed or the number of websites visited, independently from their category) was always correlated to a higher risk. If confirmed, this finding seems to suggest that there is a systematic risk of performing common actions – such as browsing the web or installing software – and the final risk would mainly depend on how many times these simple tasks are repeated by an individual or an organization. In other words, a possible direction is to try to model the risk of a compromise by using a *frequency-based* approach ⟨**R3**⟩, which is already a common solution to describe safety risks.

Measure the attack surface. In a given cyber environment, the attack surface is defined as the set of different points where an attacker can try to break into the system or exfiltrate information. As a direct consequence, reducing the attack surface by removing unnecessary services or limiting the access to parts of the infrastructure represents a way to increase the security by reducing the number of components that an attacker can target. The rationale behind this concept is that the likelihood of suffering from a security issue will raise according to the number and diversity of software, services, and systems used. While this is simple mathematics (and approached have been proposed to measure the attack surface of a system [155, 120]), the exact relationship that these variables have with cyber risk is still unknown and more experiments are needed to measure how risk actually reduces with the reduction of the attack surface ⟨**R4**⟩.

Influence of business sector, reputation, and assets of an organization. As we already mentioned above, non-technical characteristics of the target can influence the number, type, and sophistication of the adversaries it needs to face. Today it is widely accepted the hypothesis that, given enough time and resources, motivated attackers can always find a way to compromise a target. Large state-sponsored cyber attacks have shown this

to be the case also for the most secure government organizations [23, 13]. Therefore, the type of business, the sector, the reputation, and the assets owned by an organization may influence the risk of compromise more than other technical indicators, as they allow to capture the characteristics of the *attackers* (incentives, risks, and resources as proposed by [211]) instead of those of the defender. This assumption has already been shown to be valid to characterize both the number and the type of attacks, respectively by Sarabi et al. [208] and Thonnard et al. [235]. Moreover, this approach could also cover the risk of targeted attacks, whose ad-hoc natures does not allow them to be easily described by a frequency-based model [188].

Predict future events based on historical data. Historical data about claims and incidents are routinely used to estimate the risk in other insurance sectors. However, as already stated in section 2.1, the use of previously collected data to predict future cyber events faces several challenges. First of all, data on cyber incidents are scant and often biased towards those events whose disclosure is mandatory because regulated by law [20, 18]. A second challenge in this approach is to shed light on the so-called repeat players. Although previous studies found a systematic difference between costs incurred by companies that experience single or multiple incidents [198] (the so-called *repeat players*), it is still not clear whether having already been compromised is a good indicator of being again compromised in the future ⟨**R5**⟩. Finally, an additional complication is represented by the fact that attack techniques evolve very rapidly over time, making obsolete results obtained from the observation of old data. For instance, if a known vulnerability associated with a high-risk factor were to be patched, past records about events occurred because of its presence would probably not provide any contribution to capture the risk associated to new attacks.

Measure the risk that propagates through third-party relations. Outsourcing many critical business operations became a norm in the last decade. It is very typical to store and process data owned by companies on third-party cloud services and even common services such as DNS and emails are now outsourced to the cloud. This largely complicates the picture for cyber insurances as it is harder to draw a clear line of the boundaries of a company. As common sense suggests, a company that is in relation to other risky entities should have higher risk itself. While constructing sufficiently accurate service-dependency graphs of businesses is a challenging research topic by itself [95], measuring the amount of risk that propagates through this graph is an open research problem that needs attention from the community. We will come back to discuss this problem in more details in Section 3.3.3.

User's weaknesses and social engineering. One of the most common techniques used today to gain access to a network or system is social engineering: indeed, while one can think that the most successful breaches are the result of technical flaws or zero-day vulnerabilities exploitations, almost 97% of them is achieved by tricking users to reveal sensitive information using a social engineering scheme [30]. Unfortunately, while social engineering attacks can pose a tremendous threat to organizations, current approaches to IT security and risk management tend to underestimate or completely ignore the human factor in risk assessment models, tools and processes [192]. Extending existing schemes by modeling users and their behavior could largely increase their prediction accuracy⟨**R6**⟩.

Risk aggregation. All the factors we previously mentioned are likely to somehow affect (to a different and still unknown extent) the risk of cyber incidents. But even if researchers would be able to precisely identify a number of good and stable risk indicators, we would still have known very little about the aggregation procedure required to combine the different scores. This problem is exacerbated by the fact that, for practical reasons, each study looks at a single factor in isolation. But different factors are probably not independent and they can have very complex consequences and side-effects on other indicators. For instance, a good security posture may mitigate a larger attack surface, but it can be completely undermined by untrained users. Therefore, if distinct studies respectively find good predictors of risk, a constructive combination of them would still require a considerable amount of research ⟨**R7**⟩. A classic insurance solution could be to evaluate all risk indicators separately and then rely on actuarial data about past incidents to combine them in a single risk class, but as we already said this data may be very hard to put together and may become obsolete very fast. Finally, a major obstacle to risk aggregation is the different granularity of the risk computed by different approaches. Some can predict the risk of compromise of a given software artifact, other of a user, or of an individual machine. How to aggregate these values, for example, at a company level is still an open research problem ⟨**R8**⟩.

Horizontal Issues

So far, we discussed different open problems and research questions and their relevance for cyber insurance. However, we believe it is important to also highlight three important aspects about cyber risk itself that apply to all previously mentioned approaches:

1. **Cyber risk vs cyber-insurance risk** – As briefly shown in section 3.2.1, almost all the existing literature focuses on cyber risk assessment or prediction. Although these are important for the purpose of diverting security spendings towards most relevant threats, such evaluation could be misleading for cyber-insurance risk assessment. Indeed, a quantification of the first does not necessarily reflect the second, that after all is the actual value insurances are interested in: for instance, a class of events could have a high risk to harm one entity but lead to claim submissions with a very low probability. In other words, it is also important to study and measure how cyber risks translate to insurance claims in the real world ⟨**R9**⟩.

2. **Consumers vs corporations** – Since cyber-insurance products are recently made available also for the consumers market [12], it is possible that a different approach and/or set of features should be considered depending on the entity under investigation. Indeed, consumers are less active with respect to big corporations, operate in a different scenario, and may become an appealing target of cyber attacks for different reasons compared to large enterprises. However, no study exists to date to compare the risk and threats encountered by consumer vs enterprise users ⟨**R10**⟩.

3. **Risk variety** – risk assessment or prediction procedures need to be targeted towards specific categories of risk. Indeed in an insurance context, addressing cyber risk as a single-unit problem may be too generic and may not lead to meaningful results. For instance, the authors of [61, 259] predict machines and users at risk of malware infections, without providing any fine-grained categorization (after all, malware is a very generic term). In the same way, Liu et al. [150] attempt to forecast generic cyber incidents specifying no type or effect. However, as shown by Eling et al. by using actuarial data [102], different types of data breaches need to be modeled as distinct risk categories. A more fine-grained classification is needed ⟨**R11**⟩to also highlight particular categories of threats strongly coupled to the subject we are evaluating the risk for: for instance, malware targeted against banking systems are probably not very relevant for those enterprises in other business sectors.

3.3.2 Area 2: Automated Data Collection

"If you're writing policies for personal automobile or personal homeowners insurance you definitely have a lot of really good data. The worst data is probably in cyber insurance"

– Nick Economidis, Cyber liability underwriter at Beazley PLC [253]

The importance of data collection for cyber-insurance carriers does not only relate to the actuarial domain, whose issues have already been discussed in section 2.1. Data collection about prospective clients is indeed the first crucial task of policy underwriting, as it allows insurance firms to elicit a reasonable approximation of the overall security posture of the applicants, measure their level of risk, and subsequently compute premiums. The most common way to achieve this goal is to furnish organizations wishing to buy a cyber-insurance policy with security questionnaires. In a recent study, Romanosky et al. [199] analyze 44 of these questionnaires filed across the states of California, Pennsylvania, and New York, and point out commonalities that allow to group the questions into four macro categories.

The first set of questions aims at defining some general *organizational* details of the company, like its business sector and annual revenues, the kind of sensitive information stored and handled, how relationships with third-party service providers are managed, the nature and amount of IT security investments and, if any, its cyber-incident history. The second category focuses on *technical* aspects, often covering questions on security and access control measures adopted by the company and, less frequently, on its information technology and computing infrastructure. The existence of *policies and procedures* for data management is investigated in a third set of questions, in which insurance firms investigate whether data processing, retention and destruction practices are compliant with current regulation laws and procedures to maintain and strengthen information security. Finally on the *legal* side, questionnaires verify how well a variety of laws and regulations, enacted to protect consumers from the consequences of cyber incidents and data breaches, are implemented and adhered.

The information collected is then used for premium computation: while some carriers use flat-rate pricing for each first- and third-party coverage (with no differentiation by firm or industry), others incorporate more features (such as firm's sector and revenue) as factors to be multiplied in a base rate pricing. In more sophisticated policies, also the soundness and completeness of security controls and practices have a weight in the final result.

Although these questionnaires are widely adopted by cyber-insurance

firms, the measure of their accuracy as a standalone tool for defining the security posture —and as a step further the risk — of an organization is still questionable. A recent work examines 24 application forms to determine whether the collection of security checks referred by technical questions corresponds to the controls defined in two well-known standards of security best practices [254]. As result, existing forms are found to be predominantly focused on a small range of controls and the authors suggest how to extend them to be in alignment with the two information security frameworks. Nevertheless, the extent to which security standards compliance reflect the level of risk a company faces has not been yet understood ⟨**R12**⟩.

As suggested by modern approaches for data collection about cyber-insurance applicants [75], cyber questionnaires should be only one of many tools employed by insurance firms. For instance, instead of relying on self-assessment, the security posture of an organization can be automatically refined using two types of data sources: (i) *internal data*, provided by monitoring and telemetry tools installed inside the subject under investigation; and (ii) *external data*, collected from publicly available databases or by scanning Internet-facing services.

Although recent works show the feasibility of both approaches [150, 61], open questions still exist on both sides. Intuitively, internal data (if available) should provide a better accuracy to understand the cybersecurity risks of an organization. However, organizations do not exist in the void and the outcomes of internal telemetry analysis could be insufficient when assessing the security posture of an entity that maintains relationships or dependencies with external subjects – thus requiring a combination of the two approaches to cover unavailable information about these third parties.

On the other hand, in a cyber-insurance scenario, internal data could be unavailable to the insurer, who needs to base his evaluation on external data only. In this respect, the effectiveness of methods based on such sources only is not known, neither conditions and circumstances in which they can be used to achieve a good accuracy. As already depicted in Table 3.2 , studies that use external indicators to predict risk also validate their findings based on externally available ground truth. This is a big limitation, as cyber incidents are insufficiently reported and records, even if available, are often published too late and miss details and key elements. Moreover, the different precision and granularity of the ground truth make impossible to compare the results with those obtained with internal indicators. More research is therefore needed to compare the accuracy and relationship of externals indicators and internal telemetry information on the same dataset ⟨**R13**⟩. In particular, no previous work has provided insights on a combined use of

both sources, trying to answer the question whether internal data can serve as ground truth for refining the power of external indicators ⟨**R14**⟩.

3.3.3 Area 3: Catastrophe modeling

"One key challenge is accumulation. [...] We know we can write earthquake exposures in both Japan and California with the confidence that the same event will not impact all these exposures at once. We know to be wary of writing two industrial risks along the same river basin, and the role flood defenses play in mitigating loss. With cyber risks, the contours of systemic accumulation are not as clear"

– Hemant Shah, Risk Management Solutions [76].

For an insurance company, catastrophe modeling (or simply cat modeling) is a way to estimate the likelihood or frequency at which catastrophes can occur and to what extent they can impact the insurance. To decrease the likelihood of cyber catastrophes, a typical solution that is widely adopted is client diversification. The assumption here is that if the clients of the insurance company have diverse attack surfaces and diverse characteristics, a potential new zero-day vulnerability will not exist in all of them, leaving only a percentage of insureds affected by a possible cyber attack. While this may seem a reasonable conclusion, a recent unpublished work from Eling and Schnell [103] suggested that, when modeling losses with specific distributions, diversification may not be a good idea because of the heavy-tailed distribution nature of cyber risks. This would be an important and counter-intuitive finding, that needs to be confirmed by further measurements ⟨**R15**⟩.

At its core, cat modeling boils down to capturing and modeling dependencies among different entities. This, in turn, translates into the identification of the dependencies that come from the software, hardware, and services used by a company. However, obtaining such detailed and comprehensive information about a large enterprise is a very challenging task. Moreover, because of the cyber-insurance context and the transitive nature of these dependencies, this task would need to be performed by using publicly available datasets. This makes the problem even more complex and hence, we believe, it opens new directions for researchers to explore and contribute.

In an ideal scenario where all companies reveal the software, hardware, and services they use and share with the community, building the service dependency graph, identifying the nodes in this graph that might cause catastrophic events, and calculating the indirect risk that comes from these

dependencies would be a simple task. However, even in such a perfect world, the dynamicity of the graph would require to continuously report and recalculate the risk and likelihood of the existing catastrophes and the identification of new catastrophe scenarios. In other domains, if two risks are not connected (such as a fire hazard on two areas tens of thousands of miles apart) this fact is not likely to change in the near future. But in the cyber-insurance domain, the relationships among two different companies are often very ephemeral – as services providers and software libraries may change very often. But as of now, there is no existing work that studied how the dynamicity of the ecosystem could influence the whole cat modeling topic and whether (and how often) the portfolios defined by the insurance companies should also be updated ⟨**R16**⟩.

Moreover, the reality is far from this ideal scenario and even the topic of building adequately accurate service dependency graphs and modeling the catastrophes with sparse and incomplete data are research topics that need more attention from the community ⟨**R17**⟩. Altogether, this can lead to a *supply chain risk analysis* that would provide a principled foundation for catastrophe modeling.

However, the identification of all services used by a company, especially without its cooperation, is often infeasible. For instance, the presence of backup or redundancy services can remain undetected, as those only come into play when the primary provider fails. As a full and precise view of all the dependencies of a company may be impossible to obtain, then a modeling algorithm should be able to work with incomplete information, potentially inferring the missing connections from settings and relations observed elsewhere ⟨**R18**⟩. Although not done particularly for the cyber-insurance domain, there exist two works [95, 223] that aimed at building dependency graphs of popular companies by using public datasets such RIPE atlas, passive DNS records, and web crawling data. In 2017, Dell'amico et al [95] performed a large-scale study to identify the dependencies between websites and Internet services. The findings of the study confirm the monopoly problem in the current service ecosystem. To make matters worst, over time the Internet appears to be loosing its decentralized nature and the popularity of the few dominant providers is steadily increasing. In the same year, Simeonovski et. al [223] built a service dependency graph to explore what percentage of the Internet would be effected when a popular provider is attacked. The study found that by only targeting a handful of service providers it would be possible to take down 23% of the websites.

Another challenge that affects cat modeling is the lack of a mapping procedure to reliably associate measurements and public data to organi-

zations. Network scans, web crawlers, service monitoring systems, public blacklist, and other techniques that can be used to identify the software and technologies adopted by a company typically work at the level of domain names or IP addresses. On the contrary, incident reports and risk prediction operate at a company granularity. Sadly, the connection between the two is not always straightforward and new techniques are needed to link the two information ⟨**R19**⟩. For instance, Liu et al. [150] explain their attempt to perform a manual mapping and all the difficulties and caveats encountered in the process, making it evident the necessity of a clearer and automated procedure.

3.3.4 Area 4: Forensic Analysis

"I often think of the 1990s as the decade of prevention, the 2000 as the decade of detection, and this is the decade of incident response."

– Bruce Schneier, Security Specialist

After the detection of a cyber incident, the response phase requires the intervention of computer security experts to analyze and understand the detail of the event. However, computer security skills are not only required for helping the company to recover from the incident but also, from an insurer's perspective, to verify the claim, assess the damage, and confirm whether it is covered by the subscriber's policy. Indeed, forensic investigations are the norm to assess if, and to which extent, the insurance is liable for the event.

Computer forensics is a broad research field that covers the collection, analysis, and preservation of digital evidence. It is a highly developed science with its own language, modus operandi, and standardized procedures [81]. However, while the other research topics discussed in this paper have all been recently contextualized (in terms of specific problems and new challenges) to the cyber-insurance domain, no study has looked at the problem of computer forensics from a cyber-insurance perspective.

For instance, one aspect that may require special attention is information forgery. In traditional insurance sectors, fake accidents cost over 30 billion dollars per year, with several insurers reporting these frauds to account for up to 20% of claims costs [55]. However, while set-up wrecks and burning houses are sadly common practice for fraudsters to cash the insurance coverages, there is almost no mention to date about similar frauds in the cyber domain.

Current forensic approaches are mainly concerned with the possibility that an attacker can hide undetected or that important evidence and artifacts can be deleted or manipulated. In other words, the focus on evasion

and not on forgery. The lack of motivation can explain why planting fake evidence in a computer system is not yet very common, but forged incidents are extremely easy to set up for anyone with average programming skills [111]. The vast majority of the indicator of compromise used today rely on the simple existence of filesystem and registry artifacts - without any knowledge of how (and by whom) the data was created in the first place. In this setting, it is not hard to mimic a malware infection or even a targeted attack against an organization. However, with cyber insurances becoming more and more common, forged digital evidence may become a major problem in the future.

In particular, digital evidence forgery could help businesses to overcome one of the cyber-insurance most-common pitfalls: the fact that technicalities can invalidate coverage allowing insurance carriers to deny indemnity payments [8]. For instance, cyber insurance does not normally cover when employee errors (e.g., falling for phishing attacks) are the cause of a malware infection (e.g., ransomware) [10]. Since these events are instead covered under other clauses (e.g., malware installed by an external attacker), forging digital evidence would allow to "fake" a botnet infection to fall within the scenario covered by the insurance policy, thus allowing the victim to cash the indemnity.

Since today staging fake security incidents requires very little effort, researchers should not only study how to collect hidden signs of compromise, but also how to double-check and validate their authenticity ⟨**R20**⟩.

3.4 Development of cyber-insurance research

Cyber insurance is still a hot research topic among practitioners and academics. In this section, we briefly discuss follow-up works that have been published after our study and that contribute to one of the four research areas that we consider in our systematization.

As we widely discuss in section 3.2, cyber-insurance carriers make large use of questionnaires and self-assessment surveys to collect useful information from prospective clients in order to establish premiums. Nurse et al. [177] interviewed a group of cyber-insurance professionals and reported the challenges in both gathering and using collected data. Authors found that insurers are likely to investigate security-related information —such as whether the enterprise has a Chief Information Security Officer (CISO) and the extent to which employees received cybersecurity training. During the underwriting phase, practitioners also note that if too much of this data is requested to potential clients, those may choose a competitor that requests

less information, which usually translates into accepting more risk. At the same time, a crucial issue mentioned by the authors is the storage of the huge amount of data collected after an incident and the lack of systems for analyzing it in the future.

Uuganbayar et al. [240] instead proposed a game-theory-based approach to help organizations in distributing investments between self-protection measures and cyber-insurance policies. Their unique approach not only defines the amount of investments but also how those should be spent by selecting the most cost-efficient security controls.

Given the rise of interconnections among IoT-driven service organizations, Pal et al. [183] investigated the feasibility of a cyber re-insurance market able to cover the catastrophic service disruptions that propagate through these networks and that are worth billions of dollars. Authors proved through a game-theoretical analysis that it may not be economically incentive compatible to cover the aggregate cyber-losses arising due to this kind of cyber-attacks.

Finally, Woods et al. [255] published a systematization study that organizes prior work on cyber-risk quantification by using a causal model inspired by structural equation modeling. The authors analyzed past literature that attempts to quantify how much harm results from cyber incidents and whether this has changed over time. They found little evidence that either the typical size or variance of cyber harm is particularly exceptional. On the other hand, while analyzing studies that focus on the impact that different security interventions have to reduce harm, they discover that those studies do not consider important factors as the threat level and often report spurious results like increased security budgets leading to greater frequency of breach or that applying software updates increases the likelihood of web-server compromise.

3.5 Conclusions

Even with profuse investments, attacks targeting cyber infrastructures are not preventable in their entirety. As this awareness of cyber risks increases among companies and organizations, cyber insurance gains prominence as a tool to cope with residual risks and face hazards and losses resulting from the use of computer systems and services. This new insurance sector presents unique challenges and unsolved problems: quantify cyber-risk exposures and consequently price them remain major issues due to the lack of historical data and effective modeling techniques; on top of those, risk accumulation due to the interconnected nature of some attacks and catastrophe occurrence

are other two wide open research areas.

In this chapter we discussed the unique challenges that affect the cyber-insurance sector. We focus on a pure technical perspective, highlighting the limitations of current approaches, evaluating the feasibility of new solutions, and proposing research areas in which system and network security experts can play a fundamental role for the development of cyber insurance. Differently from legacy frameworks based on qualitative approaches for risk assessment and data collection, we endorse the relevance of prediction techniques based on objective measures and automatic feature gathering.

In the next chapter, we start exploring some of the open challenges proposed in this chapter. In particular, we look at risk indicators of malware encounters by analyzing real-world telemetry and by differentiating between home machines and those used in enterprise environments.

Chapter 4

A Comparison of Systemic and Systematic Risks of Malware Encounters in Consumer and Enterprise Environments

4.1 Introduction

In this chapter, we leverage real-world telemetry to conduct an exploratory study on risk indicators of malware encounters, differentiating between home machines and those used in enterprise environments.

In the first part of the chapter we carry an extensive analysis of common aspects and differences in malware encounters between the two segments. To assess the implications of different choices in security investments and policies, we quantify the malware encounter rate in consumers and enterprises and provide evidence of the most common classes and signatures observed by the two parties. We also look at the reporting frequency and different labels of popular malware families, the different incidence that PUA and Adware, and the impact that behavioral signatures have on corporate and consumer hosts.

In the second part of the chapter, we conduct an exploratory study of undiversifiable risk indicators that we were able to extract from our real-world telemetry. For instance, we assess whether the days and hours of activity together with the volume of host-generated files can serve this purpose. We also look at the effect that the number of installed software vendors has on the malware encounter rate. We assess whether being in a recidivist infected state can be a good risk predictor, and finally, we verify whether the size of an enterprise or its industrial sector can provide useful insights on the systematic risk the company encounters.

4.1.1 Systemic vs Systematic

The way in which consumers and enterprises approach security is very different: while the former follow a reactive approach, installing defenses (typically in the form of AV software) to detect and remove possible malware infections, companies are expected to work more proactively, by relying on articulated risk assessment, mitigation, and risk transfer methodologies [94]. It is also well-known that consumers invest less in security, often preferring off-the-shelf, easy-to-use solutions that offer few customization options. On the contrary, organizations tend to protect their assets and data by deploying complex and multi-faced solutions that rely on several layers of defenses, such as firewalls and security proxies, intrusion detection and prevention systems, email protection and anti-exfiltration software, together with measures to prevent insider attacks and to limit the spread of infections. Consumers and enterprises also differ from a user point of view. In fact, in addition to educating employees on the best security practices, enterprises may adopt stricter security policies about what software can be

installed— thus preventing users from running software of dubious origin that is often a vehicle for malware. For end-users, this choice is left to the sole user's security awareness and knowledge. Each of these factors may affect the risk of experiencing cyber incidents and malware infections. This risk, known as *systemic*, is strictly related to the individual security posture and to the adopted counter-measures.

However, there are also other factors to be accounted for, as both consumer and enterprise machines are not isolated entities. The interconnected nature of our society brings companies to rely on external services and to outsource computational tasks to third-party subjects, thus exposing the hosts of both parts to a potential *systematic* risk — which is a form of undiversifiable risk that is independent of how much a subject spends in security products and from its cyber hygiene.

The two terms, systematic and systemic, are commonly used in the financial sector. In particular, the systematic risks (also known as undiversifiable, volatility, or market risk) refer to the risk inherent to the entire market, that is not specific to a particular stock or industry and that therefore is impossible to completely avoid and cannot be mitigated through investment diversification. On the contrary, systemic risk (also known as nonsystematic, specific, or residual risk), is unique to a specific company, industry, or market segment. This second type of risk can be reduced by simply redirecting the investment towards multiple companies,stocks, and markets related to different sectors, thus reducing the likelihood that a failure in one of them could influence the others.

4.1.2 Why this study matters

Cyber risk estimation is a very complex and challenging problem, that to date was mainly approached from a *qualitative* perspective [94]. In this work, we aim instead at exploring *quantitative* metrics, obtained by leveraging empirical data. In fact, while it might sound obvious to the reader that some factors are correlated to higher security risks, (e.g., the fact that machines with higher activity are more likely to encounter malware, or that a broader and more diverse set of software results in higher attack surface), the exact relationships that these variables have with the risk of encountering malicious software has never been measured before. In addition, our study provides numerous insights on the different impact these factors have on consumers and enterprises environments.

It is also important to stress that the cyber security risks of consumers and enterprises have never been compared before. Although it is possible to infer some differences and similarities by looking at studies that analyzed

either the first or the second segment in isolation, those studies relied on different data sources, focused on different aspects, and were performed over disjoint timeframes, thus making it difficult to compare results. On the contrary, the internal telemetry information we use for our experiments comes from a single AV sensor, it has been collected in the same time period, and it allows us to clearly distinguish between corporate and consumer machines. Moreover, while risk assessment is one of the cornerstones of computer security, the difference among consumer vs enterprise security has never been experimentally measured before: do enterprise machines encounter less malware because they are protected by more and more diverse cyber defenses? Are enterprise users more security conscious and, therefore, less likely to visit risky websites at work? Is there some relevant difference among the malicious files encountered by end-users and large companies employees?

4.2 Datasets

This section provides a detailed description of the different data sources we used in this study, as summarized in Table 4.1. Our main source of information is the telemetry data obtained from NortonLifeLock [46] collected on Windows machines and made of different feeds. *Activity* reports provided a starting point to list all machines that had the antivirus sensor installed and opted in to share their data, allowing us to compute the number of hours each machine was active every day. *File appearance logs* helped us to identify vendors of installed programs. Using *malware encounters logs* we identify where, how many times, and which signatures were triggered for each malware encounter. Finally, we scraped the company website to retrieve a list of all existing *signatures* along with their class and description.

4.2.1 Consumers vs Enterprises

Our data contains 640 K unique enterprise identifiers. However, since big corporations can span multiple countries and comprise several subsidiaries— each of which may possess a different identifier— we use a second mapping to further group those cases to a single organization. In total, we were able to identify 45.6 K (2nd record in Table 4.1) unique organizations. We distinguish 6.5 K micro ($\leq$ 10 hosts), 12.3 K small ($\leq$ 50 hosts), 11.9 K medium ($\leq$ 250 hosts) and 14.8 K large enterprises ($>$ 250 hosts), with the biggest of them having 3.4 M machines.

In the period of our experiments, we observed a total of 144.9 M dis-

Dataset	Info About	Unique instances	
		Consumers	Enterprises
Activity	Hosts	144.9 M	226.4 M
	Enterprises		45.6 - 640 K
	Countries	239	235
File appearance	Vendors	59.9 K	40.9 K
File appearance	Vendors	59.9 K	40.9 K
Encounters	Hosts	14.2 M	27.1 M
	Enterprises		26.6 - 244.2 K
	Records	62.4 M	76.5 M
	Signatures	24.0 K	23.3 K
	Countries	239	235
Signatures	Labels	32.0 K	
	Subclasses	41	
Industrial sectors	Sectors	10 - 1215	

Table 4.1: Overview of datasets used

tinct consumer machines and 226.4 M enterprise machines. Our dataset covers 239 (for consumers) and 235 (for enterprises) of the 249 countries, territories or areas of geographical interest with an assigned ISO 3166-1 code [24]. The two tables below (Table 4.2 and Table 4.3) report the geographical breakdown of the machines in our dataset: North America is the most represented region (38% of the machines), followed by Europe (27%) and Asia (22%). In South America, Africa and Oceania we measure the lowest concentrations (< 10 % overall).

4.2.2 Host activity and file appearance

All the 371 M machines in our dataset have an anonymized identifier linked to the AV software licence and thus stable throughout the period under analysis. Each of them routinely queries a centralized system to assess the reputation of files that appear on the host. These requests are made possible thanks to the explicit consent of both consumer and enterprise users, who opted-in to share their data in an anonymized and privacy-preserving form. We leverage this process for two different purposes. First, for each machine and for each day in the time frame of this study, we computed the number of active hours. We then computed the number of active days per month by counting the days in which the machine submitted at least one request. On average, consumer and enterprise hosts are active 6.4 and 7.6 days

Consumers		Enterprises	
Country	% hosts	Country	% hosts
United States	33.87	United States	35.52
Japan	7.46	India	6.60
Germany	5.41	China	4.51
United Kingdom	4.60	Brazil	3.39
China	3.74	Japan	3.12
Brazil	3.52	United Kingdom	3.02
Canada	3.45	Germany	2.22
France	3.25	France	2.10
Australia	3.07	Canada	1.90
India	2.61	Australia	1.55
Italy	1.98	Mexico	1.52
Others	27.04	Others	34.55

Table 4.2: Host distribution per countries

Consumers		Enterprises	
Continent	% hosts	Continent	% hosts
North America	38.89	North America	42.55
Europe	27.57	Asia	27.18
Asia	22.32	Europe	19.76
South America	5.49	South America	5.80
Oceania	3.49	Africa	2.57
Africa	2.24	Oceania	2.13

Table 4.3: Host distribution per continents

per month, respectively for 2.9 and 3.7 hours per day. Second, for all executed applications we extract the vendor name (if the file is signed), thus identifying more than 40 K distinct vendor names for enterprises and around 60 K for consumers.

4.2.3 Malware Encounters

When a file is flagged as malicious by the host AV sensor, the event (including the hash and the signature identifier) is reported to the central server. We use these logs to create a register that, for each machine, records the day, the number of encounters (as the same object can be reported multiple times), and the matching signature name. Our data do not allow us to

perform a retroactive analysis of files to catch newly identified threats, but only consider those reported by existing signatures at the time of detection. Over the 140 M collected events, we identified 14.2 M distinct consumers and 27.1 M distinct enterprise hosts that encounter at least one malicious file within the year. Overall, malware was encountered by 58.3% of the enterprise.

We scraped the website of the AV vendor to obtain the list of available signatures—together with their descriptions, years of creation, and subclasses. In this way, we were able to gather information about 18143 labels classified in 41 subclasses (out of roughly 24 K signatures observed in the dataset). For a more concise classification, we decided to merge similar and smaller subclasses into seven broader groups: *Adware*, *PUA*, *Trojan*, *Ransomware*, *Worms*, *Viruses*, and *Others*. The full mapping among the different classes is reported in Table 4.4.

Class	Subclass
Adware	Adware, Adware-trojan
PUA	Misleadingapplication, Misleadingapplication-trojan, Potentiallyunwantedapp
Ransom	Ransom
Trojan	Trojanhorse, Trojanhorse-macro, Trojanhorse-virus, Trojanhorse-worm, Trojanhorse-worm-macro, Trojanhorse-worm-virus, Trojan-virus, Trojan-worm
Virus	Virus, Virus-macro
Worm	Worm, Worm-macro, Worm-virus
Others	Dialer, Dialer-adware, Dialer-hacktool, Dialer-trojan, Hacktool, Hoax, Joke, Joke-trojan, Macro, Other, Other-trojan, Other-worm, Parentalcontrol, Remoteaccess, Removalinformation, Securityassessmenttool, Securityassessmenttool-trojan, Spyware, Spyware-trojan, Trackware, Trackware-trojan

Table 4.4: Malware classes grouping

4.2.4 Enterprise industry sectors

For a subset of the anonymized enterprise identifiers, we were provided with a number of additional information; including their industry sectors and the countries in which their registered offices are based. This industry classi-

fication is available in different granularities, ranging from a fine-grained classification of up to 1215 distinct sectors to a coarse version of only 10 macro-sectors.

Table 4.5 shows the number of machines and enterprises per sector, according to the most concise classification: information technology is the prevalent industry with more than 3 M hosts and 4732 enterprises. Globally, our dataset shows good industry coverage, with all sectors having at least 200 K active machines, and half of the sectors having more than 1 M hosts.

Sector	Enterprises	Hosts
Consumer Discretionary	5030	1.99 M
Consumer Staples	1495	912.22 K
Energy	654	210.71 K
Financials	5052	2.96 M
Healthcare	2349	1.96 M
Industrials	7715	2.79 M
Information Technology	4732	3.63 M
Materials	2159	427.00 K
Telecommunication Services	314	307.08 K
Utilities	496	245.59 K

Table 4.5: General sector statistics

4.2.5 Selection Bias

The dataset we used for our study is the largest ever adopted for risk-based experiments: while the telemetry of previous works included at most 20 K consumer devices [216], and 82M machines of 28k enterprises [137], the one used in this work has been collected on more than 226M organization hosts and 144M home-user computers located in almost 250 countries. However, it is not completely unbiased. For instance, we only analyze consumers and enterprises that invest in security products: it is reasonable to believe that those without any protection should have a worse security posture, thus making our results conservative. Moreover, our datasets are obtained from a single vendor and only from those users who opted-in to share data: although this allows us to better compare the two classes of machines, software from other vendors may provide different security, and users who opted-out due to privacy concerns could be more security conscious. Finally, our telemetry is only collected on Windows hosts and so it is possible that users

running other OSes (e.g., macOS) may have a different security posture.

Consumers			
Class/Label	Hosts	Reported Events	Labels
Trojan	11.3M (79.5%)	186.7 M	3.4 K
W97M.Downloader	627.3 K	3.1 M	
Mdropper	305.8 K	1.1 M	
Dromedan	303.6 K	916.8 K	
PUA	6.3M (44.4%)	32.6 M	747
InstallCore	699.2 K	1.0 M	
DownloadSponsor	509.3 K	1.6 M	
OpenCandy	335.4 K	438.6 K	
Others	4.6M (32.4%)	12.6 M	820
Jswebcoin	148.5 K	669.8 K	
Remacc.Ammyy	101.2 K	155.5 K	
Remacc.Radmin	10.9 K	18.9 K	
Adware	770.9K (5.4%)	2.4 M	491
Browext	154.0 K	623.3 K	
DealPly	54.3 K	87.6 K	
DriverUpdater	48.2 K	56.0 K	
Worm	559.1K (3.9%)	4.1 M	1.1 K
Silly	125.3 K	353.8 K	
Ippedo	64.7 K	206.8 K	
Dunihi	53.4 K	1.5 M	
Virus	279.5K (2.0%)	15.1 M	589
Sality	56.1 K	2.1 M	
Virut	47.8 K	493.1 K	
Bursted	34.5 K	154.8 K	
Ransom	112.0K (0.8%)	416.1 K	326
Wannacry	51.4 K	299.8 K	
Crysis	15.1 K	26.7 K	
Cerber	7.5 K	10.3 K	

Enterprises				
Class/Label	Enterprises	Hosts	Reported Events	Labels
Trojan	16.1K (60.5%)	22.7M (83.8%)	217.1M	3.2 K
Dromedan	2.6 K	481.7 K	1.4 M	
W97M.Downloader	4.0 K	179.3 K	603.5 K	
JS.Downloader	1.5 K	98.9 K	187.7 K	
Others	10.9K (41.0%)	2.3M (8.5%)	7.1 M	616
Remacc.Ammyy	985	74.6 K	115.2 K	
Jswebcoin	1.6 K	70.2 K	286.0 K	
Remacc.Radmin	172	26.5 K	42.2 K	
PUA	10.5K (39.5%)	1.9M (7.0%)	3.8 M	548
InstallCore	3.6 K	245.6 K	307.9 K	
OpenCandy	3.0 K	186.7 K	231.2 K	
DriverPack	1.1 K	105.1 K	149.5 K	
Worm	4.1K (15.4%)	692.1K (2.6%)	5.1 M	884
Silly	1.8 K	164.5 K	438.5 K	
Ippedo	1.0 K	83.5 K	325.2 K	
Dunihi	1.1 K	68.5 K	2.0 M	
Virus	2.6K (9.8%)	320.9K (1.2%)	17.6 M	396
Sality	1.1 K	74.8 K	3.2 M	
Virut	933	59.2 K	739.5 K	
Bursted	639	52.8 K	232.4 K	
Ransom	1.2K (4.5%)	160.6K (0.6%)	665.8 K	307
Wannacry	550	109.2 K	546.3 K	
Crysis	210	21.5 K	37.2 K	
Locky	31	4.0 K	7.8 K	
Adware	2.8K (10.5%)	149.5K (0.5%)	444.6 K	429
Browext	1.0 K	30.5 K	121.0 K	
Lop	339	16.4 K	20.5 K	
Funshion	153	6.9 K	15.5 K	

Table 4.6: Most common malware signatures and classes for consumers and enterprises. For each malware class, percentages represent a normalization to the total number of hosts and organizations that encounter malware.

4.3 Malware Specificity

In this section, we describe similarities and differences in malware encounters among consumer and enterprise hosts. We start by analyzing the overall picture of encountered malware signatures and classes in section 4.3.1. Considerations about the number of malware classes on each host and the average age of signatures follow in sections 4.3.2 and 4.3.3. In section 4.3.4, we finally discuss how behavioral signatures, PUA, and Adware impact consumers and enterprises in a different way.

4.3.1 Overall picture

We start our analysis by measuring malware encounter prevalence in consumers and enterprises. Over the twelve months observation period at our disposal, we found that the percentage of hosts that encounter malware slightly differs between the two groups: for consumers, 14.2 M of the 144.9 M active hosts have suffered at least one encounter (9.80%), while in enterprises 27.1 M out of 226.0 M machines (12.0%) detected malicious software. We verified that this difference is statistically significant ($p<.001$) by running a Chi-squared test on a 2-by-2 contingency table obtained by considering infected and clean devices in consumers and enterprises.

It is worth noting that the malware encounter rate we measured in enterprise environments is consistent with prior works. In fact, in their conservative estimation along three years (from 2015 to 2017), Kotzias et al. [137] report a prevalence rate of 13%; the same ratio increases to 15% in the study of Yen et al. [259], who consider hosts of a large enterprise in a four-month time frame in 2013. This shows that, once averaged over a sufficient number of computers, the malware encounter rate in enterprises remained relatively constant across different studies, AV vendors, and even across multiple years.

No prior study exists instead that specifically focuses on consumer hosts encompassing every class of malware. Some measured a combined encounter rate —therefore also including enterprise machines— on a global scale [41], others restricted their analysis to only few malware classes to investigate their distribution vectors [136, 167, 74]. Although in the report published by Microsoft [41] there is no clear distinction between consumer and enterprise machines, our study reveals an encounter rate that is higher than the 6% assessed by their researchers in the security bulletin over the same period.

While the overall encounter rates are similar, a closer look at the malware families shows that there are some relevant differences between consumer and enterprise encounters. Table 4.6 summarizes the most common malware

signatures and their corresponding classes in our telemetry data, together with percentages that represent a normalization to the fraction of devices and organizations that encounter malware. Labels are sorted by the number of distinct hosts in which they appeared, after removing generic records and those for which we could not assign a class (as explained in Section 4.2.3). As a single signature could be triggered multiple times in the same machine, we also measure and report these occurrences. We complete the picture by counting the total number of distinct labels for each class and the number of enterprises in which each signature has been observed.

Results show that Trojan is by far the most popular class: these signatures alone represent 47% of total number of signatures matched for consumers and nearly 80% for enterprises. Although this malware class is also prevalent in organization environments, home users show higher infection frequency and a more diverse set of labels: on average, consumer hosts report Trojan detection events 16.46 times during the year and encounter 2.02 distinct families in the same period. Enterprise frequency and distinct labels are lower (respectively 9.56 and 1.33). Again, the differences between the two means are statistically significant (Reporting frequency: Welch's Analysis of Variance (ANOVA) F-test = 5104, p<.001; Families: Welch's ANOVA F-test = 1709257, p<.001). The most common families are respectively *W97M.Downloader*, a well-known set of malicious macros embedded in Microsoft Word document files, and *Dromedan*, a label associated with a Trojan family spread via email attachments.

Table 4.6 also highlights the completely different incidence of PUA and Adware between the two groups. Although *InstallCore* —a large family of bundlers that install Adware and PUP— and *Browext* — malicious software that shows advertisement and slows down the system to frustrate the user— are the most observed labels on both sides, PUA and Adware account upwards 29% for consumers, but not more than 7.1% for enterprises. In addition, home users report Adware and PUA detections on average 5.18 times per year, while enterprise machines only 2.05 times (Welch's ANOVA F-test = 649, p<.001). Since this is an important difference between the two groups we decided to dedicate Section 4.3.4 to investigate it in more detail.

On the contrary, Viruses and Worms (respectively 1.1% and 1.2% of all the signatures matched) appear with similar frequency in both groups. Although we register a statistically significant difference in the mere detection rate between the two segments of machines (Virus: X-squared = 12447, p<.001; Worm: X-squared = 14164, p<.001), we find no such difference when considering the reporting-event frequency and distinct-label encoun-

ters: Viruses are respectively detected on average 53.92 and 54.99 times during the year on home-user and organization machines (Welch's ANOVA F-test = 0.12, p=.12), showing the same average presence of 1.21 different signatures per host (Welch's ANOVA F-test = 2.88, p=.09). Similarly, Worms are reported 7.40 (consumers) and 7.38 (enterprises) times on average (Welch's ANOVA F-test = 0.02, p=.90), in the form of 1.24 and 1.21 distinct labels (Welch's ANOVA F-test = 1.06, p=.08). Our data reveals that the family of *Silly* Worms, that replicates through email attachments and local copies to steal sensitive information and disable other software, is the most common in its corresponding class. *Sality*, a popular malware that infects executable files acting as backdoor or botnet, dominates instead the scene when it comes to Viruses.

4.3.2 Distribution of malware subclasses

Figure 4.1 shows the cumulative distribution of the number of distinct malware subclasses observed in enterprise and consumer hosts. For each machine, a subclass is counted if at least one of its signatures is matched by the AV product. The maximum number of distinct classes (22 for consumers and 21 for enterprises) has been reported by two machines per group. While at a first sight the graph might suggest similar behaviors in the two categories, the Chi-squared tests separately considering up to 20 distinct encountered categories reported significant differences with $p < .001$. In particular, substantial differences are present in the leftmost part of the plot: while nearly 82% of enterprise hosts have encountered only a single subclass of malware, this percentage drops below 57% for consumers. This, in turn, reveals that on average consumer machines are more likely to encounter a more diverse set of malicious files than enterprise computers. As already discussed in the introduction, a possible explanation for these differences can be the adoption of stricter security policies and multiple layers of defenses present in enterprises but not in consumer environments.

Our measurements show that for most of the malware categories there was no relevant change over the year in terms of the fraction of hosts that detect them. This supports the hypothesis that different malware classes reach a plateau that they maintain over time despite the effort of security companies to mitigate them. There were only two exceptions to this rule, which we present in Figure 4.2. The first was a slight but steady decrease of *Ransomware* families, both in consumer and enterprise data. The second was a rapid increase of *Cryptominer* families, followed by a general downward trend. Ransomware and Cryptominers are the last two malware classes that emerged over the last few years and their curves show that in

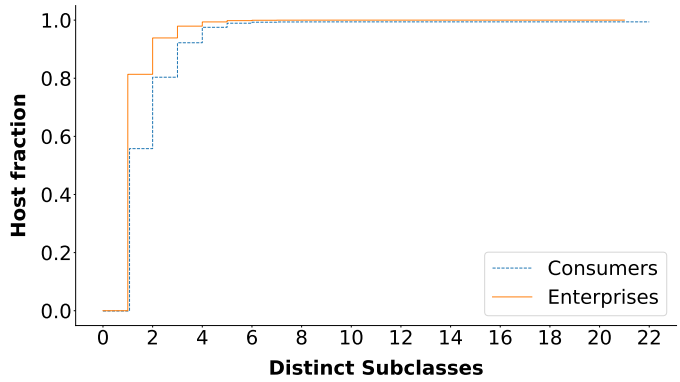


Figure 4.1: Cumulative distribution of the number of distinct subclasses per host

fact they did not yet reach a stable trajectory.

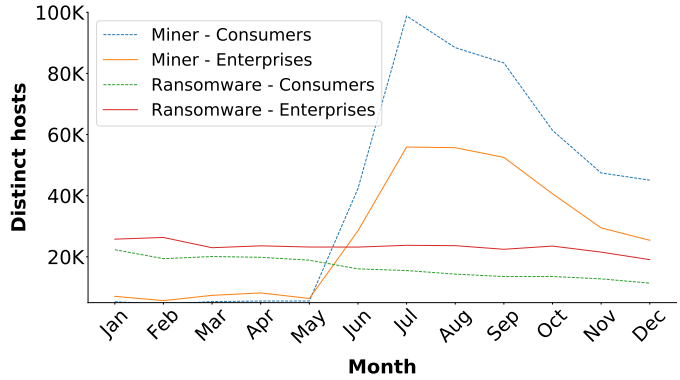


Figure 4.2: Ransomware and Miner trends for consumers and enterprises

4.3.3 Age of encountered malware

We continue our analysis by estimating how *old* the malware encountered by the hosts in our dataset is, by looking at the date in which each signature was first introduced by the vendor. Figure 4.3 depicts the average age of matched signatures in our one-year observation period. For each of the 12 months, we group all the labels based on the year in which they were created. Then, for each of the 29 years (from 1990 to 2018) we average the number of distinct records over the months and compute the 95% confidence interval. Despite a common peak of over 300 signatures written in 2014 and a drop for those developed in 2018, the number of matching

signatures present in our dataset is almost constant since 2003. This corroborates what has already been observed in other studies about the fact that it is still common to encounter today samples belonging to very old malware families [146]. In fact, about 174 K consumer hosts and 151 K corporate machines (respectively 1.0% and 0.5% of those that suffered at least a malware encounter) report encounters for signatures whose creation even predates the year 2000. Among those, the most common for consumers (4858 hosts) and enterprises (1990 hosts) is *CIH*, a 22-year-old signature to identify a computer Virus that targets Microsoft Windows 9x systems.

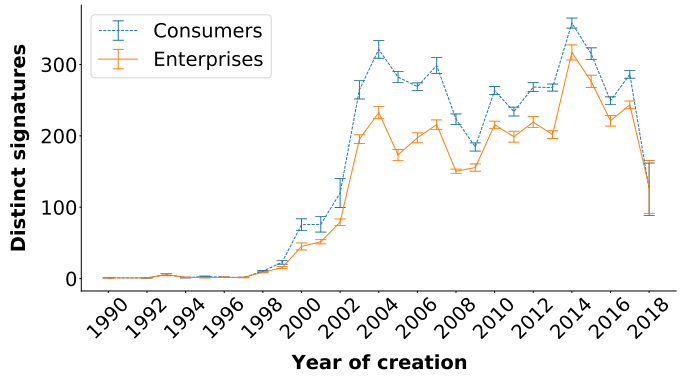


Figure 4.3: Average number of different signatures per year of creation. Error bars provide a 95% confidence interval

4.3.4 Behavioral signatures, Adware, and PUA prevalence

So far, in this manuscript we have used the word *signature* to indicate without distinction the set of unique data that allows an AV software to detect, quarantine, and remove specific malware. However, two main approaches exist to create a signature: the older pattern-based methodology in which a model was built to match a particular family of malware, and the more recent behavioral-based approach in which generic heuristics are used to capture different aspects of malicious behavior. While the first leverages object attributes to create a unique fingerprint, the latter typically evaluates an object based on its runtime actions [84].

In our dataset, we identified 6.7 K behavioral signatures by using their label and report their prevalence for consumers and enterprises in Figure 4.4. The reported percentages are obtained by dividing the number of distinct hosts with at least one behavioral-based detection and the number of distinct hosts that have suffered one or more encounters of any kind. We verify that

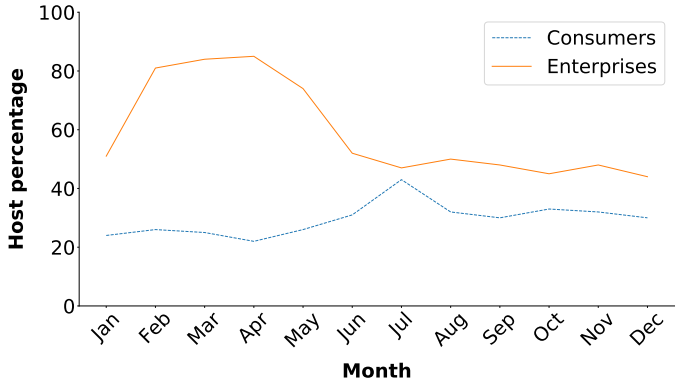


Figure 4.4: Prevalence of behavioral signatures in consumer and enterprise machines

all the monthly differences are statistically significant ($p < .001$) by running a Chi-squared test on the contingency table obtained by considering devices that trigger behavioral signatures and those that do not, in consumers and enterprises.

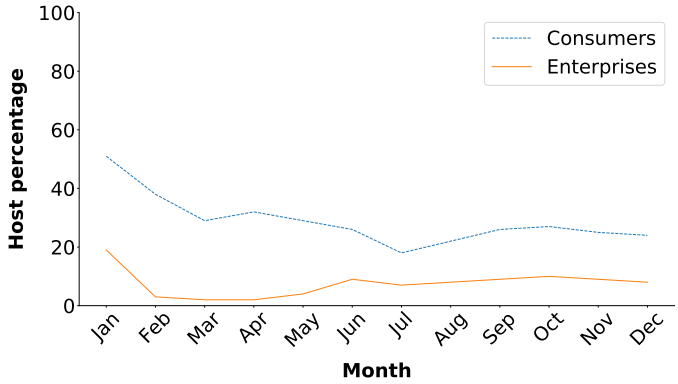


Figure 4.5: Prevalence of PUA and Adware signatures in consumer and enterprise machines

The curve for enterprise hosts lies considerably above the one of consumers, a sign that behavioral signatures match much more in the former environment (an average of 59% of hosts in enterprise vs 30% in consumer hosts). This could be due to the presence of less popular software and of custom applications built and compiled on corporate machines, for which the AV has not been tested against to whitelist or tune its behavioral signatures. On the contrary, consumer machines mostly run well-known appli-

cations that are therefore accounted for by AV vendors. However, since the totality of behavioral signatures is categorized as Trojan by the AV vendor, we speculate that this difference could also be due to sophisticated malware that targets specifically certain enterprises, which could not be easily detected with a traditional pattern-based signature.

In Figure 4.5, the trends are inverted when considering Adware and PUA. In fact, their prevalence in consumer hosts is constantly higher (6.06 times on average with a statistically significant difference for each month — $p < .001$) than in enterprises. As already discussed in section 4.3.2, a very likely explanation can be found in the freedom that consumer users have to install any kind of software, whereas more rigorous rules are enforced in enterprises.

4.4 Undiversifiable risk analysis

We now shift our focus to the analysis of systematic risks, to investigate whether this kind of risk exists in the cyber domain and identify correlated indicators for consumers and enterprises that can help us to measure its significance together with the differences between the two classes.

To this end, we employ regression analysis by constructing several models that simultaneously use a combination of host attributes as regressors, thus controlling for conflicting explanatory variables when modeling the risk of encountering malware. We detail the model generation in section 4.4.1 and deeply discuss each risk factor in the subsequent sections.

4.4.1 Model generation

We postulate that the monthly risk of encountering malware for one host is influenced by a combination of the following seven independent variables: active days and hours, file-request volume, reputation and number of installed vendors, geographical location and whether or not malware has already been detected on the machine the month before.

Our objective is to obtain a Log-Odds distribution for the dependent variable Y, that expresses the odds —the ratio of successes (host encounters malware) and failures (host is clean)— as a linear combination of the regression variables. Since Y is monthly given in our telemetry as a boolean value (i.e., host encounters malware or is clean), we transform it as to obtain a count by bucketing numerical variables (days, hours, files created, vendor number and enterprise size) into bins to reduce granularity, grouping all the machines that share the same combination of values, and counting how

Host Attribute	Bin Category	Malware family	Consumers Monthly Odds		Enterprises Monthly Odds	
			μ	σ	μ	σ
Activity Days Ref: [0-4]	8-12	Any	2.78	0.45	1.59	0.30
	4-8	Any	2.10	0.19	1.44	0.19
	12-16	Any	3.26	0.67	1.82	0.57
	16-20	Any	3.58	0.85	1.91	0.71
	20-24	Any	4.01	1.11	1.97	0.75
	24-28	Any	4.15	1.25	1.79	0.73
	28+	Any	4.51	1.33	1.85	0.48
	28+	Virus	4.26	1.40	3.73	1.89
Activity Hours Ref: [0-3]	3-6	Any	1.34	0.09	1.02	0.20
	6-9	Any	1.57	0.32	0.95	0.14
	12-15	Any	1.25	0.47	0.88	0.21
	15-18	Any	1.35	0.38	0.98	0.23
	18-21	Any	1.59	0.49	0.99	0.39
	21+	Any	2.65	1.67	1.32	0.56
	18-21	Adware	1.68	1.32	0.63	1.46
	21+	Adware	3.30	2.23	0.08	0.25
File-volume Activity Ref: [0-1K]	1K-2K	Any	1.05	0.07	1.19	0.26
	3K-4K	Any	1.64	0.33	1.33	0.54
	5K-10K	Any	2.21	0.55	1.59	0.90
	10K-50K	Any	3.19	1.05	1.85	0.79
	50K+	Any	4.77	1.23	2.34	1.38
	10K-50K	Adware	9.67	3.87	2.62	1.77
	50K+	Adware	13.52	4,71	9.79	3.76
Vendors Ref: [0-20]	20-40	Any	1.11	0.04	1.09	0.12
	40-60	Any	1.22	0.06	1.30	0.28
	60+	Any	1.39	0.09	1.54	0.55
	60+	Adware	1.46	0.31	4.86	3.74
	60+	PUP	1.56	0.09	3.37	1.06
Reputable vendors only Ref: No	Yes	Any	1.00	0.05	0.99	0.25
	Yes	PUP	0.98	0.05	0.82	0.06
	Yes	Virus	0.64	0.04	0.70	0.09
Repeat player Ref: No	Yes	Any	1.77	0.77	1.33	0.49
	Yes	Adware	8.33	3.15	5.86	1.14
	Yes	Virus	2.21	1.03	5.50	2.56
	Yes	Worm	10.56	2.45	8.44	2.82
Geographical location Ref: NA	AF	Virus	6.35	2.10	12.14	2.49
	AS	Virus	4.19	0.51	9.72	1.21
	AF	Worm	20.77	2.61	18.59	4.61
	AS	Worm	5.39	0.23	9.49	2.31
	OC	PUP	0.86	0.18	1.25	0.45
	OC	Trojan	1.04	0.10	0.82	0.22

Table 4.7: Odds ratios of encountering malware according to our regression models

many of them are infected or clean.

We then make use of Generalized Linear Model (GLM)s [78], test them in different configurations, and analyze the outcome of several goodness-of-fit quantities (Pseudo R-Squared, Log-Likelihood, Dispersion, and the estimation provided by the Akaike Information Criterion (AIC)). We achieve the best results when modeling the risk of malware encounters Y as a Binomial distribution using a Logit link function. The analysis of the pseudo-R-Squared values obtained when modeling the different malware classes along the year revealed that, on average, between 68.4% and 89.9% of variance in the encounter rate is explainable by the chosen control variables.

Once the model has been fitted to the data, the extent to which the independent variables influence the dependent variable is captured by their regression coefficients. In particular, for each regressor, we select a bin (e.g., 0-4 days) or categorical value (e.g., North America) as a reference baseline, and express the odds ratio of other bins or values to derive the attribute's importance.

We separately model consumer and enterprise machines. We are aware that comparing the *magnitude* of odds ratio from models that use different samples from different populations may introduce an error [166]. However, our ultimate goal is to analyze the *trends* within each segment —odds ratio increase, decrease or fluctuations— and hereinafter we never directly compare the magnitude of the coefficients between consumers and enterprises.

In our experiments, we consider each month separately, as data are monthly aggregated due to anonymity constraints. We run a separate model for each month starting from February, by only considering hosts that have been active all the 12 months (11.7M consumer machines and 2.8M hosts of 33.7K enterprises), as to have information of the previous-month clean/infected state always available. At first, we define one host being targeted by malware if it encounters any kind of malware in that specific month. In addition, we separately consider and model five different malware classes —Adware, Trojan, PUP, Virus, Worm— to explore any variations in host-attribute importance or differences between consumers and enterprises when narrowing down the analysis to a specific class. In Table 4.7, we report the average μ and the standard deviation σ of the odds ratio along the 11-month period for the most explanatory cases that we discuss in the following sections. We note that all reported values are statistically significant with $p < .001$ for all the months of the considered period.

We do not include the enterprise size (i.e., number of hosts) and its industrial sector in the previous experiment, as these regressor variables are not available for consumers. In fact, the odds analysis of models that have

been constructed with different variables is statistically unsound [166]. We instead repeat the experiment by isolating enterprise machines and simultaneously modeling all the 9 attributes at our disposal for this segment of hosts. In Table 4.11, we only report the odds ratio of the two features that were added at this step. Also in this case, all reported values —including those that are not reported in Table 4.11— are statistically significant with $p < .001$ for all the months of the considered period.

4.4.2 Time-based activity

It is reasonable to expect that the longer a machine is active, the more likely it is to encounter malware. Indeed, the odds of detecting malware for consumers linearly increase with the number of active days, reaching a 4.51 factor with respect to the reference class for those active on average more than 28 days per month. While a similar relationship also exists for enterprises, the effect is much less pronounced and the odds reach a peak of only 1.97 when considering those hosts active between 20 and 24 days per month. Activity days represents a stable indicator along the months, as detailed by the low standard deviation in relative odds. A similar trend exists also with respect to the number of hours of activity per day – but in this case, both enterprise and consumers show a comparable magnitude and a similar random behavior for those machines active on average more than 9 hours per day.

To better understand this phenomenon, we separately assess the influence of activity days and hours in Figure 4.6. We split the machines based on their average uptime days and, for each of the 31 days, we compute the percentage of hosts that detect malware. Regardless of the number of days, we repeated the same task for the number of uptime hours. While for consumers the plot suggests that malware detection rates keep increasing alongside the number of active days, for enterprises this growth stops at around 20 days (roughly the number of working days in a month), but then the curve considerably drops for machines that are always running. The same trend is exhibited by looking at the daily hours of activity. In this case, the growth of the encounter rate stops at around eight hours for both groups (which again seems to align with the number of working hours in a day). As we clarify later in the section, these values seem to suggest that the active time changes with the *role* of the machine, and different roles may have very different encounter rates.

With these results in mind, we identified a set of machines for which the time-to-risk relationship was more regular. These include machines with up to eight hours of activity per day and, for enterprises, hosts that are

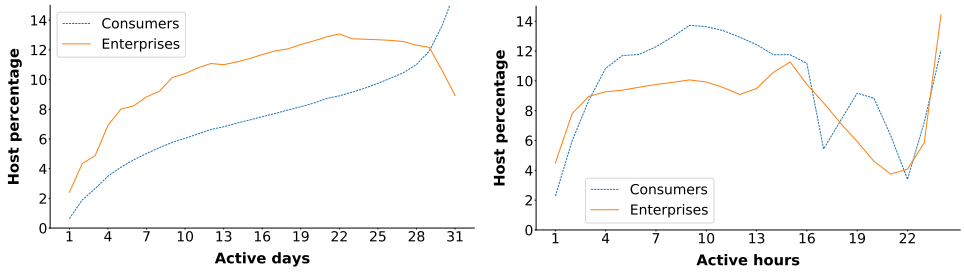
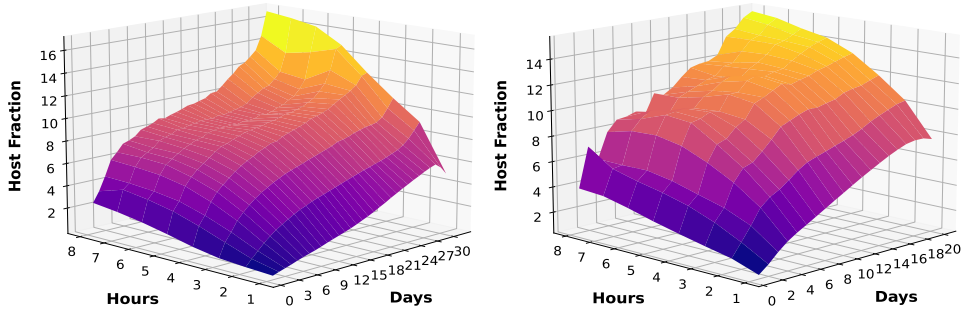


Figure 4.6: Disjoint influence of activity days and hours on malware encounters

active no more than 20 days per month. This group accounts respectively for 96% of the consumer hosts and for 73% of the enterprise machines. Figure 4.7 shows the joint influence that activity time has on the Regular Group: for each day X and each hour Y, the point on the 3D surface is given by selecting the machines active for X days and Y hours on average, and computing the percentage of those that detect malware. Interestingly, both plots follow a smooth behavior according to the one of the two control variables, confirming the goodness of time activity as a risk indicator for this type of machine.



(a) Consumers (b) Enterprises

Figure 4.7: Joint influence of activity days and hours on malware encounters

For machines in the regular group, we also computed to what extent each additional day or hour of activity increases the odds of encountering malware by fitting a model that considers days and hours as integer variables, while keeping unchanged the other regressors. We measure that for any additional day of activity the odds of encountering malware increases by 4% for consumers and 3% for enterprise machines. An additional hour

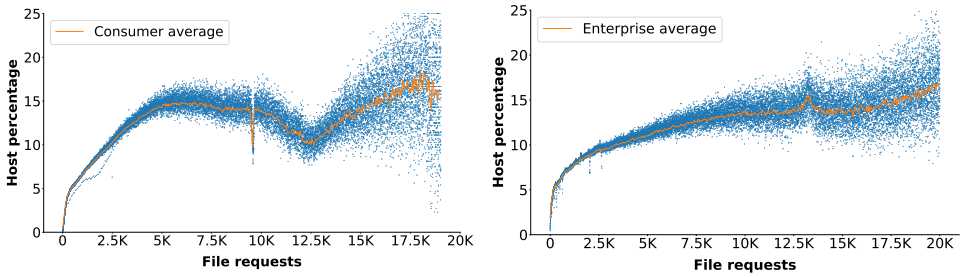
of daily activity results instead in an additional 17% and 6% extra risk. At first, both results suggest that adding more daily uptime has a stronger impact than adding more days of activity, but we need to keep in mind that machines in the regular group have a maximum of 8 hours of daily activity vs 20 (for enterprise) and 31 (for consumers) days per month. If we repeat the experiment by considering only specific classes of malware, in the case of number of active days we find a consistent behavior with the general case in both odds magnitude and increasing trends.

The odds related to number of active hours per day deserve instead special attention. In fact, we observe that for enterprise machines running more than 8 hours per day (i.e., the threshold we identified for the regular groups), the odds across all malware classes are lower than for hosts active fewer hours per day. We speculate that the reason is that those machines are likely dedicated to performing not-interactive tasks (e.g., servers). This hypothesis is confirmed by looking at the almost-zero odds of encountering Adware in enterprise machines that are always running: since this particular malware is usually shipped during software installations or web-browsing activity, very low odds of suffering this kind of infection can be explained by the lack of this kind of tasks. On the contrary, we observe a decrease-increase behavior for consumers, an indicator that those machines are probably used in both automated and interactive fashion.

4.4.3 File-based activity

As we already mentioned, the machines in our dataset routinely query a centralized system to assess the reputation of new objects: by monthly counting the number of these requests, we build a second metric for host activity and correlate its magnitude to the odds of malicious program detection. We find that the odds of detecting malware steadily increase with the level of activity in terms of files generated for both consumer and enterprise hosts and across malware families. This relationships does not vary month by month, as confirmed by the very low standard deviation reported with the mean. While we observe a similar magnitude in the odds of machines that generate less than 5K files per month, the effect of a greater file-volume activity (5K+) more consistently impacts consumer hosts. At its extreme, we observe that the odds of infection reach twice those of enterprises when selecting machines that generate a very high file-volume activity (50K+).

In Figure 4.8, we provide the reader with a visual representation of the relationship between files generated and malware encounters: for a given number X of file requests, we group the machines that queried the centralized system exactly X times in a month, and compute the percentage Y of



(a) Consumers

(b) Enterprises

Figure 4.8: File volume influence on malware encounters

those that encounter malware. The orange curve in the graphs provides an indication of the underlying trend, and it has been obtained by sampling the percentage every 100 values. The two figures reveal a similar logarithmic trend for both corporate and consumer machines. While for a low number of queries (up to roughly 5 K for consumers and 2.5 K for enterprises) a rise in the file-based activity entails a severe increase in the malware encounter rate, this effect gets weaker as we move in the right part of the plot.

4.4.4 Software vendors

We now measure to what extent various machine profiles might have an impact on the overall risk. We achieve this by looking at the set of software installed on the computers, extracting the vendor name from the publisher subject that can be obtained from signed binaries. On the vast majority of computers (around 80% for both groups), we identify software that is signed by between 10 and 15 different publishers. The maximum numbers of publishers identified on a single machine were 2312 for consumers and 349 for corporations. We first test whether an increasing number of software vendors implies a higher risk of detecting malicious programs. The rationale behind including the vendor number as a regressor in our model is that the odds of encountering malware –and in turn suffering from security issues– may raise according to the number and diversity of software installed in a system.

Our modeling reveals that a relationship exists between the two variables, and that enlarging the set of software installed on a machine results in higher odds of encountering malware. For instance, consumer and enterprise machines with a number of vendors between 20 and 40 are 1.11 and 1.07 times more likely to be targeted by malware than those with less than 20 signers. Odds increase to 1.22 and 1.18 for hosts with a number of

vendors between 40 and 60, and reach 1.39 and 1.22 for those with more than 60 vendors. Once again we measure a very low standard deviation, which suggests that results persist for all the considered months. When restricting to Adware and PUA, we find that the presence of a very high number of vendors entails higher odds ratios in enterprises (4.69 and 3.44). This difference is not reflected for home users, for which the magnitude of odds follows the general case.

We further dive into the relationship between a diverse set of programs and malware encounters by dividing both consumer and enterprise machines into groups based on the number of different software vendors installed. For each group in which we have at least 100 elements, we compute the fraction of hosts that encountered malware at least once.

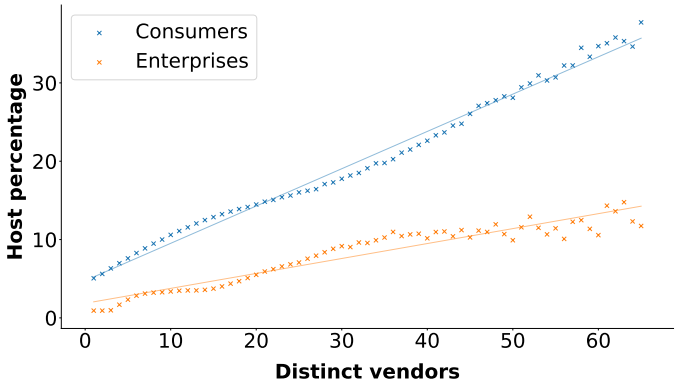


Figure 4.9: Relationship between the number of distinct vendors installed and hosts that encounter malware.

In Figure 4.9 we report these percentages together with straight lines, that represent linear regressions obtained using the least-squares method with a mean squared error of 1.86 for consumers and 4.75 for enterprises. Again, the diversity of software installed on the computers positively and *linearly* correlates with the rate of malware encounters. This is true both for consumers and enterprises, with the difference that the slope associated with the consumer trend is steeper than the one of enterprises. This discrepancy is also reflected by the higher odds ratios in the former group. We can justify this behavior with the fact that in enterprise contexts, even if a user were to install a diverse set of applications, each of them would probably serve the purpose to carry some tasks related to her job: indeed, with the existence of security policies, users are less likely to install software from dubious origin on the machines provided by their employers.

As a further insight, we also consider whether the nature of the installed

software influences the odds of malware encounters and whether that could be used to profile the role of the machines that installed them. We rank the top 20 vendors in our dataset based on the number of hosts on which they appear and report their list in Table 4.8.

Consumers	Enterprises
Microsoft Corporation	Microsoft Corporation
Symantec Corporation	Symantec Corporation
Google Inc	Google Inc
Apple Inc	Adobe
Adobe	Intel
Dell Inc	Oracle America Inc
Mozilla Corporation	Citrix Systems Inc
Intel Corporation	VMware Inc
NVIDIA Corporation	ESET
HP Inc	Mozilla Corporation
McAfee Inc	Cisco Systems Inc
Dropbox Inc	Hewlett Packard Company
Hewlett Packard Company	Lenovo
OracleAmerica Inc	Pulse Secure LLC
ESET	Dell
Garmin International Inc	Sun Microsystems Inc
Wild Tangent Inc	Apple Inc
Valve	NVIDIA Corporation
CyberLink	LogMeIn Inc
Lenovo	CrowdStrike Inc

Table 4.8: Top-20 vendors for consumers and enterprises

Our hypothesis here is that the machines that installed only those could be used as a control group, as they might belong to regular user profiles who only use common software, such as browsers, document editing tools and such. We therefore create two different profiles, isolating machines with only top-20 vendors installed from the rest: while for enterprises this set is composed of around 12% of the active hosts, this percentage rises above 42% for consumers. In our tests, we found that a higher vendor reputation has a negligible contribution to lowering the odds of encountering malware. Indeed, we register no changes in odds for consumers ($\mu = 1.00$ and $\sigma = 0.05$) and a small decrease for enterprises ($\mu = 0.99$ and $\sigma = 0.18$). However, we register a more significant impact when modeling malware classes as PUA (0.98 consumers - 0.77 enterprises) and Adware (0.64 consumers -

0.65 enterprises): in this case, the presence of only reputable vendors is an important factor that contributes in lowering the odds of encountering particular families that are usually shipped with dubious software.

4.4.5 Repeat players

We now assess whether being a repeat player has an impact on the odds of encountering malware. When fitting the model for a specific month, we consider a machine being a repeat player if malicious software was detected on it the month before. Our hypothesis is that repeated encounters with malware can be a sign of users' hazardous behaviors or of their poor security practices during the year under analysis.

In fact, we found a difference ($\mu = 1.77$ for consumers and $\mu = 1.48$ for enterprises) between the odds that a recidivist host will encounter malware versus a clean machine. The importance of this risk factor and the differences between home and corporate users increase when considering malware classes as Adware, Worm and Virus. When looking at consumers and at the first two cases, repeat players are 8.33 and 10.56 times more likely to encounter malicious software than machines that were clean the previous month. Although odds increase also for enterprises, here we register factors of lower factors (3.01 and 7.06).

4.4.6 Geographical location

Previous works show that the number and types of malware that computers encounter vary greatly across countries [163, 74, 162]. To verify these findings, we consider the continent in which one host is located as a regressor variable, and model how the odds of encountering malware vary with the geographical location.

When considering all malware categories, we register the same order of odds magnitude both across countries and types of machines. On the contrary, geographical location constitutes a considerable risk factor when restricting to Worms and Viruses. For those classes, we measure comparable odds in North America, South America, Europe and Oceania, but register a massive increase in continents like Africa (> 20 for Worms, > 6 for Viruses) and Asia (> 5 for Worms, > 4 for Viruses) for both consumers and enterprises. We refer the interested reader to Tables 4.9 and 4.10 for a complete geographical breakdown of malware classes.

This result is in line with what reported in a previous study by Mezzour et al. [162], who found a predominant prevalence of these two classes in Sub-Saharan Africa and South Asia. In the opposite direction, we find that

	Consumers					
Country	Trojan	PUA	OT	Worm	Adware	Virus
Africa	64.01	15.36	7.15	**9.3**	0.8	**3.38**
Asia	**66.71**	14.47	10.94	3.81	1.47	2.61
South America	60.17	23.62	12.49	1.5	1.65	0.57
Europe	56.57	**25.51**	14.4	0.99	2.22	0.31
North America	58.04	24.8	13.68	0.61	2.59	0.27
Oceania	58.17	22.19	**15.67**	0.93	**2.8**	0.25

Table 4.9: Geographical breakdown of malware classes for consumers

	Enterprises					
Country	Trojan	PUA	OT	Worm	Adware	Virus
Africa	77.05	**5.80**	4.87	**9.57**	0.34	**2.36**
Asia	81.76	5.59	**6.63**	3.28	**0.45**	2.29
Europe	87.42	4.88	6.21	0.94	0.3	0.26
North America	**93.39**	2.34	3.53	0.44	0.17	0.12
Oceania	90.88	3.17	4.84	0.72	0.2	0.2
South America	86.64	5.18	5.72	1.42	0.4	0.65

Table 4.10: Geographical breakdown of malware classes for enterprises

machines in Oceania have lower odds of encountering Trojans and PUA. Here, we find that the odds home-users facing PUA are reduced by a factor of 0.86 with respect to those in North America. A similar result holds for corporate machines whose odds ratio of encountering Trojan is 0.79.

We also tested whether there exist geographical regions where many machines encounter some malware families that appear very rarely elsewhere. To analyze this aspect, we first ranked all the signatures in our dataset based on the number of distinct hosts on which they have been detected. We then isolated the top-100 labels among behavioral signatures, PUA and Adware, and the remaining set of malware and, for each label, we broke down the machines that have encountered it across continents (Figure 4.10).

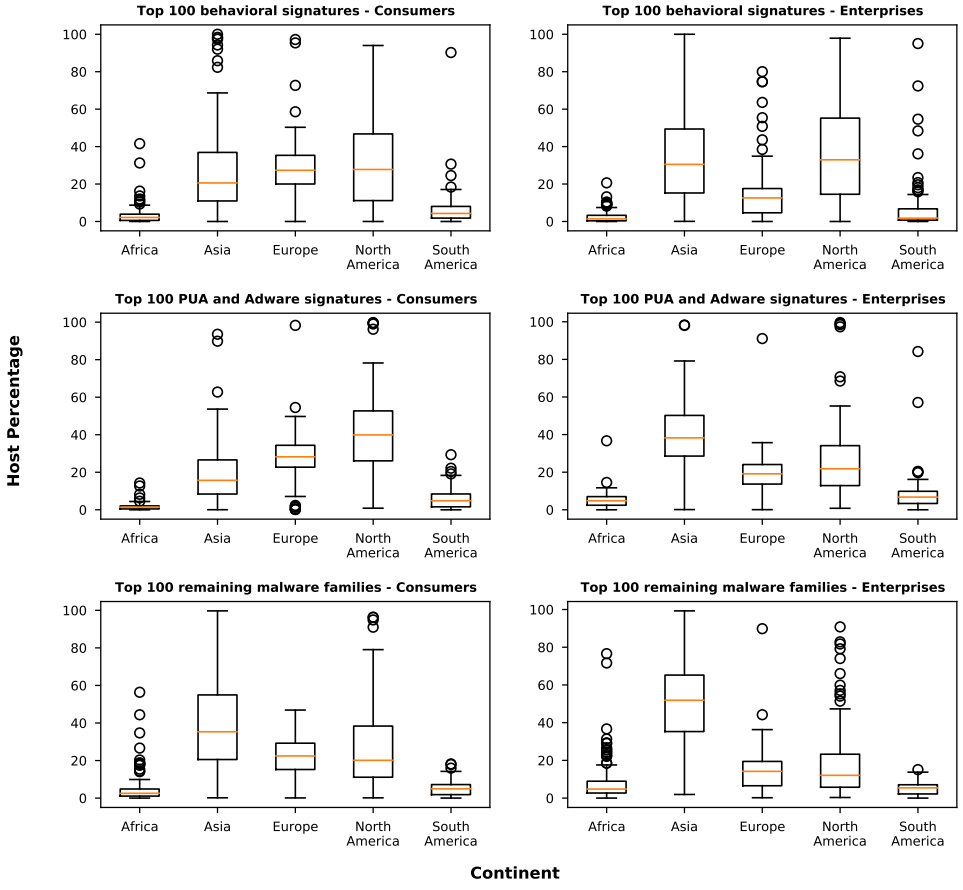


Figure 4.10: Breakdown of top 100 behavioral signatures, PUA and Adware, and remaining malware families. Percentages are sorted according to the number of distinct hosts on which the signatures have been detected.

Although we identified some differences, machines located in North America, Europe and Asia encountered the top-100 signatures with a similar frequency, while Africa and South America follow different behaviors.

After discarding generic cases, we observe that the family of the trackware *TransitGuide* (218 K hosts), developed to monitor browser activity of the targets, and of the Trojan *Kotver* (122 K hosts), that performs click-fraud operations in order to generate revenue for its authors, are almost exclusively detected in consumers located in North America (97% and 92%). At the same time, the Adware families of *KpZip* (22 K machines for consumers and enterprises) and *Funshion* (19 K consumer and 11 K enterprise machines), both created with the aim of displaying ads to profit from user clicks, are mostly encountered by computers located in Asia (92%).

4.4.7 Enterprise size and industrial sector

We finally focus our analysis on the risk profiles of enterprises with different sizes and industrial sectors. As reported in Table 4.11, the odd ratios related to small, medium and large organizations slightly differ from the baseline of micro firms. In addition, we do not observe any trend that relates an increasing number of hosts to higher or lower odds of malware detection, but instead register a fluctuating value when considering any malware class as well as when narrowing to specific categories. This suggests that the enterprise size is not correlated with the likelihood of malicious software encounters.

To get a clearer picture of this relationship, we decide to separately consider the enterprise size as a risk factor. Figure 4.11b shows a scatterplot in which each blue dot represents a separate enterprise, and on the axis we report its size (i.e., number of computers) and the fraction of its machines that encountered malware at least once in the one-year period of our experiments. Green crosses indicate clean enterprises, i.e., companies whose hosts do not encounter malware in the considered timeframe. The orange line shows the average among companies of the same size, considering both clean entities and those that encounter malware. We also plot a dotted line showing the average consumer rate—i.e., the ratio of consumer machines that had at least one encounter (9.8%)— with the aim of detecting whether the consumer encounters distribution is more similar to that observed in enterprises with a particular size.

In line with the insights gathered analyzing the odds ratio, the figure depicts an almost constant trend, slightly above the consumer line, with a flexion of the curve for those companies with sizes lower than 50 machines or higher than 100 K hosts. This may seem to suggest that small (<50) and

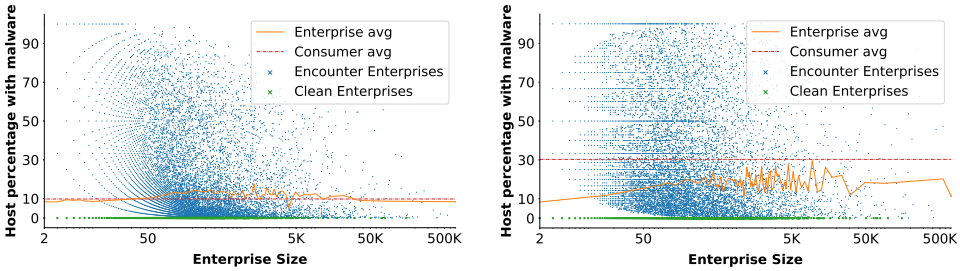
Host Attribute	Bin Category	Malware family	Enterprises Monthly Odds μ	σ
Enterprise Size Ref: [0-10]	10-50	Any	1.09	0.20
	50-250	Any	1.04	0.19
	250+	Any	0.98	0.49
Enterprise Sector Ref: Information Technology	Consumer Discretionary	Any	1.55	0.47
	Consumer Staples	Any	1.02	0.25
	Energy	Any	1.82	0.51
	Financials	Any	1.01	0.36
	Healthcare	Any	0.93	0.30
	Industrials	Any	1.32	0.28
	Materials	Any	1.48	0.38
	Telecommunication	Any	1.37	0.66
	Utilities	Any	1.56	0.26

Table 4.11: Odds ratios of encountering malware according to our regression models for enterprise size and industrial sector

large enterprises (>100 K) tend, in proportion, to have a smaller number of computers that encounter malware. However, the difference is very small and the Pearson correlation coefficient for size and the fraction of hosts that encounter malware is 0.01, indicating a negligible relationship between the two. Once again, this is a sign the number of machines in enterprises is not correlated to how much malware is detected.

To further investigate this aspect, we decided to focus our analysis only on those machines that were active for each of the 12 months of our experiment (2.8 M hosts of 33.7 K distinct enterprises). The rationale behind this choice is that hosts active only few months have less likelihood of reporting detections, thus lowering the average encounters rate we are interested in measuring. In Figure 4.11b, companies are still represented by blue dots. However, while the x-coordinate indicates the enterprise size (as in the previous case, obtained considering all machines), the y-coordinate is computed by considering only hosts active 12 months, and thus dividing those that encounter malware by their total number.

Interestingly, the effect of this filtering is more pronounced for consumers, where the percentage of machines that encounter malware raises to 30.3% (+ 20.4%), while in enterprises we register an average of 21.5% (+9.5%). We also observe a discrete gap between small organizations (<50) and those with a number of hosts comprised within 50 and 500 K: while



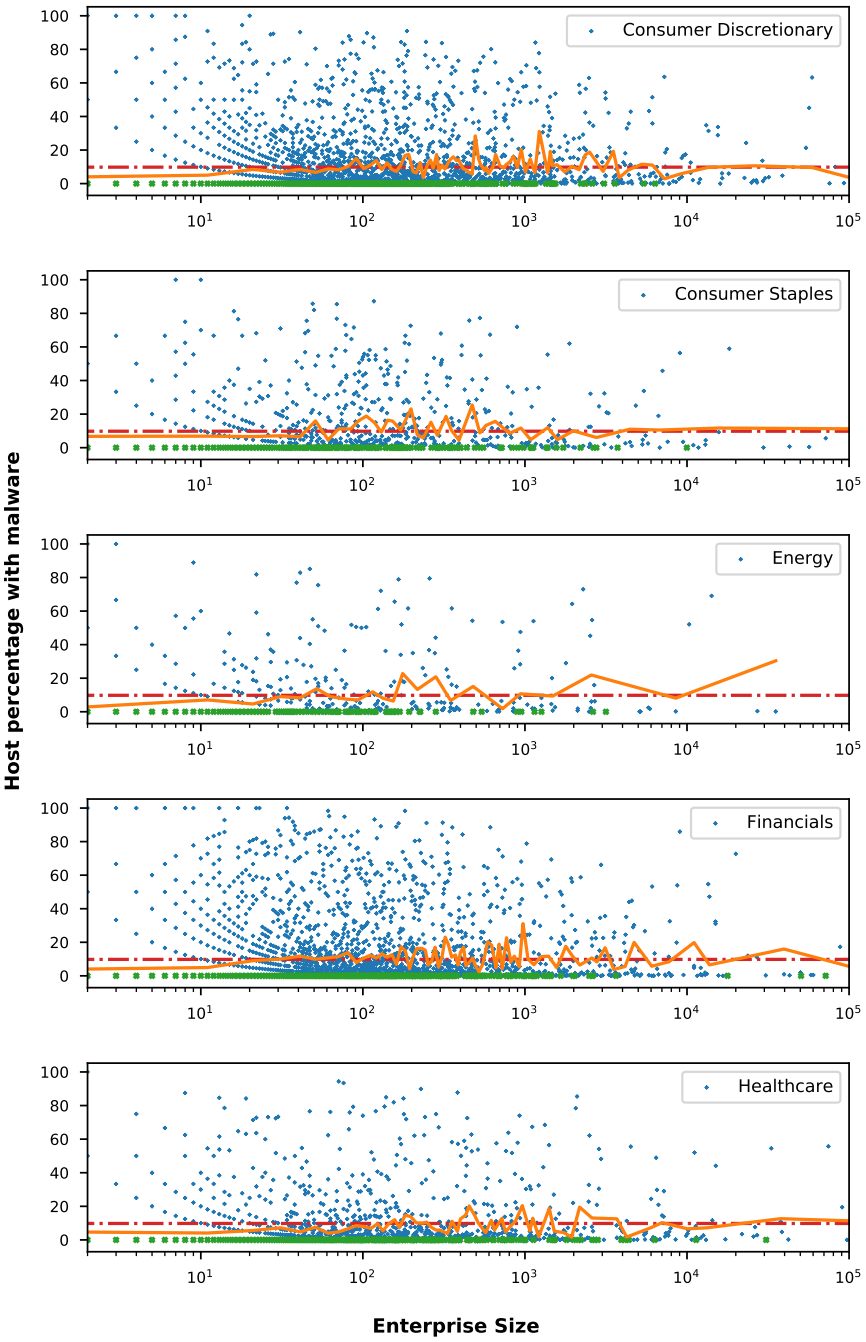
(a) All hosts

(b) Hosts active 12 months

Figure 4.11: Relationship between enterprise size and the fraction of hosts that encounter malware, computed for any host and for those active each of the 12 months of our experiment

for the former the mean stays around 16%, in the other case it reaches 23%. While this may indicate the existence of a relationship between enterprise size and malware detection rate, overall we still observe a very low Pearson correlation coefficient (0.02). In fact, excluding the companies with less than 50 machines, the remaining set of organizations (> 50 and $< 500K$, i.e., 92% of the total) exhibit an almost constant trend regardless of their size.

To gather further insights, we verify whether the industrial sector affects the relationship between the size of enterprises and the malware encounter rate. For this, we compute the Pearson correlation coefficient to measure the extent to which an increase in enterprise size leads to a higher number of hosts that detect malware. We also report in Figure 4.12 a plot for each sector. Again, we do not observe any general correlation, similar to the one obtained by looking at all enterprises (0.01), a sign that the number of hosts alone does not play a very important role in explaining the encounter rate.


Consumer Discretionary
Consumer Staples
Energy
Financials
Healthcare
Host percentage with malware
Enterprise Size
0
20
40
60
80
100
10^1
10^2
10^3
10^4
10^5

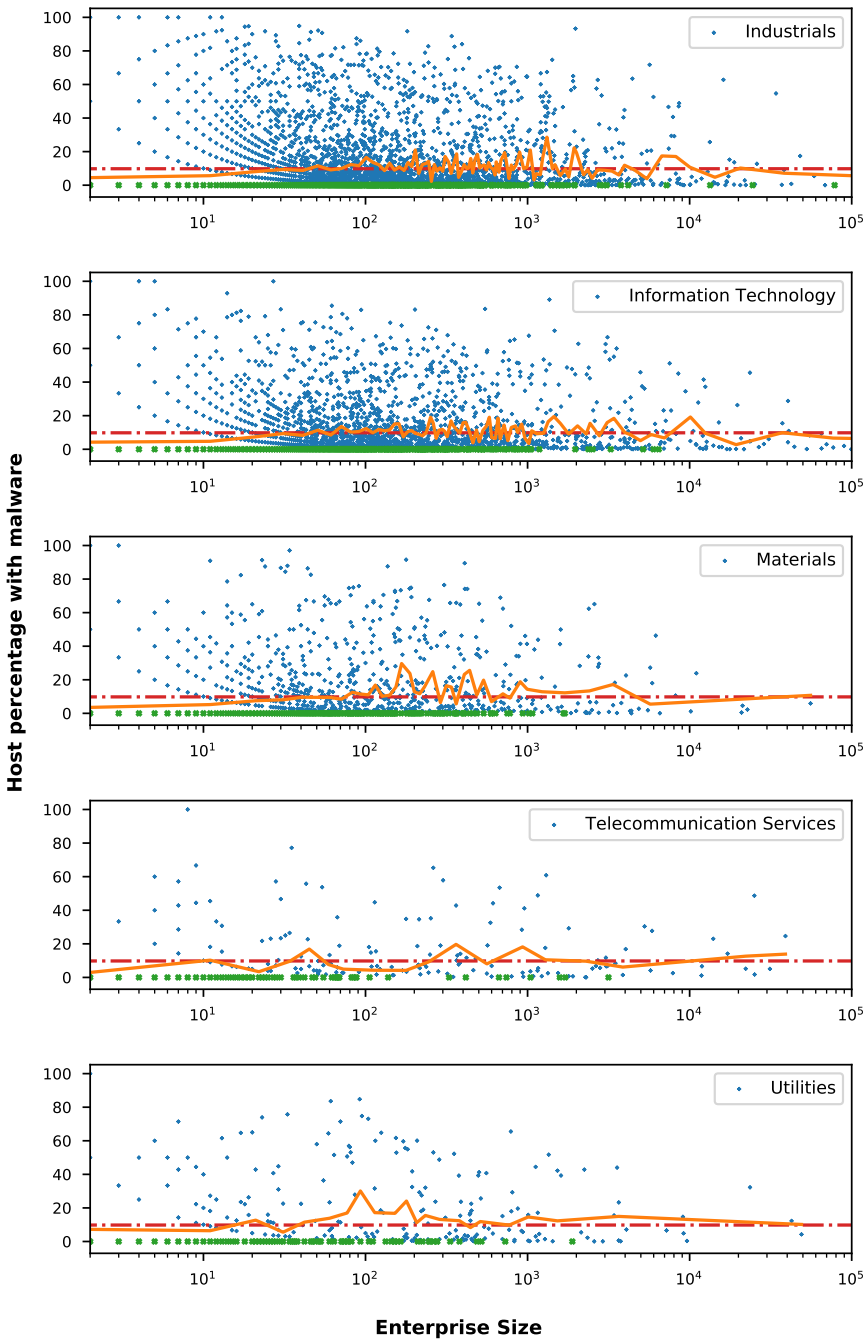


Figure 4.12: Relationship between the enterprise size and the fraction of hosts that encounter malware. Each plot represents a distinct sector.

In addition, we conduct a test to verify whether a statistically significant difference exists among the distributions in Figure 4.12 across industrial sectors. We opt for a non-parametric Kruskal-Wallis one-way ANOVA rather than a parametric one-way ANOVA, as this allows us to relax the one-way ANOVA assumption of data normality, which is not met in our case. The Kruskal-Wallis test assumes that a) the independent variable (enterprise size) has two or more independent groups; b) the measurement scale of the dependent variable (ratio between hosts that encounter malware and enterprise size) is ordinal, ratio or interval; c) the observations within a group and among groups must be independent. d) no data distribution assumptions if the test is used as a test of dominance, i.e., to verify whether at least one group stochastically dominates another one. With those assumptions verified, we run the test, our null hypothesis being that the samples come from populations with the same distribution. We obtain a test statistic H = 13.75 (p=.13), values that do not allow us to reject the null hypothesis: we conclude, once again, that the malware encounter rate based on enterprise size is not influenced by its industrial sector.

To conclude the study of enterprise environments, we evaluate how the risk of encountering malware varies across organizations in the different fields. In this case, we consider IT as a baseline for comparisons when evaluating odds ratios. We measure that machines of firms in the fields of Consumer Staples and Financials show negligible differences with those in the IT segment (1.5% higher likelihood of infection). Overall, we also find that the Healthcare industry is the best sector with the odds ratio with respect to the reference segment being 0.93. On the other hand, firms dealing with Energy, Consumer Discretionary, Utilities, Industrials, Materials and Telecommunications reveal a higher likelihood of encountering malware, We end up to similar conclusions when narrowing the analysis down to specific malware classes.

4.5 Discussion and conclusions

Home-computer users and enterprises tend to face malware in two different ways: while consumers approach the problem in a reactive fashion, often relying on a single AV product to detect and block possible malware infections, corporations act in a proactive manner, installing multiple security products, activating several layers of defenses, and establishing policies among employees.

In the first part of our work, we investigate whether the different measures in the two environments have an impact on their risks. In other

words, we want to answer the question whether more security products, tools, policies and restrictions in the enterprise segment are effective to lower the risk of malware encounters. Globally, we measure for 144.9 M consumer machines and 226.4 M corporate hosts an encounter rate of 9.8% and 12.0% respectively. According to these results, home-machine users encounter slightly less malware than the counterpart, suggesting, at first glance, that all the choices that enterprises adopt are not effective in practice. However, we believe this first impression to be misleading: when considering all the available hosts in our dataset, a lot of them have been found to be active for only a few months, or even a few days, and these low-activity hosts are more prevalent among end-users than corporate machines.

When we restrict the two sets of machines to only those active every month of the year, we find an opposite result: around 30% of consumer hosts report malware encounters vs 21% of enterprise machines. If we go one step further and select only those machines that are active more than 20 days and 15 hours per day, the gap widens as 89% of consumers encounter malware against 53% of corporate machines. Moreover, we also found that the average consumer machine encounters a more diverse set of malicious files compared with its corporate counterpart, and this finding holds for all the malware classes considered in our study.

Security policies and restrictions also seem to have a relevant impact on reducing risks. Indeed, when analyzing the presence of Adware and PUAs, we report a concentration of such malware families 6 times higher in consumers, due to the freedom in installing any kind of software that this group of users has. Since the presence of less reputable programs is often a vehicle for malware, we believe the same findings apply also when considering other families. On the opposite, generic behavioral signatures (who might match unknown threats or suspicious files) are twice as likely to trigger in enterprise environments than in consumers hosts.

If on the one hand a good security posture and a better cyber hygiene are important to reduce the risk of malware encounters, on the other hand it is not the only factor to take into account. Indeed, the interconnected nature of our society, the use of third-party software and the sharing of the same networks expose all the classes of machines to undiversifiable and systematic risk, regardless of the number and type of security measures and policies in place.

In the second part of this work, we leverage the data at our disposal to investigate whether this portion of risk exists and provide quantitative indicators that can be used to measure its significance: for this purpose, we extract seven indicators for each consumer machine and nine for each

enterprise host that carry no information about its security level, and test their correlation with malware encounter risks. Interestingly, we find that height of them serve this purpose: host uptime days and hours can act as control variables for the encounter rate of a subset of regular machines; with a logarithmic relationship, the same holds for file-based activity; encountering malware over and over and being recidivist along time represents an important risk factor, which is even more pronounced when considering malicious categories as Adware, Virus and Worm; for the same classes, host geographical location can explain the risk of suffering from higher encounter rate; finally, we also verify the effectiveness of vendor number and reputation; For organization environments, we compare industrial sectors and spot those that have higher odds of reporting malicious software; we fail, instead, in proving any correlation between enterprise size and malware encounter rate, even when separately considering each industrial sector.

To our knowledge, no scientific or empirical work has looked at the systematic nature of cyber risks, although the topic is largely discussed in other domains. The existence and quantification of systematic cyber risks is an emerging problem among risk management experts and cyber-insurance underwriters, as the number of events that simultaneously affect a large number of hosts across different enterprises and countries is increasing every year. Hypotheses to explain it have also been advanced considering global-scale incidents and the subsequent market reactions: experts agree that factors and events such as common widespread vulnerabilities, infrastructure failure cascade, loss of integrity of trusted systems, concentrated dependencies and indirect attacks to central actors characterize its nature [33, 32, 37, 92].

Despite these conclusions, systematic risks need a deeper understanding for what concerns their underlying factors and likelihood. An objective analysis of the extent to which these indicators can explain cyber risks would definitely be beneficial for particular tasks, such as premium establishment for cyber insurance policies [199]. Indeed, in order to compute premiums, insurance carriers scale a base rate by factors depending on the enterprise size, industrial sector, and by considering whether or not the company had already suffered cyber security events (i.e., it is a repeat player). In this respect, our study shows that an assessment done considering the enterprise size as a factor may not be appropriate - we find no correlation with malware encounter- and that different indicators are needed to come up with a correct assessment.

In this work, we try, for the first time, to shed light on systematic risk indicators, by shifting the analysis at the host level and by using real-world

data telemetry. With the findings previously discussed in mind, we support the hypothesis that this portion of risk exists in the cyber scenario —in both consumer and enterprise context— and that the factors we identified can be used as good indicators to quantify it. We believe these insights can help both companies and academic researchers to better understand the global picture of malware encounters in the wild, and that our study can be used as a foundation for future works in the area of systematic risk.

Chapter 5

When data matters: Web Tracking From the Users' Perspective

5.1 Introduction

In this last contribution, we emphasize the importance of data sources and analysis perspective when assessing cyber risks. In particular, we focus on the web-tracking practice and its effects on users' privacy risks. Although this phenomenon has been extensively studied by web practitioners and researchers, all the previous works looked at it from the trackers' standpoint, leaving a lot of unknowns regarding the real impact of tracking on real users. In this chapter we fill this gap by considering real-world telemetry as data source, and analyze web-tracking and related risks from the users' eyes.

We split this chapter in two parts. First, we look at web-tracking from a time and frequency perspective (Section 5.4): for each user in our dataset, we estimate how long it takes to encounter a significant fraction of the trackers. We then perform a correlation analysis to understand what increases the privacy risk, discovering that there is an interesting relationship among privacy and security risks on the web.

In the second part of the study, we estimate what percentage of the user's browsing history is known to trackers and investigate how much this knowledge could be extended through real or hypothetical collaborations among different tracking companies (Section 5.5). For instance, our experiments show that the actual knowledge popular trackers have of the users' histories is almost double the estimate obtained by crawling the top Alexa popular domains, thus confirming the importance of a holistic approach when measuring cyber risks. We also shed light on the most efficient monitoring strategy and what sensitive information could be learned about the users because of their visits to particular classes of websites.

5.1.1 Why this study matters

Third-party web tracking was first introduced to support web analytics and advertisement [145] but evolved over the years into a very widespread phenomenon employed for a wide range of purposes. Currently, more than 90% of the websites include at least one tracking script [104, 204], resulting in a multi-billion dollar business [142, 165, 122, 260] where many companies earn huge amounts of money by selling or leveraging the data collected from users.

Previous works showed that users are aware of this practice and have rightfully started to complain about the amounts of online tracking present on the web [200, 247]. On the other hand, those studies also reported that participants are surprised when confronted with detailed information about the extent and prevalence of web tracking [247, 161]: once aware of the

actual impact, users' general attitudes often resulted in being at odds with such practices [161], and in stronger intentions to take privacy-protective actions [247].

The scarcity of works that investigate how impactful web-tracking is for Internet users can explain why, despite being aware of the practice, only a few are conscious of the actual implications and take the appropriate actions to protect themselves. For example, only 7.74% of the browsers' market share belongs to privacy-centered browsers [133], 8.5% of the users reported the use of tracker-blocking tools [247], and just 0.59% of them use privacy-preserving search engines [209]. We believe that studies that look at the problem from the users' perspective to identify concrete evidence for its seriousness could be immensely helpful to the general population.

In fact, as mentioned earlier, web tracking is not a new phenomenon on the Internet and a wide corpus of previous works have analyzed both the impact and the prevalence of web tracking. However, previous studies have assessed its size by measuring how many websites contain trackers, or how many websites are known to a given tracking company [172, 49, 224, 125]. As we will demonstrate in this work, knowing in how many websites a tracker is detected is difficult to translate into how much the tracker knows about the average user. More than that, our experiments show that measuring the coverage by only crawling top-ranked websites results in gross underestimation. In reality, users visit only a tiny fraction of the Internet websites – typically composed of a mix of popular (such as social networks, search engines, news) and less popular sites (such as regional pages, friends blogs, or specific work-related sources). As a result, it is still unknown what fraction of the user's browsing history is known to web trackers or what fraction of trackers are encountered by each user.

In this work, we aim at filling this gap by complementing the current knowledge on web tracking with real-user browsing behaviors. We leverage the telemetry of 250K users and the information collected by a large-scale crawling experiment to analyze the impact that web tracking has end-users located all around the world. Differently from previous studies, whose results are based on the analysis of the top websites listed on publicly available services [42, 43], the use of browsing telemetry allows us to exactly know when and which websites are accessed by users, without the need for distribution approximations. This allows us to precisely understand how often users encounter new trackers, how many different ones, and what amount of information each tracker knows about them.

5.2 Data Sources and Methodology

As in the previous work, our main dataset comes from the telemetry of NortonLifeLock [46]. The data, collected on the consumer hosts about the users' *web-browsing activity* is described in Section 5.2.1. We acquire the *category and risk score* (Section 5.2.3) for each domain in the telemetry, and detect the *trackers* present on the webpages by using a custom crawler (Section 5.2.2). We also take advantage of a *linkage graph* published by Sanchez-Rola et al. [203] about the information-sharing relationships among different trackers (Section 5.2.4).

5.2.1 Web-browsing telemetry

This dataset contains the web-browsing history of 250K users. The telemetry spans a period of 8 days and was collected from October 14th to 21st of 2019. The data includes a code that reports the country registered by the user when installing the AV software, a daily log with the list of domains browsed by each user, and the hour in which the request was performed. Overall, we count 2.35M distinct websites (0.8% were not accessible or offline), which finally accounted for 107M entries in the users' browsing history.

5.2.2 Website trackers

We identify the trackers that exist on the websites in our dataset through a custom crawling framework. The crawler is based on the open-source web browser Chromium, and uses a custom instrumentation developed using the Chrome debugging protocol (CDP) [82]. By connecting into its network tracing processes, we gather all the requests and responses performed by the browser during a web access. In order to avoid possible detections of our automated browser, we implemented the most recently-proposed methods [205, 206, 207, 98], also leveraged by other recent studies [201, 242]. When third-party scripts were loaded into each page we analyze the request, extract the destination domain and verify that the loaded entities were actually trackers by leveraging the tracker list used by Mozilla Firefox [168], and EasyPrivacy [99]. The two monitor different forms of tracking, such as web bugs, tracking scripts, and information collectors. Once the tracking domains are identified, we map the domain names to organizations based on three manually-curated lists: Disconnect [96], WhoTracks.me [83] and webxray [148].

We scanned the 2.33M websites in our telemetry using a server located in

the US and discovered 6,320 distinct tracker names. To account for tracker variability due to geographic locations, we deployed additional crawlers in three different countries from three continents. For this, we leverage a commercial Virtual Private Network (VPN) service [45]. Specifically, we looked at browsing histories of users from France (6213), Brazil (5152), and Australia (5603), and crawled 130.70K, 67.81K, and 126.73K websites from the respective country. We report the results and compare them with the data collected from the US in Table 5.1. We found that on average 80.3% of the websites include exactly the same trackers, while another 6.9% has only one additional tracker. To obtain further insights into the remaining websites that have more than one different tracker ($\sim$ 12%), we compute the Intersection over Union (IoU) coefficient between the two sets of trackers obtained by crawling from US and the respective location: the rationale is that a result close to 1 (e.g., $> .8$) refers to very similar organization lists; on the other hand, a value close to 0 (e.g., $< .2$) implies the opposite. We finally assess that around 95.5% of the websites show no or subtle differences in the trackers detected, whereas we detect a diverse tracking ecosystem only on a very small subset of 4.5% domains.

		US		
Country	Same trackers	± 1 tracker	IoU > 0.8	IoU < 0.2
France	84.42%	5.52%	0.46%	4.97%
Brazil	79.28%	6.84%	1.14%	4.56%
Australia	77.20%	8.04%	1.84%	4.04%

Table 5.1: Comparison summary between trackers detected crawling websites from US and France, Brazil and Australia

5.2.3 Website categories and risk

By using the public classification service from the same security vendor, we were able to assign a category to the websites in our telemetry [230]. The service supports over 60 languages and is composed of more than 300 specialized modules that disassemble web pages and analyze their components. The main features used to feed the classification algorithm are: webpage language, source code language, document type, character set, external link categories, content words, scripts and iframes. In addition, the categorization is fine-tuned by an offline system, which simultaneously analyzes multiple pages looking for connections and additional evidence to supplement what was collected in real time. Hypertext Transfer Protocol (HTTP)

referrer headers and hyperlinks are examples of attributes used in this phase.

In addition, to better investigate the impact of tracking and the prevalence of different trackers on websites that could be related to user's sensitive information, we selected a set of *sensitive categories*: Health, Legal, Financial, Sexuality, Political, and Religion. Our decision was guided by categories defined as sensitive in various data protection laws [105, 106, 140], and used in recent studies [202, 158].

Finally, we additionally assigned a security-related risk level to each distinct website in the telemetry by leveraging the rating service from the security vendor [229]. The service uses cloud-based Artificial Intelligence (AI) engines to categorize websites by combining multiple data sources. At first, historical information of the domain is used to detect the existence of malicious behaviors, e.g., whether its DNS resolutions belong to malicious networks and the website has already been identified as source of malware, scams or phishing. The webpage is then queried and the characteristics of its content together with features extracted from the server behavior are analyzed (e.g., shady file content, network errors, lie detector analysis). The AI algorithm then outputs a risk score between 1 and 10, going from websites with huge traffic and long history of good behavior (risk 1), through webpages with evidence of shady behavior (risk 5), to domains with solid evidence of maliciousness (risk 10).

5.2.4 Tracker relationships

A previous study [203] investigated the relationships among 810K actors during the creation and sharing of cookies through cookie chains. In particular, the authors shed light on the role of those acting as dispatchers of information, receivers, or cookies direct creators.

We manually extracted the dependency relationships of the top trackers from the linkage graph and its related table in their manuscript, and used them to evaluate information sharing between a sender and a receiver organization. In this measurement, we assume that this happens in all the cases, i.e., the former always shares any data with the latter: although for many of the relationships this does not match the reality —trackers share part of the information and not for all the webpages—, in our discussion we consider it as an upper bound in order to evaluate the worst-case scenario for some of our findings.

5.3 Dataset Statistics

The users in our telemetry span 214 of the 249 countries with an assigned ISO 3166-1 code [24]. More than 44% of the users are located in North America (with 38% of them in the United States). Asia and Europe follow with about 20% of the users each. In South America, Africa and Oceania we find the lowest percentages (less than 17% overall). We report the complete geographical breakdown in Table 5.2.

Continent and Countries	% Users	% Trackers	Categories
North America	44.16		
United States	37.08	80.30	92
Canada	2.63	59.84	92
Mexico	2.61	43.75	91
Asia	20.87		
Philippines	6.27	58.54	93
India	3.97	51.16	92
Malaysia	2.70	40.16	88
Europe	18.69		
Great Britain	4.10	66.33	91
France	2.35	51.17	90
Italy	1.72	46.71	88
South America	9.33		
Peru	2.05	43.83	91
Brazil	1.95	39.40	90
Colombia	1.91	38.62	85
Africa	5.00		
Nigeria	1.59	30.08	83
South Africa	1.15	39.76	89
Egypt	0.66	34.05	91
Oceania	2.74		
Australia	2.12	53.34	91
New Zealand	0.44	37.83	90
Fiji	0.10	22.37	75

Table 5.2: Overview of the continents and their top-3 countries ordered by percentage of users in our dataset.

On average, the median user is active slightly less than 6 days out of 8, and for a number of hours per day that ranges from 3 to 10. We report a graphical summary of users' activity in terms of mean browsing days and

hours in Figure 5.1.

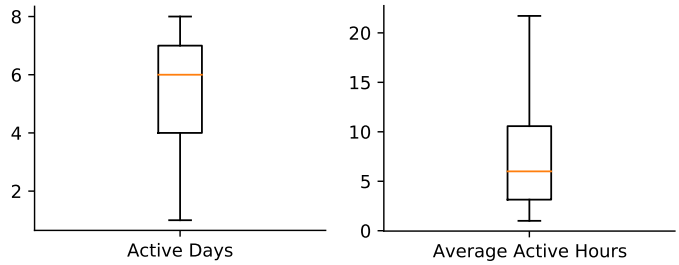


Figure 5.1: Overview of the average number of active days and average number of active hours per day for all the users in our dataset.

We further look at the aggregated users' browsing behaviors in our dataset: we detect that on average during the 8 days, users present a history with 406 entries, browse 19 distinct categories, 118 different webpages, visit more than once 59 of them, and encounter 3,170 trackers from 177 distinct organizations. Additionally, we measure that 93% of them have less than 10 trackers, and for a single webpage visited, users encounter on average 3.5 different trackers.

In Table 5.3 we provide a summary of both sensitive and top-10 categories in our dataset, sorted by the number of websites they encompass. Webpages related to users' *Health* are the most frequent among the sensitive categories, also reporting the longest list of trackers encountered (34% of the 6,320 trackers). On the contrary, the *Political* category, the smallest among the sensitive category in terms of number of websites, visiting users, and different trackers detected, shows the highest average of trackers. This suggests that fewer organizations focus on political websites but more consistently. We will come back to this comparison in Section 5.5, when we will discuss in more detail which and how much sensitive information the different trackers can obtain about users. Regarding the other, non-sensitive, categories almost the totality of users browse websites classified in the *Technology/Internet* and *Business/Economy* groups: we indeed detect in the pages of these two categories almost 50% of the tracking organizations.

We finally analyze the coverage of the top 20 trackers in our dataset, reporting the percentage of known history, websites and users who encounter them in Table 5.4, together with the average values for all the trackers. We point out to the reader the subtle difference between two recurrent concepts throughout the manuscript: when computing the known history percentage by a tracker, we refer to the portion of entries in our telemetry in which

Category	% Websites	Avg Trackers	% Trackers	% Users
Health	4.89	10.80	34.78	33.89
Sexuality	2.89	2.82	24.75	17.97
Financial	2.00	7.77	29.11	53.86
Legal	1.95	2.64	19.73	34.62
Religion	1.91	8.29	20.41	19.84
Political	0.52	14.25	16.66	11.58
Business/Economy	11.62	8.64	48.94	83.30
Technology/Internet	6.55	9.36	46.06	99.18
Shopping	6.37	14.52	38.32	58.56
Education	4.44	7.01	30.97	50.68
Suspicious	3.79	1.45	28.84	40.49
Entertainment	3.47	13.84	40.41	53.34
Travel	2.76	8.27	31.33	33.36
Search Engines	2.43	3.43	26.41	94.32
Restaurants/Food	2.24	18.90	27.07	21.85
Personal Sites	2.18	8.90	26.61	19.66

Table 5.3: Overview of sensitive (above) and top-10 (below) categories in our dataset

we detect the tracker —thus also considering revisited websites across hours and days. On the contrary, when reporting the known website percentage, we only consider the fraction of unique website IDs —i.e., we do not take into account revisited webpages.

At a glance, Google clearly stands out, being directly present in almost 73% of the websites in our dataset. The other top-20 tracking organizations cover on average 15.27% of users' history and 8.45% of the websites. From the users' perspective, almost all of them encounter at least once one of the top organizations in Table 5.4. Interestingly, while the average number of users reached by a single tracker is 3%, we measure that almost the totality encounters at least one tracker. The few exceptions – 419 users corresponding to 0.16% of the total – have a clean and not-tracked history. However, the small number together with the fact that those users only browsed an average of two different websites in 8 days, suggests that in practice *everyone* who browses the web is tracked to some extent.

It is also interesting to observe the difference between the two middle columns, i.e., the coverage in terms of unique websites and the one in terms of entries in the users browsing history. Google is the only tracker in which

Tracker	% History	% Websites	% Users
Google	63.07	72.33	99.76
Facebook	30.05	26.53	98.33
Microsoft	22.97	4.11	97.56
Adobe	19.92	7.83	97.42
Appnexus	18.91	5.27	97.58
Yahoo!	17.36	5.33	97.05
Twitter	16.73	6.10	96.85
Rubiconproject	15.16	4.52	96.79
Thetradedesk	14.54	3.61	96.37
Rapleaf	13.92	4.19	96.12
Casalemedia	13.68	4.26	96.61
Pubmatic	13.30	4.08	96.45
Openx	13.09	4.16	96.40
Mediamath	12.69	2.49	96.56
Drawbridge	12.41	3.30	94.39
Amazon.com	12.00	2.69	95.14
Akamaitechnologies	11.53	1.11	95.48
Linkedin	11.33	2.09	94.13
Quantcast	10.84	3.68	95.81
Taboola	9.65	1.45	94.24
Average	0.14	0.06	3.00
Untracked	20.07	23.11	0.16

Table 5.4: Coverage overview for the top-20 companies involved in tracking in our dataset

the first is bigger than the second, meaning that it is the only company that also covers many less popular websites that do not receive many visits. Microsoft is instead an example of a company that seems to focus mostly on popular sites, as shown by the fact that its history coverage is more than five times the one of websites.

5.3.1 Dataset Limitations

Although our telemetry is large and contains hundreds of thousands of users from almost every region in the world, it may still be subject to some selection biases. For instance, it only includes users who protected themselves by installing an AV product and opted in to share their data: users who decided not to opt-in due to privacy concerns could behave differently, being more

conscious with respect to tracking and high-risk websites. Furthermore, our entire telemetry comes from Windows machines. It is possible that users running other OSes (e.g., macOS and Linux) or browsing through mobile devices may exhibit a different behavior. Moreover, our data covers only 8 days of users' browsing experiences. As we will discuss in the following sections, users encounter the vast majority of the trackers already in the first day of browsing. Therefore, it is very unlikely that the final results would significantly get impacted with more data.

5.4 Standing in users' shoes

We start our analysis of web tracking by looking at the trends from the users' perspective. Our goal is to use our telemetry information to estimate how much, and how fast, real users encounter web trackers during their daily activity. We are also interested in finding whether some users are more exposed than others, or whether a certain class of online behavior leads to higher or lower privacy risks.

5.4.1 How long does it take for a user to encounter trackers?

To answer this first question we investigate the relationship between the time a user spends browsing the Web and the number of new trackers she encounters. To this end, we initialize a *cumulative* tracker set for each user. Then, for each cumulative i^{th} hour spent browsing, we add the new trackers encountered to the set and register its length variation from the previous time interval. Each i^{th} point of the blue curve in Figure 5.2 is then obtained by averaging the i^{th} values of all the users active at least i hours. For the j^{th} daily curve, the i^{th} hours close to the boundary with the next day refer to users active i hours in the j^{th} day —thus active almost all the j^{th} day. Activity does not refer to the day fraction (i.e., active only late night).

In a similar way, we maintain also a *daily* set for each user. For every j^{th} day, we add new trackers and register variations as for the cumulative case. We finally compute each of the i^{th} points for a j^{th} daily curve in Figure 5.2, by averaging the values of users active at least i hours in the j^{th} day. We do not include the daily plot of the 8th day in our telemetry because our data does not cover all its 24 hours.

The analysis of Figure 5.2 provides three important findings. First, the curve of new trackers per hour of activity follows a decreasing exponential distribution, with a drastic drop in the first 12 hours. Indeed, the average of new trackers encountered falls below 5 after 12 hours, below 2 after 22

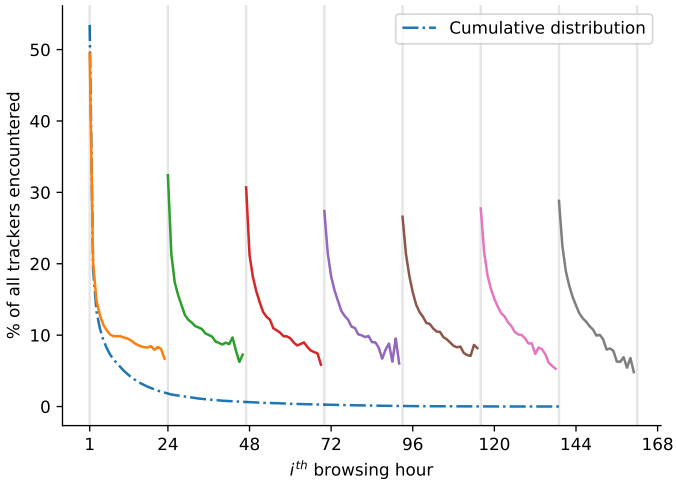


Figure 5.2: Cumulative and daily distribution of new trackers encountered per hour of activity.

hours and users encounter almost no new tracker after 35 hours of activity.

Another way to look at this data is to compute how many hours it takes for users to encounter a given percentage of all the trackers they encountered during the week under analysis (on average 177 trackers per user). In this case, on average after 2, 12, and 24 hours of activity users have already encountered respectively 50%, 84%, and 94% of their trackers.

The second interesting finding is that given a window of i hours (e.g., 24), users who are active for more consecutive hours encounters a higher number of trackers with respect to the others. This discrepancy is clearly visible in Figure 5.2, when comparing the first part of the cumulative curve with the daily curve of the first day.

For instance, we can consider two users that both have three hours of activity over a 24h window. The first browses the Web in three separate sessions of one hour each – in the morning, afternoon, and evening. The second browses instead for three hours straight in a single session. In our experiments, we noticed that the second user is more likely to encounter a higher number of unique trackers. And the reason is that sessions that are far apart are more likely to have larger intersections in the visited websites. In other words, the likelihood of revisiting the same websites and running into already encountered trackers is higher in those cases. On the contrary, users characterized by longer browsing sessions show higher variability in the websites and trackers encountered.

The third observation we can make from Figure 5.2 is that all daily

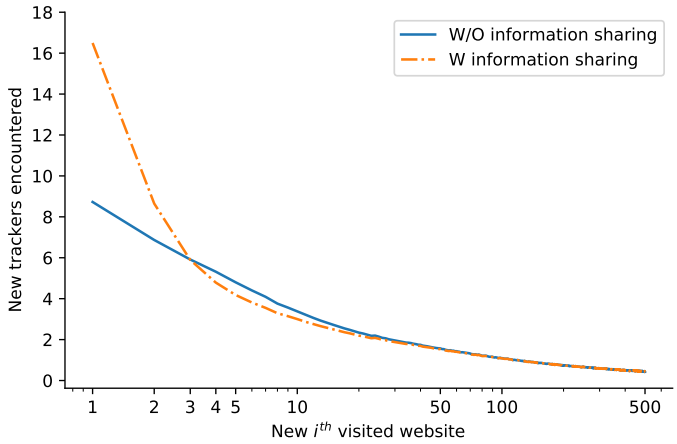


Figure 5.3: Average number of new trackers per new website

curves have really similar shapes, with a sudden decrease in the number of new encountered trackers in the very first hours. This suggests that, even if the user would restart with a clean browsing history every day, it would only take two hours on average to re-encounter 50% of all trackers. In other words, if a user encounters on average 177 different trackers per week, half of them are regularly encountered every day within the first two hours of web browsing.

So far we have captured the users' activity by counting the time they spend browsing. Another way to do that is to count the number of visited sites. The trend of how the newly encountered trackers evolves for each new website visited is summarized in Figure 5.3. The points on the blue curve are obtained by averaging the number of new trackers encountered for the i^{th} new visited website, among users who browse at least i distinct websites. The distribution in Figure 5.3 shows a similar trend of the corresponding cumulative curve when considering the hours of activity (Figure 5.2). The exponential shape has a maximum at 9 — suggesting that users encounter more than the average of 3.5 trackers when visiting the very first website, probably indicating a popular page with multiple trackers—, and quickly drops: after 20 different websites, users only encounter on average 2 new trackers. When computed in percentages, our data shows that by visiting 22, 100, and 300 distinct websites, the trackers encountered are respectively 50%, 75%, and 85% of the total encountered over the week.

However, this represents a *best-case scenario* that considers each tracker in isolation. In reality, trackers also exchange data with one another. Therefore, we complement our analysis by plotting a second curve, but this time

considering the relationships among the different actors indicated in Section 5.2.4. In this case, when we add a new encountered tracker to the set, we also add all other trackers that directly receive information from it [203]. This curve, in orange in the graph, represents a *worst-case scenario*. In fact, the fact that a relationship exists between two trackers does not imply that the two companies share *all* data about all users on all websites. Therefore, reality lies somewhere in between the two curves.

Even in the worst-case scenario, it is interesting to observe that the data shared among trackers exposes the users to a higher number of tracking companies for the first few visited websites. However, after around 20 websites the two curves overlap, showing that at that point the number of new trackers encountered by the user is independent from possible collaborations among trackers.

5.4.2 Is there a correlation among distinct visited websites and encountered trackers?

We now look at the correlation between the total number of distinct websites visited by a user and the number of encountered trackers. In particular, we are interested in finding (and comparing) those users that encounter a disproportionate number of trackers despite visiting a few websites, and those that instead encounter a few trackers while visiting many different pages.

To begin with, we compute the two attributes (distinct websites and distinct trackers) and plot them for each user in Figure 5.4: a point (x, y) on the red curve represents the average number y of trackers encountered for users who visit x different websites, and the green area defines the 95% confidence interval.

The total number of visited websites positively correlates with the trackers encountered (Pearson Correlation Coefficient: 0.98, $p < 0.001$). However, Figure 5.4 exhibits two classes of outliers, whose attributes fall out outside the confidence interval boundaries. Specifically, we define *Upper Outliers (UO)* those with an abnormal-higher ratio between encountered trackers and visited websites (blue dots in the picture, users that encounter a lot of trackers while not visiting many websites). On the contrary, we report in orange the *Lower Outliers (LO)* (users that browse a lot but encounter less trackers), for which this ratio is lower than the average and outside the confidence interval. The UO and LO sets contain respectively 6,726 and 5,552 users, which together account for 4.6% of the users in our dataset.

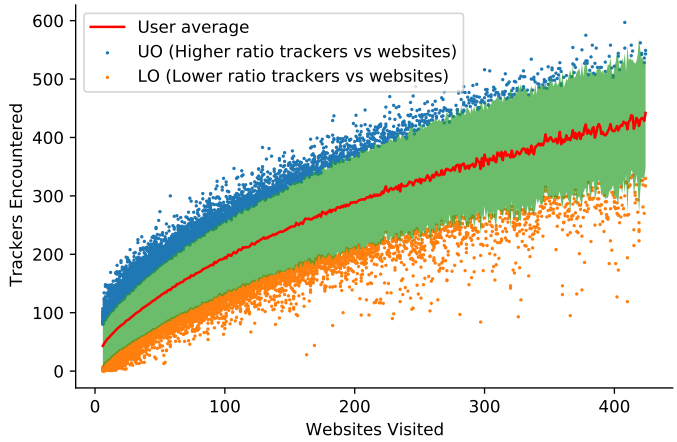


Figure 5.4: Correlation trend between the number of visited websites and encountered trackers

To investigate whether any significant difference exists in the websites visited by the two groups of outliers, we use two metrics: popularity and security risk score. We compute the popularity of each website in our telemetry by simply considering the number of times it appears in different users' browsing histories. This score is conceptually similar to the reputation returned by online rating services [42, 43, 44, 47], and it is strictly related to the data in our experiment.

Given a popularity x, we separately plot for each group the sum of visits that each distinct website with reputation x receives (Figure 5.5). We next compute the weighted average for UO and LO according to the following criterion:

$$Wavg = \frac{\sum_{reput=1}^{max\_reput} reput * visits(reput)}{sum(visits)}$$

The two averages, represented by the vertical lines in the figure, show that users that encounter fewer trackers (LO group) are indeed visiting less popular websites. Instead, users who browse fewer websites but encounter on average more trackers mainly visit popular web pages: this is the case, for instance, of very popular news websites, social media, and online marketplaces, which incorporate a large number of advertisers, and a myriad of analytics services. For those users within the green zone in Figure 5.4, the reputation score falls between the one of UO and LO (i.e., 3,997), confirming our hypothesis that reputable websites are more tracked.

To compute the security risk score we leverage the website risk score

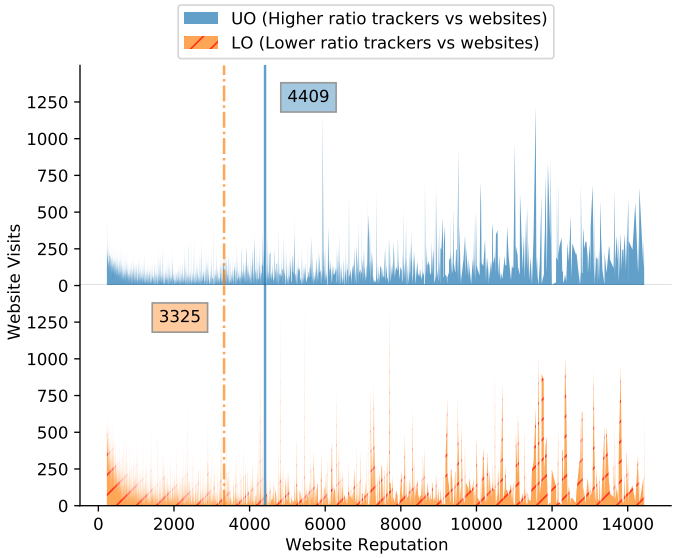


Figure 5.5: Website reputation distribution for *UO* and *LO*. The difference between the means in the two groups is significantly different ($Welch'st = 113.06, p < 0.001$).

provided by the AV vendor. Then, for each set of users, we split the websites they visited according to their risk value, and plot a histogram with the percentage of the total history they account for (Figure 5.6). The figure also includes the weighted average of both groups, computed by following the same procedure described for Figure 5.5.

The plot shows that users in the UO group mainly browse benign websites. In our dataset, not a single website visited by these users had a rating that classifies it as either suspicious or malicious (>= 6). On the other end of the spectrum, users in the LO group visit a larger percentage of dangerous sites. Similarly, the users in the green zone visit websites with low-risk scores however slightly higher than those UO users (2.6 risk score).

Overall, we found that websites that include no trackers are often less popular and characterized by a higher security risk. Table 5.5 reports the top and bottom website categories, sorted by the percentage of webpages in which we do not detect any trackers. The top categories show a considerably higher risk score (6.96 on average) than the bottom (3.90 on average) suggesting that the former often present suspicious or malicious content rather than the latter (confirmed also by the category names). A clear exception in top half of the table is represented by the *Business/Economy* category, which is both low-risk and low-tracking. This category represents websites

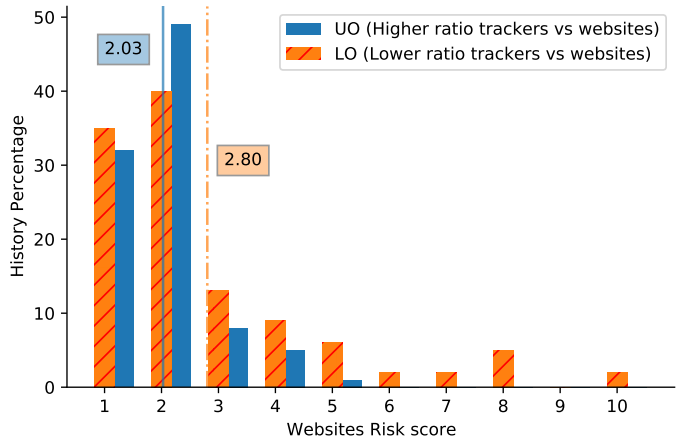


Figure 5.6: Website risk score distribution for UO and LO. The difference between the means in the two groups is significantly different ($Welch'st = 432.41, p < 0.001$).

devoted to businesses (including information and management) that are not linked to any selling activity. Taking this into account, a possible explanation is that websites in this group are directly related to customers or employees, so they do not include any type of tracking.

5.4.3 How Frequently do Users Encounter the Same Trackers?

So far we only looked at how often users encounter new trackers. But the key point of tracking is identifying the same user across different websites. So, if a user encounters a specific tracker only once a day, then deleting its cookie at the end of the browsing session could prevent the tracker to connect the different visited sites. It is clearly possible that some trackers perform some type of browser fingerprinting [125] in order to be able to track users around. In these cases, deleting cookies would not avoid tracking. However, as cookies are still the de-facto tracking method on the web [203], we wanted to investigate how effective the cookie cleaning option could be to improve users' privacy posture.

To better understand this aspect we looked at how frequently each tracker was encountered by each user. In Table 5.6, we report the percentage of users for which the top-5 most recurrent trackers appear with a frequency lower than 2 hours. Google, for instance, is encountered on average every 1.11 hours. This means that to fully prevent the largest company

Category	% with zero trackers	Avg risk
Malicious Outbound Data/Botnets	90.23	9.40
Business/Economy	70.45	3.93
Potentially Unwanted Software	56.75	7.00
Spam	56.21	7.00
Placeholders	55.81	6.00
Suspicious	53.41	7.58
Scam/Questionable/Illegal	49.24	7.35
Email	43.69	4.47
Malicious Sources/Malnets	42.89	9.99
Social Networking	12.92	4.09
E-Card/Invitations	12.69	3.44
Informational	12.39	3.93
Alcohol	12.11	4.03
Translation	12.02	3.29
Restaurants/Food	11.85	4.19
Charitable Organizations	11.78	3.95
News/Media	10.93	3.77

Table 5.5: Zero-Tracker website percentage and risk score for top and bottom 0-tracker categories

in our dataset from being involved in tracking practices, a user should delete the cookies after every single browsing hour, which is obviously not realistic. Figure 5.7 and Figure 5.8 respectively report the cumulative distributions for a time-based and site-based perspective. The plots show that 50% of the trackers are repeatedly encountered every 8 hours or 60 websites. In other words, if the cookies are cleaned up every 8 hours or after 60 website visits, only half of the trackers could be prevented from tracking. However, cookie cleaning is clearly is not an absolute solution for those privacy conscious users who do not want to be tracked by any means: this practice is not effective against big players that can know much more and are encountered much frequently (on average every 1.34 browsing hours)

5.5 The Knowledge of Trackers

In the previous section, we have seen that the average users encounter 84% of the trackers within just half a day of web browsing. While this is very concerning for the privacy of Internet users, the impact on their privacy

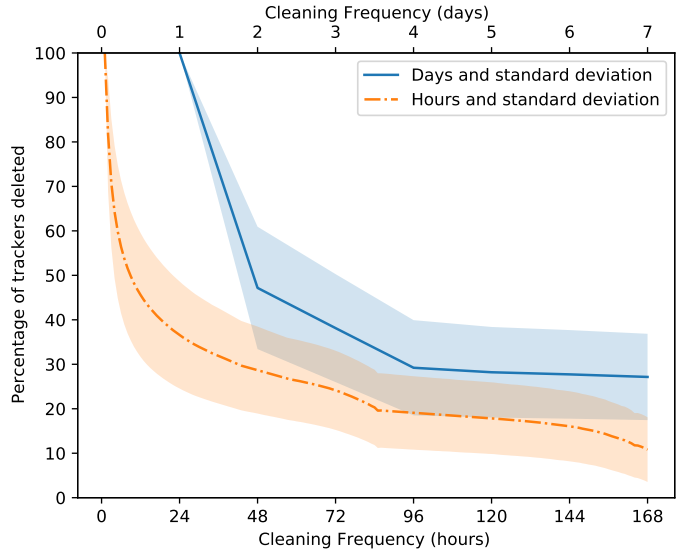


Figure 5.7: Percentage of trackers deleted according to the frequency (browsing hours and days) of cookie cleaning.

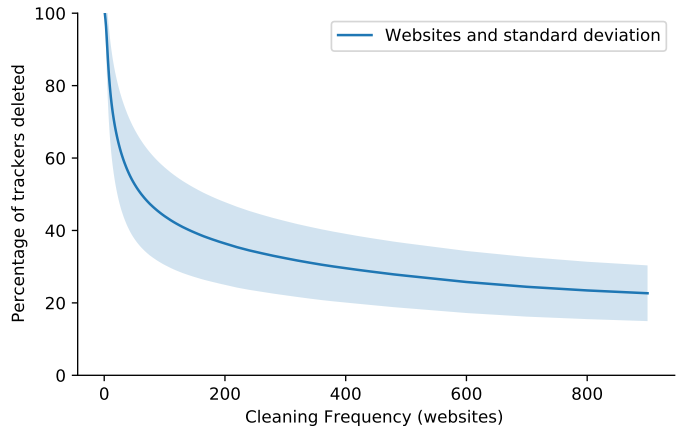


Figure 5.8: Percentage of trackers deleted according to the frequency (browsed websites) of cookie cleaning.

might not be as significant and worrying unless those trackers can compromise a significant fraction of the users' browsing history. In this section, we take a closer look to estimate how much information about users is known (or potentially known) by those trackers. We first assess to what extent main trackers on the visited websites know about the users' browsing histories, and then, how much additional coverage they could gain by sharing

Tracker	% Users	Avg frequency (hours)
Google	80.41	1.11
Microsoft	67.61	1.29
Twitter	67.18	1.41
Yahoo!	66.25	1.43
Rubiconproject	62.68	1.44

Table 5.6: Top-5 trackers according to the frequency (browsing hours) of appearance. % Users refers to users for which the tracker appears with a frequency < 2 browsing hours.

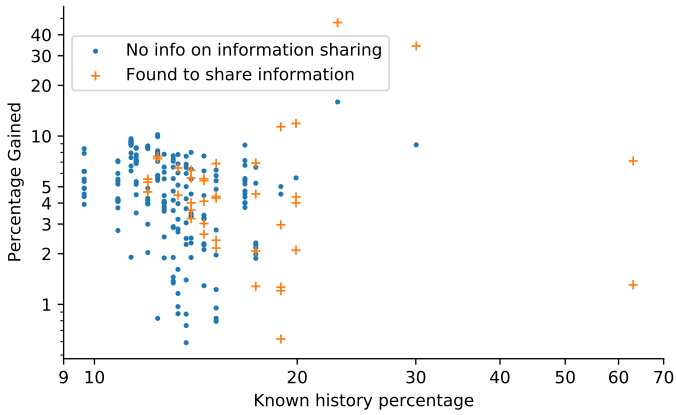
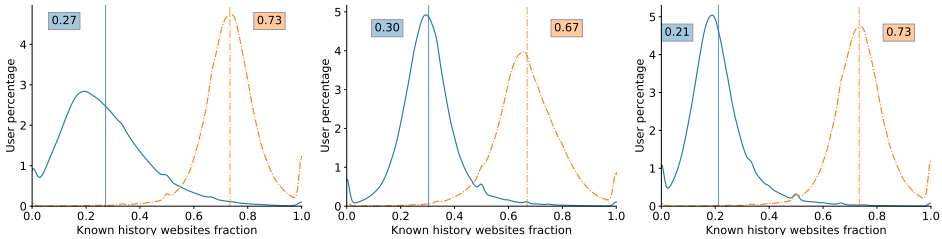


Figure 5.9: Possible browsing history gain through collaboration

information among one another. We also investigate the type of information that could be learned about the identity of users through regularly browsing particular types of websites. Finally, we conclude the section with an optimal tracking strategy analysis.

Tracker A (Receiver)	Tracker B (Sender)	Coverage Tracker A	Coverage Tracker A+B	Gain	% B overlapping
Linkedin	Amazon.com	11.33	20.62	9.29	22.57
Amazon.com	Linkedin	12.00	20.62	8.62	23.90
Microsoft	Google	22.97	70.19	47.22	25.13
Taboola	Linkedin	9.65	17.54	7.89	30.40
Linkedin	Openx	11.33	20.43	9.10	30.49
		...			
Rubiconproject	Casalemedia	15.16	15.95	0.79	94.21
Casalemedia	Openx	13.68	14.43	0.75	94.28
Casalemedia	Pubmatic	13.68	14.27	0.59	95.55
Google	Facebook	63.07	64.38	1.31	95.65
Appnexus	Rubiconproject	18.91	19.54	0.62	95.89
Microsoft	Google	22.97	70.19	47.22	25.13
Facebook	Google	30.05	64.38	34.33	45.56
Microsoft	Facebook	22.97	38.92	15.95	46.91
Adobe	Microsoft	19.92	31.79	11.87	48.30
Appnexus	Microsoft	18.91	30.27	11.35	50.56
		...			
Drawbridge	Linkedin	12.41	13.24	0.83	92.71
Rubiconproject	Casalemedia	15.16	15.95	0.79	94.21
Casalemedia	Openx	13.68	14.43	0.75	94.28
Appnexus	Rubiconproject	18.91	19.54	0.62	95.89
Casalemedia	Pubmatic	13.68	14.27	0.59	95.55

Table 5.7: Upper (Lower) part: top and bottom 5 relationships sorted by ascending overlapping (descending gain)



(a) Microsoft (b) Facebook (c) Appnexus

Figure 5.10: Known-history percentage distribution of the trackers that directly appear the most in users' history without (solid line) and with information sharing (dotted line). The percentage of users' history without any tracker is 20.07%.

5.5.1 How much do trackers know about you?

For each tracker we identified in our dataset, we computed the average fraction of browsing history known, the percentage of websites in which they are present, and also the fraction of users who encounter them. On average, each tracker tracks 3% of the users and knows 0.14% of their browsing history. However, the top trackers (such as Google, Facebook, and Microsoft) are quite far from the average. In fact, they are able to track nearly all users, as can be seen from Table 5.4, and they know on average 47% of each user's browsing history. Google alone, which is the biggest player in the tracking ecosystem, covers 64% of the average users' history logs. The percentage increases to 80% for 9.73% of the users, and reaches a stunning 100% for 2% of them.

5.5.2 How much can trackers know about you through collaboration?

Collaboration among trackers is not a new phenomenon [107, 184, 203]. It allows them to merge the user data with another tracker, reconstructing users browsing history, and bypassing the same-origin policy [234]. In order to do it, tracking companies can use multiple methods, with cookie sharing/synchronization being the most common one. For example, a tracker can include its cookie in the request of another third party, facilitating an information-sharing channel even if not directly present in that specific website. Our goal here is to estimate the concrete impact of such collaborations on users' browsing history, which was not explored before by other studies.

In the previous section, we have seen that with the exception of Google, none of the other trackers knows more than 30 percent of the average user's browsing history. Clearly, if Google shared its knowledge with any other tracker, they could also achieve similar coverage. However, this is not a very realistic scenario from a strategic point of view. On the other hand, collaboration among smaller players in the ecosystem might make more sense. Therefore, to understand how much information trackers could gain through collaboration, we calculated the browsing history gain for all possible pairs of companies among the top 20 trackers in our dataset and plotted the percentage of gain versus known history percentage in Figure 5.9. If the two companies were already known to collaborate according to previous measurements [203], we colored them in orange. If you remove the top three players, in general most trackers over the top 20 can know between 10 and 20% of the browsing history of the users. Through collaboration, they can increase their knowledge of an additional 5 to 10% (mean gain is 5.3%) in the best case scenario unless they can collaborate with Google.

In Table 5.7 we also provide concrete examples for some of the interesting collaboration options. Similarly, those collaborations that are known to exist by other means are marked in gray. The most obvious gain examples come from the collaboration among the biggest players. Because in most of the websites in which we observe Facebook, we also encounter Google (95.65%), Google gains not much (1.31%) from getting information from Facebook. However, Facebook could immensely increase its knowledge, up to 64.38%, from a potential collaboration with Google. Another interesting observation is that Microsoft and Google do not target similar sets of websites, therefore a possible collaboration would have a much larger impact. On the contrary, the overlap among the top 20 trackers ranges between 23 and 96% (mean overlap of 64%). This clearly indicates that many of them are tracking users in a very similar set of websites.

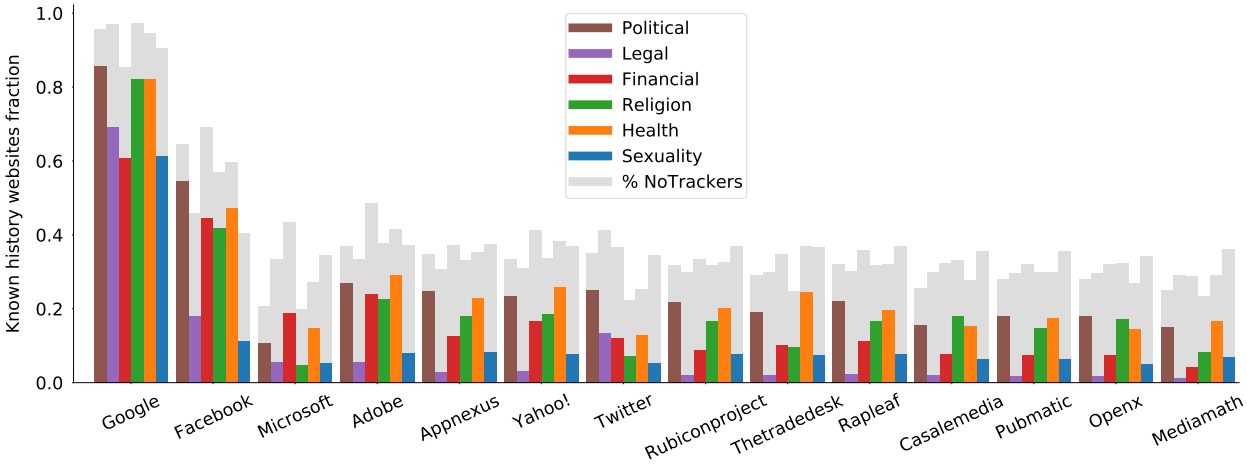


Figure 5.11: Known history percentages of the 6 sensitive categories by the top trackers.

112

Now let's look at the worst-case scenario, in which we assume that all trackers that were identified to be sharing information according to recent studies collaborate to increase their knowledge as much as possible. In Figure 5.10, we provide three examples of how much information can be potentially gained in such a scenario. It is interesting to see that Microsoft could potentially already know up to 73% of the users browsing history (instead of the 27% it has if it was completely disconnected from other players). Another similar spike is observed in Appnexus (from 21 to 73%). While the gain for Microsoft is mostly due to its relationship with Google, Appnexus receives information from a variety of other trackers including Microsoft, Adobe, Yahoo!, and more. Again, these numbers assume a complete share of all tracking information among the companies, so in reality the numbers are likely somewhere in between the two scenarios (no collaboration and full collaboration).

Continent	Websites Percentage						
	All	Sexuality	Health	Religion	Financial	Legal	Political
Africa	**6.03**	1.74	0.84	0.49	1.45	0.70	0.11
Asia	**5.84**	0.96	0.89	0.42	2.07	1.27	0.09
Europe	**5.45**	1.91	0.87	0.50	1.80	0.86	0.12
North America	**5.00**	1.27	1.20	0.75	3.12	0.77	0.18
Oceania	**4.85**	1.72	1.00	0.52	2.36	0.88	0.11
South America	**5.38**	1.04	1.26	0.42	1.95	1.93	0.14

Table 5.8: Sensitive website prevalence in users' history.

Continent	Average trackers						
	All	Sexuality	Health	Religion	Financial	Legal	Political
Africa	**7.01**	4.50	10.34	7.50	6.64	2.53	7.53
Asia	**6.96**	5.96	14.44	7.31	5.54	1.79	7.39
Europe	**7.12**	5.56	7.52	5.87	4.81	2.82	6.10
North America	**8.55**	6.79	12.97	8.68	8.23	3.70	14.69
Oceania	**7.69**	6.59	12.47	8.84	6.33	2.60	8.52
South America	**7.20**	6.02	10.06	9.44	4.87	2.23	12.85

Table 5.9: Sensitive website average number of trackers in users' history.

5.5.3 What type of sensitive information can be obtained about you?

Visiting or regularly browsing particular types of websites could reveal sensitive information about users. In this part of our analysis, we focus on websites that could fall into sensitive categories and check which trackers are present on those sites and could therefore gain access to private users' information. In particular, we identified six categories that are widely considered to be sensitive (see Section 5.3) and we computed the portion known by top trackers. Figure 5.11 reports the averages over the whole dataset. In gray, we represent the percentage of history in which we do not detect any trackers.

At a glance, we observe that the tracking activity is not uniform among the six sensitive categories: while the percentage of untracked history is very low in the *Health*, *Religion*, and *Political* categories (respectively 12, 15 and 10%), the fraction doubles for the *Sexuality*, *Financial*, and *Legal* classes (30, 24, 28%).

A first interesting case is the *Political* category: although it presents the lowest number of websites and users who browse it (see Table 5.3), it turns out to be the category the top trackers know the most about. In fact, our crawler detects multiple trackers on average on each of these pages, with top trackers uniformly present on most of them.

The *Legal* category results in the opposite case: top organizations on average know less than 5% of sites in this category, with the exclusion of Google (69.10%): we measure an average presence of 2.64 trackers for websites in this group.

More concretely, if looking at the per-tracker details in the graph, the figure presents similar trends and known history percentages, except for Google and Facebook. Since in general Google knows over 60% of the users' history, it is not very surprising that it also covers a good fraction of the browsing history related to the sensitive categories. However, the Facebook case is utterly interesting. On the general data, it only knows up to 30% of the users browsing history, which is in line with other top players. Despite that, it covers almost 60% of the browsing on the *Political* sector, and around 50% of the *Health* category. This seems to indicate, for example, that Facebook puts a particular effort in tracking specific website classes. On the other side of the spectrum lies Microsoft, which on the general data has a much larger coverage (over 20%) than its presence on sensitive website categories.

We also investigate whether the prevalence and tracking of sensitive websites are uniform across continents. For each of them, in Tables 5.8 and

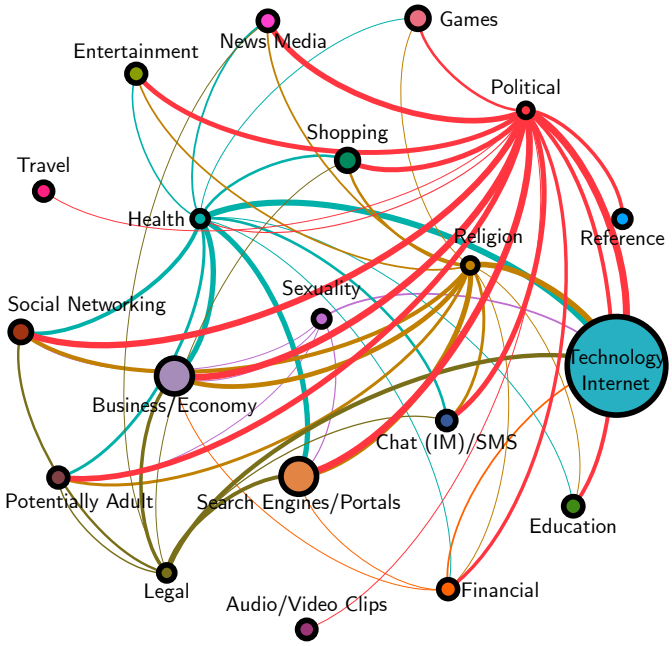


Figure 5.12: Relationships among sensitive and top categories

5.9 we respectively report the average percentage of browsed websites per sensitive category together with the average number of trackers encountered. Results show no substantial differences across continents and confirm that sensitive information about *Health*, *Religion*, and *Political* is more subject to tracking practices, although their prevalence is very small in users' histories. The only comforting difference is observed in Europe. Very likely thanks to the GDPR, the average number of trackers found in websites is lower than others.

As a next step, we investigate how much more information can be identified about a user's identity by connecting the pieces. For example, if a tracker knows that a user follows a particular political party or religious belief, can we estimate the likelihood of them knowing also about the user's travel plans, health interests, etc? To this end, we build a linkage graph among the sensitive categories and other website categories. We consider each user at the time, and isolate the history containing webpages of the sensitive category (SI) under analysis from the remaining part (RI) — note that the group also contains other sensitive categories besides the one we investigated so far. For each webpage in SI, we extract the list of trackers and check their presence in the remaining webpages of RI. Given the list of

matched websites, we detect their categories and increase a counter for each of them. Once n webpages in *SI* have been analyzed, we divide each of the category counters by n, obtaining a ratio. For a single user, a ratio close to 1 between a sensitive category a and another one b means that, each time we encounter a website in a, the trackers also know that the user visited b.

We plot the resulting linkage graph in Figure 5.12. Node sizes represent the percentage of history that falls in the category: the biggest category is *Technology/Internet* (39% users' history), the smallest is *Political*, accounting for 0.13%. Each edge between two nodes expresses the average category correlation for all the users in our dataset. To increase the readability, the graph only includes the sensitive and the most prominent ten categories that have at least one ingoing edge with a weight greater than 70%. We observe that the strongest correlation percentage (95.55%) holds between *Political* and *Technology/Internet*, while the weakest (70.01%) between *Legal* and *Chat (IM)/SMS*.

We also see that some categories are much less connected with the others. For instance, *Sexuality* and *Financial* have very few connections with other categories, and those connections are very small. On the other hand, *Political* has many strong connections with many other categories found in the dataset. In the middle, we find cases like *Health*, *Religion* and *Legal*, that despite having more connections than the first two, only have a couple of strong connections with others. We also verify how the linkage graph varies according to users' geographical location, and find that relationships are stable across continents except from Asia, in which we see *Health* has stronger connections than *Political*.

Another interesting point is that sensitive categories do not seem to have many connections among them. However, we have to note that, not having a direct connection in the graph does not necessarily indicate that trackers could not connect them through their relations to other common categories. For example, both *Political* and *Health* are connected to *Potentially Adult*, which could be used as a hub.

5.5.4 What is the optimal tracking strategy?

Earlier in this section, we have made estimations on how much browsing history knowledge could be obtained through collaboration among trackers, concluding that unless collaboration happens with Google, it is hard to gain a significant fraction of the browsing histories. An alternative option for the trackers to achieve the same goal is to plant themselves on *key websites*. For an optimal tracking strategy, the trackers need to build a list of *popular websites* such that the minimum number of them is required in order to cover

Tracker	% key websites
Google	66.04
Facebook	35.50
Adobe	21.54
Appnexus	19.02
Yahoo!	18.44
Microsoft	17.04
Rapleaf	16.00
Thetradedesk	15.56
Drawbridge	15.50
Rubiconproject	14.90
Twitter	13.92
Casalemedia	13.22
Openx	12.30
Pubmatic	11.80
Amazon.com	11.46
Linkedin	11.32
Mediamath	10.50
Quantcast	8.92
Akamaitech	6.98
Taboola	6.86

Category	% key websites
Technology/Internet	22.96
Business Economy	12.94
Shopping	6.60
News Media	5.82
Travel	5.04
Entertainment	3.90
Games	3.20
Suspicious	3.04
Financial Services	3.02
Education	2.70
Search Engines	2.36
Reference	2.28
Pornography	2.20
Mixed Content	1.74
Social Networking	1.38
Placeholders	1.38
Health	1.10
Restaurants/Food	1.08
Sports	0.98
Government	0.94

Table 5.10: Top-10 prevalence in the 5K *key-websites*: trackers (left) and categories (right)

a certain percentage of the whole users' history. To assess the effectiveness of this option, we created a sorted list of the 5K most reputable websites, according to the definition provided in Section 5.4.

In Figure 5.13, we plot how the known history percentage grows in relation to how many key websites the trackers need to work with. We also plot the existing presence of the top three trackers on those top 5K sites. The blue curve shows that, by cherry-picking only 200 websites, a company could observe 50% of the users browsing history. This value increases to 65% and 78% when extending the set of key webpages to the first 1K and 5K respectively (over a total of 2.33M in our dataset), indicating that being able to add a tracker to the top sites brings much more additional information than collaborating with other trackers.

As seen in Table 5.10, top players already show a significant presence on the *key websites*. When we look closer at the tracking strategy of Google,

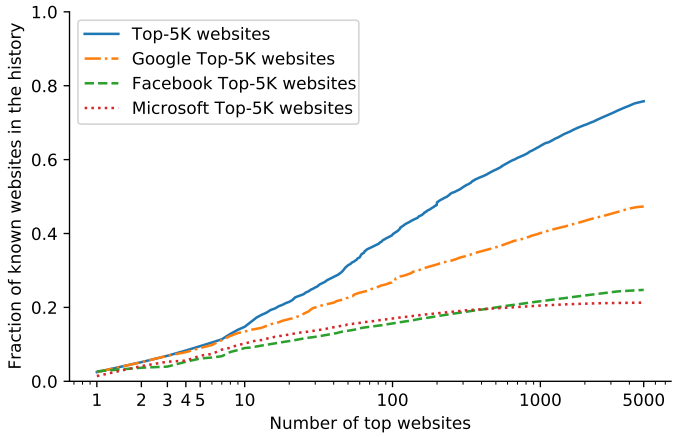


Figure 5.13: Optimal tracking strategy on *key-websites* vs top-3-tracker strategy on their top-5K websites

Facebook, and Microsoft (Figure 5.13), we identify interesting differences. First, although present in 66% of the key websites, the coverage of Google diverges from the optimal curve after considering only 10 websites (0.2% of the 5K): a sign that its presence is more prominent in the less reputable website of the group. We also noticed that Microsoft had a better coverage strategy than Facebook: although the two organizations show a similar trend in Figure 5.13, the first is only present in 17% of the key websites —half the percentage of the second—, suggesting that it appears in more reputable websites. In Table 5.10, we report the breakdown of categories together with the fraction of key websites for each of them: *Technology/Internet* and *Business/Economy* group a sheer number of webpages, being the two most popular categories overall.

5.6 Comparison and Discussion

While web tracking is widely considered a common phenomenon, the results that we obtained by studying web tracking from the users' perspective show that it is considerably more widespread than previously thought. Previous studies [203, 204, 131, 104] attempted to quantify its scale by conducting large-scale measurements on open datasets, such as Alexa 1M [42] or Tranco [189]. However, while one would expect that crawling the most popular websites should provide an upper bound approximation of exposure, we found this to be wrong. For example, Google was found to track user activities on 46% [204, 131] of the top domains, but our study reveals that

the actual knowledge of the users' histories reaches 73%. In the same way, Facebook prevalence was estimated around 18% [203], but our measurement shows it to be almost twice that value.

One of the main results of our study is to show that if the impact of web-tracking is measured only by considering top websites, the fraction of known browsing history would be largely under-estimated. Moreover, the relationship among the two is not always the same. As an example, Microsoft and Pubmatic appear both in 4% of the analyzed websites, but the former covers on average almost twice the users' browsing history compared to the latter (Table 5.4). The use of telemetry makes it also possible to quantify the exact impact of collaborations among organizations on end users. Previous studies discovered that 66% of the top-100 trackers share cookies [104] and that users with a larger browsing profile are tracked by more identifier sharing domains [107]. Thanks to our analysis we now know that this practice could increase the knowledge that trackers have of the users' activity by almost 50%.

Another advantage of our method with respect to previous works is that it also allows us to shed light on the timing and frequency with which users are tracked, thus unveiling insights on research areas that have never been explored so far and whose investigation is impossible by crawling top websites. For instance, we show that users encounter almost all the tracking organizations in just half a day of activity. Even more worryingly, we show that the frequency with which some of the top trackers are encountered makes it infeasible to prevent their monitoring by simply deleting the cookie history.

The knowledge that tracker organizations have of users' browsing interests, habits, recurrence, location and hourly activity enables the creation of powerful profiles that get more and more refined and available to many players willing to purchase them. As a result, users risk to lose control of their private information and face several serious consequences. For example, a known use of tracking is the personalization of search results based on users' interests and the creation of the so-called Filter Bubble [251], a personalized search where an algorithm guesses what results the user would like to see based on previously collected information. Web tracking is also massively used to serve targeted advertising, facilitate marketing, and increase sales profit by influencing customer purchasing behaviors. In this respect, tracking can be used to modify product prices according to the geographical location and the financial situation of potential customers [59, 245]. Many companies also leverage this information to assess users' financial credibility [231, 232] and establish insurance coverage [123].

5.6.1 What can users do to protect themselves?

As web-tracking closely concerns users and their activities, several tools and strategies exist to defend against this practice, being the most important: cookie clearing, list/rules-based blocking, and network-level masking.

Cookie clearing – In order to significantly reduce cookie-based tracking, users could delete the cookies stored in their browsers. However, this approach is complex to strictly follow in the long term and it would require a lot of effort: users must delete cookies with high frequency (i.e., less than one hour according to our findings in Section 5.4) and cherry-pick the ones to delete in each case.

List/rules-based blocking – The most common solution is the use of browser extensions or privacy-centered browsers that maintain an up-to-date list of tracking domains or rules and block all the connections towards them, thus preventing data collection about browsing sessions. Some of them rely on large-scale crawls to analyze how the ecosystem evolves [96], and some others principally have a crowdsourcing model [99]. These kind of solutions are easy to setup (i.e., install and forget) and avoid the need for manually deleting cookies on a regular basis. However, blocking resources can sometimes generate unexpected functionality problems in the page. In order to avoid them, solutions generally offer a page-specific disable option, but as indicated in Section 5.5, a large percentage of the browsing history in sensible categories is being tracked by multiple companies, so users should be extremely careful when disabling protection tools in them.

Although these solutions exist, and are practical and effective, extension or application-based blocking is not yet widely adopted: privacy-centered browsers only represent 7.74% of the market, and only 8.5% of the users adopt tracker-blocking tools [133, 247]. Therefore raising awareness about the extent of web-tracking is crucial to increase these percentages and we believe that the quantitative insights presented in this work could be immensely helpful to serve this purpose.

Network-level masking – Section 5.3 shows that the knowledge of tracker organizations spans a high percentage of the users' browsing history, reaching up to 63% in the case of Google. Therefore, some protections can be implemented at the network level to protect a larger portion of users and devices. Protective measures can be installed in home routers [187] or adopted as a privacy layer in companies. Despite being flexible and allowing protection of multiple devices at the same time, those tools are more difficult to set up, and require users to regularly maintain them, discouraging regular web users in adopting them.

There are also some solutions to mask the user's real IP address from

the remote site, thus preventing IP-based tracking. This goal is achieved by using anonymous proxy servers (which act as intermediary and offer anonymization services by removing sensitive information), virtual private networks (VPNs) (whose nodes result to be hosts of a single network, regardless of their physical locations), or Tor [21] (whose browser prevents tracking by routing the traffic through a chain of relays which protects the real user's IP address). Even if this is only one part of online tracking, some studies have already proven that a large percentage of users retain their same IP addresses for more than a month [164], allowing companies to use it as an identifier. When adopting this type of solutions, users should additionally use a list/rules-based blocking tool on top, to also avoid general types of tracking.

5.7 Conclusions

Despite the existence of these solutions and the users' awareness of online tracking practices, the adoption of such countermeasures is still limited. A possible reason is that users might feel they are not directly impacted. The goal of this study is to provide a more accurate measure of how web-tracking directly impacts them, and with evidence about how their online privacy is affected. We hope our findings can enable better decision making and foster a larger adoption of existing privacy-preserving services.

Chapter 6

Future Work and Conclusion

6.1 Future work

The challenges that affect cyber-risk quantification are far from being solved. In addition, the dynamic and ever-changing nature that characterizes the cyber-threat landscape makes it difficult to find a definitive solution and requires a constant refinement of methodologies and techniques that adapt to this evolution.

This thesis shows that data plays an important role in providing both researchers and practitioners with actionable information for cyber-risk quantification. In the first contribution, we enumerated several research directions that build on top of data but we were able to investigate only part of them. For instance, future work may look at the effectiveness of externally-collected indicators that reflect the security posture of a given entity when internal telemetry is not available. Another key obstacle when conducting cyber-risk measurements consists of merging all the different assessments that look at the problem from a particular perspective and that need to be aggregated to provide a comprehensive view to the analysts: risk aggregation is still a major challenge to address in the cyber domain. The interconnection of cyber risks and the catastrophic consequences that arise from major events overly complicate cyber-risk quantification. Future work may want to look at cascading effects of cyber incidents and how risks propagate to all the parties that depend on each other because they share third-party services, software or infrastructures.

The study on systemic and systematic indicators presented in this manuscript showed that host-extracted features carry useful information to explain why certain classes of machines are more likely to encounter malware. Although in our work we comment on the predictive power that those features have on the different malware families, future studies may want to narrow the analysis down to specific cases. For example, it is important to understand whether some indicators may explain the likelihood of suffering from particular cyber incidents (e.g., data breaches) or being targeted by more sophisticated malware classes (e.g., banking trojan or spyware). With cyber-insurance as an ultimate goal, it is also fundamental to understand whether any correlation exists between objective features and falling for incidents that require an insurance claim. In this respect, it is important to distinguish between those events that can be covered by an insurance policy (e.g., ransomware attack, DoS) and those that can not (e.g., Adware or PUA infections).

Finally, the last contribution highlighted the importance that different data sources and a holistic perspective have when evaluating cyber risks.

Although our analysis spans several countries and leverages the telemetry of 250K users, we only focus on those who browse by using regular desktop machines running Windows. Future contributions that aim at quantifying the extent of web tracking may focus on different OSs that were not considered in this analysis and well as look at the mobile ecosystem.

6.2 Conclusion

This thesis uncovers existing challenges of cyber risk quantification and provides contributions to measuring particular classes of cyber risks, such as malware-infection and privacy risks.

In Chapter 3, we detailed the complex challenges that researchers and practitioners have to face when quantifying the cyber risks of a given entity and contextualized them in the cyber-insurance domain where those are fundamental for an exhaustive definition of premiums. We also listed several research directions that system-security experts can explore to refine and improve the process of risk measurement.

In Chapter 4, we investigated the systemic and systematic nature of cyber risks by differentiating between consumer and enterprise environments. We show how the different security postures and defensive measures adopted by the two parties contribute to lower the risk of encountering particular malware classes, such as Adware and PUA. Our measurements also reveal that some host-extracted indicators —related to activity, installed-software reputation, geographic location and enterprise size —carry useful information when explaining the systematic nature of cyber risks.

In our last Chapter, we show how important a holistic approach together with reliable data sources are when assessing the extent of particular risk classes. We indeed show that privacy risks due to web tracking are underestimated if carrying the measurements by only crawling top websites as this excludes the users' perspective.

We really hope that the insights and contributions of this thesis can serve as a starting point for future work in the field, and assist academics and practitioners to better deal with cyber-risk quantification.

References

[1] Anatomy of a cyber risks stress test. https://www.zurich.com/en/knowledge/articles/2016/11/anatomy-of-a-cyber-risks-stress-test. Accessed: 2019-07-31.

[2] Cyber risk stress testing. https://www.firstbase.co.uk/cyber-resilience/cyber-risks-stress-testing/.

[3] Cyberattack insurance a challenge for business. https://www.nytimes.com/2014/06/09/business/cyberattack-insurance-a-challenge-for-business.html.

[4] Cybersecurity incident simulation exercises. https://www.ey.com/Publication/vwLUAssets/EY_-_Cybersecurity_Incident_Simulation_Exercises/$FILE/EY-cybersecurity-incident-simulation-exercises-scored.pdf. Accessed: 2019-07-31.

[5] Cybersecurity stress-testing: Don't stress about your company's safety. https://axiomcyber.com/cybersecurity/cybersecurity-stress-testing-dont-stress-about-your-companys-safety/.

[6] Eiopa insurance stress tests to assess cyber risk response. https://www.out-law.com/en/articles/2018/may/eiopa-insurance-stress-tests-cyber-risk/.

[7] Financial catastrophe research and stress test scenarios. https://www.jbs.cam.ac.uk/fileadmin/user_upload/research/centres/risk/downloads/2016risksummit-riskculture-slides-skelton.pdf. Accessed: 2019-07-31.

[8] Five pitfalls of cybersecurity insurance: Lessons from the united states. https://www.lexology.com/library/detail.aspx?g=0afce621-4d25-448e-bb38-1aaeof6c50c7. Accessed: 2019-07-31.

[9] Hackmageddon: Information security timelines and statistics. https://www.hackmageddon.com. Accessed: 2019-07-31.

[10] Heads-up: Cyber insurance does not pay out for human error. https://community.spiceworks.com/topic/1999873-heads-up-cyber-insurance-does-not-pay-out-for-human-error. Accessed: 2019-07-31.

[11] How to stress test your cyber risk management. https://www.marsh.com/uk/insights/risk-in-context/how-to-stress-test-your-cyber-risk-management.html. Accessed: 2019-07-31.

[12] Hsb introduces first cyber insurance for consumers. https://www.munichre.com/HSB/first-personal-cyber-insurance/index.html. Accessed: 2019-07-31.

[13] Inside the cunning, unprecedented hack of ukraine's power grid. https://www.wired.com/2016/03/inside-cunning-unprecedented-hack-ukraines-power-grid/.

[14] Insurers told to conduct stress tests for cyber attacks. https://www.ft.com/content/92ec137a-6185-11e7-8814-0ac7eb84e5f1. Accessed: 2019-07-31.

[15] Isaca - cobit. http://www.isaca.org/COBIT/Pages/default.aspx. Accessed: 2019-07-31.

[16] The iso/iec 27000 family of information security standards. https://www.itgovernance.co.uk/iso27000-family#4. Accessed: 2019-07-31.

[17] Microsoft solutions for security and compliance and microsoft security center of excellence - the security risk management guide. https://technet.microsoft.com/en-us/library/cc163143.aspx. Accessed: 2019-07-31.

[18] An overview of the main changes under gpdr and how they differ from the previous directive. https://www.eugdpr.org/key-changes.html. Accessed: 2019-07-31.

[19] Owasp wasc web hacking incidents database project. https://www.owasp.org/index.php/OWASP_WASC_Web_Hacking_Incidents_Database_Project. Accessed: 2019-07-31.

[20] Security breach notification laws. http://www.ncsl.org/research/telecommunications-and-information-technology/security-breach-notification-laws.aspx. Accessed: 2019-07-31.

[21] TOR. https://www.torproject.org/download/.

[22] The veris community database (vcdb). http://veriscommunity.net/vcdb.html. Accessed: 2019-07-31.

[23] What is stuxnet, who created it and how does it work? https://www.csoonline.com/article/3218104/malware/what-is-stuxnet-who-created-it-and-how-does-it-work.html.

[24] ISO 3166-1. https://en.wikipedia.org/wiki/ISO_3166-1, 1997. Accessed: 2020-05-04.

[25] Guide for conducting risk assessments. Technical report, NIST - National Institute of Standards and Technology, 2012.

[26] Managing Cyber Security as a Business Risk: Cyber Insurance in the Digital Age. Technical report, Ponemon institute, Aug. 2013. Available at: https://www.ponemon.org/local/upload/file/Cyber%20Insurance%20white%20paper%20FINAL%207.pdf.

[27] Insurability of cyber risk. Technical report, The Geneva Associations, Aug. 2014. Available at: https://www.genevaassociation.org/sites/default/files/research-topics-document-type/pdf_public/ga2014-if14-biener_elingwirfs.pdf.

[28] Cyber Risk: Too Big to Insure? Technical report, Institute of Insurance Economics and Swiss Re, June 2016. Available at: https://www.ivw.unisg.ch/~/media/internet/content/dateien/instituteundcenters/ivw/studien/cyberrisk2016.pdf.

[29] 2017 Cyber Risk Landscape. Technical report, Risk Management Solutions, Inc and Cambridge Centre for Risk Studies, 2017. Available at: https://www.jbs.cam.ac.uk/fileadmin/user_upload/research/centres/risk/downloads/crs-rms-cyber-risk-landscape-2017.pdf.

[30] 2017 Data Breach Investigations Report. Technical report, Verizon Enterprise, Apr. 2017. Available at:https://www.ictsecuritymagazine.com/wp-content/uploads/2017-Data-Breach-Investigations-Report.pdf.

[31] Global cyber market overview. Technical report, Aon Inpoint, June 2017. Available at: https://www.aon.com/inpoint/bin/pdfs/white-papers/Cyber.pdf.

[32] Is cyber risk systemic? https://www.aig.com/content/dam/aig/america-canada/us/documents/business/cyber/aig-cyber-risk-systemic-final.pdf, 2017. Accessed: 2020-05-04.

[33] Cyber insurance and systemic market risk. https://www.eastwest.ngo/sites/default/files/ideas-files/cyber-insurance-and-systemic-market-risk.pdf, 2018. Accessed: 2020-05-04.

[34] Cyber insurance claims: Ransomware disrupts business. Technical report, American International Group AIG, Mar. 2018. Available at:https://www.aig.co.uk/content/dam/aig/emea/united-kingdom/documents/Insights/cyber-claims-report-may-18.pdf.

[35] insurers pay large claims for high profile cyber attacks. Technical report, jardine lloyd thompson group plc, June 2018. Available at: https://www.jlt.com/insurance-risk/cyber-insurance/insights/insurers-pay-large-claims-for-high-profile-cyber-attacks.

[36] Kaspersky security bulletin 2018. threat predictions for 2019. https://bit.ly/2Wq5eIw, 2018. Accessed: 2020-05-04.

[37] Quantifying systemic cyber risk. http://web.stanford.edu/~csimoiu/doc/Global_CRQ_Network_Report.pdf, 2018. Accessed: 2020-05-04.

[38] Views from the C-suite Survey 2018. Technical report, FICO, Mar. 2018. Available at:https://www.fico.com/en/resource-download-file/6341.

[39] Cisco annual cybersecurity report. https://www.cisco.com/c/dam/m/hu_hu/campaigns/security-hub/pdf/acr-2018.pdf, 2019. Accessed: 2020-05-04.

[40] Internet security threat report. https://www-west.symantec.com/content/dam/symantec/docs/reports/istr-24-2019-en.pdf, 2019. Accessed: 2020-05-04.

[41] Microsoft security intelligence report. https://www.microsoft.com/security/blog/2019/02/28/microsoft-security-intelligence-report-volume-24-is-now-available, 2019.

[42] Alexa, "seo and competitive analysis software.". https://www.alexa.com, 2021.

[43] Cisco umbrella, "umbrella popularity list.". https://umbrella-static.s3-us-west-1.amazonaws.com/index.html, 2021.

[44] Majestic, "the majestic million.". https://majestic.com/reports/majestic-million, 2021.

[45] Mullvad. https://mullvad.net, 2021.

[46] Nortonlifelock, 2021. https://www.nortonlifelock.com.

[47] Quantcast, "audience insights that help you tell better stories.". https://www.quantcast.com/top-sites, 2021.

[48] L. A. Gordon, M. Loeb, and T. Sohail. A framework for using insurance for cyber-risk management. 46:81–85, 03 2003.

[49] G. Acar, M. Juarez, N. Nikiforakis, C. Diaz, S. Gürses, F. Piessens, and B. Preneel. FPDetective: dusting the web for fingerprinters. In *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security (CCS)*, 2013.

[50] C. J. Alberts and A. Dorofee. *Managing information security risks: the OCTAVE approach*. Addison-Wesley Longman Publishing Co., Inc., 2002.

[51] O. H. Alhazmi, Y. K. Malaiya, and I. Ray. Measuring, analyzing and predicting security vulnerabilities in software systems. *Computers & Security*, 26(3):219–228, 2007.

[52] M. Amutio, J. Candau, and J. Mañas. Magerit-version 3, methodology for information systems risk analysis and management, book i-the method. *Ministerio de administraciones públicas*, 2014.

[53] Artemis. Market unprepared for 'silent' cyber loss aggregation: Welsh, sciemus. http://www.artemis.bm/blog/2016/12/01/market-unprepared-for-silent-cyber-loss-aggregation-welsh-sciemus/.

[54] A. B. Association and FSSCC. Cyber insurance buying guide. https://www.aba.com/Tools/Function/Documents/2016Cyber-Insurance-Buying-Guide_FINAL.pdf. Accessed: 2019-07-31.

[55] B. Bailey. Fake accidents cost insurance industry billions each year, experts say. https://newsok.com/article/5518524/fake-accidents-cost-insurance-industry-billions-each-year-experts-say.

[56] L. Bailey. Mitigating moral hazard in cyber-risk insurance. *JL & Cyber Warfare*, 3:1, 2014.

[57] D. Balzarotti, M. Monga, and S. Sicari. Assessing the risk of using vulnerable components. In D. Gollmann, F. Massacci, and A. Yautsiukhin, editors, *Quality of Protection: Security Measurements and Metrics (QoP)*, Advances in Information Security, pages 65–78. Springer, 2006.

[58] T. Bandyopadhyay, V. S. Mookerjee, and R. C. Rao. Why it managers don't go for cyber-insurance products. *Communications of the ACM*, 52(11):68–73, 2009.

[59] P. Belleflamme and W. Vergote. Monopoly price discrimination and privacy: The hidden cost of hiding. *Economics Letters*, 149:141–144, 2016.

[60] C. Biener, M. Eling, and J. H. Wirfs. Insurability of cyber risk: An empirical analysis. *The Geneva Papers on Risk and Insurance-Issues and Practice*, 40(1):131–158, 2015.

[61] L. Bilge, Y. Han, and M. Dell'Amico. Riskteller: Predicting the risk of cyber incidents. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, pages 1299–1311. ACM, 2017.

[62] S. Bird, I. Segall, and M. Lopatka. Replication: Why we still can't browse in peace: On the uniqueness and reidentifiability of web browsing histories. In *Symposium on Usable Privacy and Security (SOUPS)*, 2020.

[63] K. P. Birman and F. B. Schneider. The monoculture risk put into context. *IEEE Security & Privacy*, 7(1):14–17, 2009.

[64] R. Böhme and G. Kataria. On the limits of cyber-insurance. In *International Conference on Trust, Privacy and Security in Digital Business*, pages 31–40. Springer, 2006.

[65] R. Böhme, G. Schwartz, et al. Modeling cyber-insurance: Towards a unifying framework. In *WEIS*, 2010.

[66] J. Bolot and M. Lelarge. Cyber insurance as an incentive for internet security. In *Managing information risk and the economics of security*, pages 269–290. Springer, 2009.

[67] J.-C. Bolot and M. Lelarge. A new perspective on internet security using insurance. In *INFOCOM 2008. The 27th Conference on Computer Communications. IEEE*, pages 1948–1956. IEEE, 2008.

[68] A. M. Bossler and T. J. Holt. On-line activities, guardianship, and malware infection: An examination of routine activities theory. *International Journal of Cyber Criminology*, 3(1), 2009.

[69] M. Bozorgi, L. K. Saul, S. Savage, and G. M. Voelker. Beyond heuristics: learning to classify vulnerabilities and predict exploits. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 105–114. ACM, 2010.

[70] H. K. Browne, W. A. Arbaugh, J. McHugh, and W. L. Fithen. A trend analysis of exploitations. In *Security and Privacy, 2001. S&P 2001. Proceedings. 2001 IEEE Symposium on*, pages 214–229. IEEE, 2001.

[71] M. S. Bulletin. Vulnerability in server service could allow remote code execution. https://docs.microsoft.com/en-us/security-updates/SecurityBulletins/2008/ms08-067. Accessed: 2019-07-31.

[72] S. A. Butler. Security attribute evaluation method: a cost-benefit approach. In *Proceedings of the 24th international conference on Software engineering*, pages 232–240. ACM, 2002.

[73] E. J. Byres, M. Franz, and D. Miller. The use of attack trees in assessing vulnerabilities in scada systems. In *Proceedings of the international infrastructure survivability workshop*, 2004.

[74] J. Caballero, C. Grier, C. Kreibich, and V. Paxson. Measuring pay-per-install: the commoditization of malware distribution. In *USENIX security symposium*, volume 13, 2011.

[75] A. Caglar. A new approach to risk assessment for cyber insurance. https://technet.microsoft.com/en-us/library/cc163143.aspx. Accessed: 2019-07-31.

[76] I. Cambridge Centre for Risk Studies & Risk Management Solutions. Managing cyber insurance accumulation risk. https://www.jbs.cam.ac.uk/fileadmin/user_upload/research/centres/risk/downloads/crs-rms-managing-cyber-insurance-accumulation-risk.pdf. Accessed: 2019-07-31.

[77] I. Cambridge Centre for Risk Studies & Risk Management Solutions. Sybil logic bomb cyber catastrophe scenario. https://www.jbs.cam.ac.uk/fileadmin/user_upload/research/centres/risk/downloads/crs-sybil-logic-bomb-cyber-catastrophe-stress-test.pdf. Accessed: 2019-07-31.

[78] A. C. Cameron and P. K. Trivedi. *Regression analysis of count data*, volume 53. Cambridge university press, 2013.

[79] D. Canali, L. Bilge, and D. Balzarotti. On the effectiveness of risk prediction based on users browsing behavior. In *Proceedings of the 9th ACM symposium on Information, computer and communications security*, pages 171–182. ACM, 2014.

[80] R. A. Caralli, J. F. Stevens, L. R. Young, and W. R. Wilson. Introducing octave allegro: Improving the information security risk assessment process. Technical report, CARNEGIE-MELLON UNIV PITTSBURGH PA SOFTWARE ENGINEERING INST, 2007.

[81] E. Casey. *Digital evidence and computer crime: Forensic science, computers, and the internet*. Academic press, 2011.

[82] ChromeDevTools. DevTools Protocol API. https://github.com/ChromeDevTools/debugger-protocol-viewer, 2020.

[83] Cliqz GmbH. WhoTracks.me: Bringing Transparency to Online Tracking. https://github.com/cliqz-oss/whotracks.me, 2019.

[84] J. Cloonan. Advanced malware detection - signatures vs. behavior analysis. https://www.infosecurity-magazine.com/opinions/malware-detection-signatures/, 2017. Accessed: 2020-05-04.

[85] C. club da la sécurité de l'information français. Risk management: concept and methods. Technical report, 2009.

[86] M. CLUSIF. Processing guide for risk analysis and management. *Club De La Securite De L'Information Francias, 2nd Edition (April 2011)*, 2010.

[87] I. I. E. Commission. Analysis techniques for system reliability - procedure for failure mode and effects analysis (fmea). Technical report, 2006.

[88] I. I. E. Commission. Iec bs en 61882:2016 - hazard and operability study (hazop studies) - application guide. Technical report, 2016.

[89] E. Condon, M. Cukier, and T. He. Applying software reliability models on security incidents. In *Software Reliability, 2007. ISSRE'07. The 18th IEEE International Symposium on*, pages 159–168. IEEE, 2007.

[90] E. Condon, A. He, and M. Cukier. Analysis of computer security incident data using time series models. In *Software Reliability Engineering, 2008. ISSRE 2008. 19th International Symposium on*, pages 77–86. IEEE, 2008.

[91] A. Cook, R. Smith, L. Maglaras, and H. Janicke. Measuring the risk of cyber attack in industrial control systems. BCS eWiC, 2016.

[92] S. Corbet and C. Gurdgiev. What the hack: Systematic risk contagion from cyber events. *International Review of Financial Analysis*, 65:101386, 2019.

[93] S. Corporation. Ssl 3.0 poodle attack vulnerability. https://support.symantec.com/en_US/article.TECH226102.html. Accessed: 2019-07-31.

[94] S. Dambra, L. Bilge, and D. Balzarotti. Sok: Cyber insurance–technical challenges and a system security roadmap. In *2020 IEEE Symposium on Security and Privacy (SP)*, pages 293–309.

[95] M. Dell'Amico, L. Bilge, A. Kayyoor, P. Efstathopoulos, and P.-A. Vervier. Lean on me: Mining internet service dependencies from large-scale dns data. In *Proceedings of the 33rd Annual Computer Security Applications Conference*, ACSAC 2017, pages 449–460, New York, NY, USA, 2017. ACM.

[96] Disconnect. Make the web faster, more private, and more secure. https://github.com/disconnectme, 2019.

[97] I. Domo. Data never sleeps 5.0, 2018. Available at: https://www.domo.com/learn/infographic/data-never-sleeps-5.

[98] Dymo. Missing Accept_languages in Request for Headless Mode. https://bugs.chromium.org/p/chromium/issues/detail?id=775911, 2017.

[99] EasyPrivacy. Easyprivacy filter subscription. https://github.com/easylist/easylist/tree/master/easyprivacy, 2020.

[100] T. Economist. The world's most valuable resource is no longer oil, but data, 2017. Available at: https://www.economist.com/leaders/2017/05/06/the-worlds-most-valuable-resource-is-no-longer-oil-but-data.

[101] B. Edwards, S. Hofmeyr, and S. Forrest. Hype and heavy tails: A closer look at data breaches. *Journal of Cybersecurity*, 2(1):3–14, 2016.

[102] M. Eling and N. Loperfido. Data breaches: Goodness of fit, pricing, and risk measurement. *Insurance: Mathematics and Economics*, 75:126–136, 2017.

[103] M. Eling and W. Schnell. Extreme cyber risks and the non-diversification trap.

[104] S. Englehardt and A. Narayanan. Online tracking: A 1-million-site measurement and analysis. In *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security (CCS)*, 2016.

[105] Directive 2009/136/EC of the European Parliament and of the Council of 25 November 2009. *Official Journal of the European Union*, 2009.

[106] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). *Official Journal of the European Union*, 2016.

[107] M. Falahrastegar, H. Haddadi, S. Uhlig, and R. Mortier. Tracking personal identifiers across the web. In *Conference on Passive and Active Network Measurement (PAM)*, 2016.

[108] L. Finance. Promoting uk cyber prosperity: Public-private cyber-catastrophe reinsurance, 2015.

[109] U. Franke. The cyber insurance market in sweden. *Computers & Security*, 68:130–144, 2017.

[110] D. Geer, R. Bace, P. Gutmann, P. Metzger, C. P. Pfleeger, J. S. Quarterman, and B. Schneier. Cyberinsecurity: The cost of monopoly. *Computer and Communications Industry Association (CCIA)*, 2003.

[111] N. Gelernter, Y. Grinstein, and A. Herzberg. Cross-site framing attacks. In *Proceedings of the 31st Annual Computer Security Applications Conference*, pages 161–170. ACM, 2015.

[112] Gemalto. Cyber insurance: The challenges facing actuaries in measuring cyber risk. https://blog.gemalto.com/security/2017/10/17/cyber-insurance-challenges-facing-actuaries-measuring-cyber-risk/. Accessed: 2019-07-31.

[113] J. Grossklags, N. Christin, and J. Chuang. Secure or insure?: a game-theoretic analysis of information security games. In *Proceedings of the 17th international conference on World Wide Web*, pages 209–218. ACM, 2008.

[114] S. Hakim, G. F. Rengert, and Y. Shachamurove. *Knowing your odds: Home burglary and the odds ratio*. University of Pennsylvania, Center for Analytic Research in Economics and the Social Sciences, 2000.

[115] A. E. Hassan. Predicting faults using the complexity of code changes. In *Proceedings of the 31st International Conference on Software Engineering*, pages 78–88. IEEE Computer Society, 2009.

[116] A. Hedrick. Cyberinsurance: a risk management tool? In *Proceedings of the 4th annual conference on Information security curriculum development*, page 20. ACM, 2007.

[117] H. Herath and T. Herath. Copula-based actuarial model for pricing cyber-insurance policies. *Insurance Markets and Companies: Analyses and Actuarial Computations*, 2(1):7–20, 2011.

[118] L. Heslault. Actuaries beware: Pricing cyber insurance is a different ballgame. https://www.linkedin.com/pulse/actuaries-beware-pricing-cyber-insurance-different-laurent-heslault. Accessed: 2019-07-31.

[119] K. J. Higgins. More than a half million servers exposed to heartbleed flaw. https://www.darkreading.com/informationweek-home/more-than-a-half-million-servers-exposed-to-heartbleed-flaw/d/d-id/1204318. Accessed: 2019-07-31.

[120] M. Howard, J. Pincus, and J. M. Wing. Measuring relative attack surfaces. In *Computer security in the 21st century*, pages 109–137. Springer, 2005.

[121] X. Hu, G. S. de Tangil, and N. Sastry. Multi-country study of third party trackers from real browser histories. In *2020 IEEE European Symposium on Security and Privacy (EuroS&P)*, pages 70–86. IEEE, 2020.

[122] IAB. The socioeconomic impact of internet tracking. https://www.iab.com/wp-content/uploads/2020/02/The-Socio-Economic-Impact-of-Internet-Tracking.pdf, 2020.

[123] P. International. Social media intelligence and profiling in the insurance industry.... https://medium.com/privacy-international/social-media-intelligence-and-profiling-in-the-insurance-industry-4958fd11f86f, 2017.

[124] Internet Archive. Wayback machine. https://archive.org/, 2020.

[125] U. Iqbal, S. Englehardt, and Z. Shafiq. Fingerprinting the fingerprinters: Learning to detect browser fingerprinting behaviors. *arXiv preprint arXiv:2008.04480*, 2020.

[126] Information technology - Security techniques - Information security risk management. Standard, International Organization for Standardization, June 2011.

[127] ISO 31000:2018 - Risk management: Principles and guidelines. Standard, International Organization for Standardization, Feb. 2018.

[128] J. Jacobs. Analyzing ponemon cost of data breach. *Data Driven Security*, 11, 2014.

[129] B. Johnson, R. Böhme, and J. Grossklags. Security games with market insurance. In *International Conference on Decision and Game Theory for Security*, pages 117–130. Springer, 2011.

[130] B. Karabacak and I. Sogukpinar. Isram: information security risk analysis method. *Computers & Security*, 24(2):147–159, 2005.

[131] A. Karaj, S. Macbeth, R. Berson, and J. M. Pujol. Whotracks. me: Shedding light on the opaque world of online tracking. *arXiv preprint arXiv:1804.08959*, 2018.

[132] J. P. Kesan, R. P. Majuca, and W. J. Yurcik. The economic case for cyberinsurance. 2004.

[133] Kinsta. Global desktop browser market share for 2020. https://kinsta.com/browser-market-share/, 2020.

[134] P. Kocher, D. Genkin, D. Gruss, W. Haas, M. Hamburg, M. Lipp, S. Mangard, T. Prescher, M. Schwarz, and Y. Yarom. Spectre attacks: Exploiting speculative execution. *arXiv preprint arXiv:1801.01203*, 2018.

[135] B. Kordy, S. Mauw, S. Radomirović, and P. Schweitzer. Foundations of attack–defense trees. In *International Workshop on Formal Aspects in Security and Trust*, pages 80–95. Springer, 2010.

[136] P. Kotzias, L. Bilge, and J. Caballero. Measuring PUP prevalence and PUP distribution through pay-per-install services. In *25th USENIX Security Symposium*, pages 739–756, 2016.

[137] P. Kotzias, L. Bilge, P.-A. Vervier, and J. Caballero. Mind your own business: A longitudinal study of threats and vulnerabilities in enterprises.

[138] B. Krishnamurthy and C. Wills. Privacy diffusion on the web: a longitudinal perspective. In *The Web Conference (WWW)*, 2009.

[139] F. Lalonde Levesque, J. Nsiempba, J. M. Fernandez, S. Chiasson, and A. Somayaji. A clinical study of risk factors related to malware infections. In *Proceedings of the 2013 ACM SIGSAC conference on Computer & communications security*, pages 97–108. ACM, 2013.

[140] I. Lapowsky. California unanimously passes historic privacy bill. Wired, 06 2018.

[141] A. Laszka and J. Grossklags. Should cyber-insurance providers invest in software security? In *European Symposium on Research in Computer Security*, pages 483–502. Springer, 2015.

[142] Y. Lau. A brief primer on the economics of targeted advertising. Technical report, Technical report, 2020.

[143] R. Layton and P. A. Watters. A methodology for estimating the tangible cost of data breaches. *Journal of Information Security and Applications*, 19(6):321–330, 2014.

[144] M. Lelarge and J. Bolot. Economic incentives to increase security in the internet: The case for insurance. In *INFOCOM 2009, IEEE*, pages 1494–1502. IEEE, 2009.

[145] A. Lerner, A. K. Simpson, T. Kohno, and F. Roesner. Internet jones and the raiders of the lost trackers: An archaeological study of web tracking from 1996 to 2016. In *USENIX Security Symposium*, 2016.

[146] C. Lever, P. Kotzias, D. Balzarotti, J. Caballero, and M. Antonakakis. A Lustrum of Malware Network Communication: Evolution and Insights. In *Proceedings of the IEEE Symposium on Security and Privacy*. IEEE Computer Society, May 2017.

[147] F. L. Lévesque, J. M. Fernandez, and A. Somayaji. Risk prediction of malware victimization based on user behavior. In *2014 9th International Conference on Malicious and Unwanted Software: The Americas (MALWARE)*, pages 128–134. IEEE, 2014.

[148] T. Libert. Webxray, a tool for analyzing third-party content on webpages and identifying the companies which collect user data. https://github.com/timlib/webXray, 2019.

[149] M. Lipp, M. Schwarz, D. Gruss, T. Prescher, W. Haas, S. Mangard, P. Kocher, D. Genkin, Y. Yarom, and M. Hamburg. Meltdown. *arXiv preprint arXiv:1801.01207*, 2018.

[150] Y. Liu, A. Sarabi, J. Zhang, P. Naghizadeh, M. Karir, M. Bailey, and M. Liu. Cloudy with a chance of breach: Forecasting cyber security incidents. In *USENIX Security Symposium*, pages 1009–1024, 2015.

[151] Y. Liu, J. Zhang, A. Sarabi, M. Liu, M. Karir, and M. Bailey. Predicting cyber security incidents using feature-based characterization of network-level malicious activities. In *Proceedings of the 2015 ACM International Workshop on International Workshop on Security and Privacy Analytics*, pages 3–9, 2015.

[152] M. S. Lund, B. Solhaug, and K. Stølen. *Model-driven risk analysis: the CORAS approach*. Springer Science & Business Media, 2010.

[153] T. Maillart and D. Sornette. Heavy-tailed distribution of cyber-risks. *The European Physical Journal B*, 75(3):357–364, 2010.

[154] R. P. Majuca, W. Yurcik, and J. P. Kesan. The evolution of cyberinsurance. *arXiv preprint cs/0601020*, 2006.

[155] P. K. Manadhata and J. M. Wing. An attack surface metric. *IEEE Transactions on Software Engineering*, (3):371–386, 2010.

[156] A. Marotta, F. Martinelli, S. Nanni, A. Orlando, and A. Yautsiukhin. Cyber-insurance survey. *Computer Science Review*, 24:35–61, 2017.

[157] F. Martinelli, A. Orlando, G. Uuganbayar, and A. Yautsiukhin. Preventing the drop in security investments for non-competitive cyber-insurance market. In *International Conference on Risks and Security of Internet and Systems*, pages 159–174. Springer, 2017.

[158] S. Matic, C. Iordanou, G. Smaragdakis, and N. Laoutaris. Identifying sensitive urls at web-scale. In *Proceedings of the ACM Internet Measurement Conference*, pages 619–633, 2020.

[159] J. R. Mayer and J. C. Mitchell. Third-party web tracking: Policy and technology. In *IEEE Symposium on Security and Privacy (Oakland)*, 2012.

[160] P. H. Meland, I. A. Tøndel, M. Moe, and F. Seehusen. Facing uncertainty in cyber insurance policies. In *International Workshop on Security and Trust Management*, pages 89–100. Springer, 2017.

[161] W. Melicher, M. Sharif, J. Tan, L. Bauer, M. Christodorescu, and P. G. Leon. (do not) track me sometimes: Users' contextual preferences for web tracking. *Proceedings on Privacy Enhancing Technologies*, 2016(2):135–154, 2016.

[162] G. Mezzour, K. M. Carley, and L. R. Carley. An empirical study of global malware encounters. In *Proceedings of the 2015 Symposium and Bootcamp on the Science of Security*, pages 1–11, 2015.

[163] G. Mezzour, L. Carley, and K. M. Carley. Global mapping of cyber attacks. *Available at SSRN 2729302*, 2014.

[164] V. Mishra, P. Laperdrix, A. Vastel, W. Rudametkin, R. Rouvoy, and M. Lopatka. Don't count me out: On the relevance of ip address in the tracking ecosystem. In *The World Wide Web Conference (WWW)*, 2020.

[165] R. Molla. Advertisers will spend $40 billion more on internet ads than on tv ads this year. Recide, 2018.

[166] C. Mood. Logistic regression: Why we cannot do what we think we can do, and what we can do about it. *European sociological review*, 26(1):67–82, 2010.

[167] A. Moshchuk, T. Bragin, S. D. Gribble, and H. M. Levy. A crawler-based study of spyware in the web. In *NDSS*, volume 1, page 2, 2006.

[168] Mozilla Foundation. Security/Tracking protection. https://wiki.mozilla.org/Security/Tracking_protection, 2020.

[169] P. Naghizadeh and M. Liu. Voluntary participation in cyber-insurance markets. In *Workshop on the Economics of Information Security (WEIS)*, 2014.

[170] R. news. Cyber re/insurance market "frustratingly immature". https://www.reinsurancene.ws/cyber-re-insurance-market-frustratingly-immature-inga-beale-lloyds/.

[171] P. Nickinson. The 'stagefright' exploit: What you need to know. https://www.androidcentral.com/stagefright. Accessed: 2019-07-31.

[172] N. Nikiforakis, A. Kapravelos, W. Joosen, C. Kruegel, F. Piessens, and G. Vigna. Cookieless monster: Exploring the ecosystem of web-based device fingerprinting. In *Proceedings of IEEE Symposium on Security and Privacy (Oakland)*, 2013.

[173] S. Noel, S. Jajodia, L. Wang, and A. Singhal. Measuring security risk of networks using attack graphs. *International Journal of Next-Generation Computing*, 1(1):135–147, 2010.

[174] P. Nohe. Re-hashed: 2018 cybercrime statistics: A closer look at the web of profit. https://www.thesslstore.com/blog/2018-cybercrime-statistics/. Accessed: 2019-07-31.

[175] P. Nohe. Study: 66% of smbs would shut down if hit by a data breach. https://www.thesslstore.com/blog/study-66-smbs-shut-hit-data-breach. Accessed: 2019-07-31.

[176] NortonLifeLock. Nortonlifelock global privacy statement, 2021. Available at: https://www.nortonlifelock.com/us/en/privacy/global-privacy-statement/.

[177] R. J. Nurse, L. Axon, A. Erola, I. Agrafiotis, M. Goldsmith, and S. Creese. The data that drives cyber insurance: A study into the underwriting and claims processes. In *2020 International Conference on Cyber Situational Awareness, Data Analytics and Assessment (CyberSA)*, pages 1–8. IEEE, 2020.

[178] I. R. C. of the International Actuarial Association et al. Stress testing and scenario analysis. http://www.actuaries.org/CTTEES_SOLV/Documents/StressTestingPaper.pdf, 2013. Accessed: 2019-07-31.

[179] H. Ogut, N. Menon, and S. Raghunathan. Cyber insurance and it security investment: Impact of interdependence risk. In *WEIS*, 2005.

[180] L. Olejnik, C. Castelluccia, and A. Janc. Why johnny can't browse in peace: On the uniqueness of web browsing history patterns. 2012.

[181] R. Pal and L. Golubchik. On the economics of information security: the problem of designing optimal cyber-insurance contracts. *ACM SIGMETRICS Performance Evaluation Review*, 38(2):51–53, 2010.

[182] R. Pal, L. Golubchik, K. Psounis, and P. Hui. Will cyber-insurance improve network security? a market analysis. In *INFOCOM, 2014 Proceedings IEEE*, pages 235–243. IEEE, 2014.

[183] R. Pal, Z. Huang, S. Lototsky, X. Yin, M. Liu, J. Crowcroft, N. Sastry, S. De, and B. Nag. Will catastrophic cyber-risk aggregation thrive in the iot age? a cautionary economics tale for (re-) insurers and likes. *ACM Transactions on Management Information Systems (TMIS)*, 12(2):1–36, 2021.

[184] P. Papadopoulos, N. Kourtellis, and E. Markatos. Cookie synchronization: Everything you always wanted to know but were afraid to ask. In *The World Wide Web Conference (WWW)*, 2019.

[185] P. Papadopoulos, N. Kourtellis, and E. P. Markatos. The cost of digital advertisement: Comparing user and advertiser views. In *The World Wide Web Conference (WWW)*, 2018.

[186] J. H. Pardue and P. Patidar. Thrats to healthcare date: A threat tree for risk assessment. *Issues in Information Systems*, 12(1):106–113, 2011.

[187] Pi-Hole. Network-wide ad blocking. https://pi-hole.net/, 2021.

[188] W. Pieters, Z. Lukszo, D. Hadziosmanovic, and J. van den Berg. Reconciling malicious and accidental risk in cyber security. *J. Internet Serv. Inf. Secur.*, 4(2):4–26, 2014.

[189] V. L. Pochat, T. Van Goethem, S. Tajalizadehkhoob, M. Korczyński, and W. Joosen. Tranco: A research-oriented top sites ranking hardened against manipulation. *arXiv preprint arXiv:1806.01156*, 2018.

[190] N. Poolsappasit, R. Dewri, and I. Ray. Dynamic security risk management using bayesian attack graphs. *IEEE Transactions on Dependable and Secure Computing*, 9(1):61–74, 2012.

[191] M. Potdar. Cyber insurance market overview. https://www.alliedmarketresearch.com/cyber-insurance-market. Accessed: 2019-07-31.

[192] R. Puricelli. The underestimated social engineering threat in it security governance and management. https://www.isaca.org/Journal/archives/2015/Volume-3/Pages/the-underestimated-social-engineering-threat.aspx.

[193] P. Radanliev, D. De Roure, S. Cannady, R. M. Montalvo, R. Nicolescu, and M. Huth. Economic impact of iot cyber risk-analysing past and present to predict the future developments in iot risk analysis and iot cyber insurance. 2018.

[194] P. A. Ralston, J. H. Graham, and J. L. Hieb. Cyber security risk assessment for scada and dcs networks. *ISA transactions*, 46(4):583–594, 2007.

[195] I. Ray and N. Poolsapassit. Using attack trees to identify malicious attacks from authorized insiders. In *European Symposium on Research in Computer Security*, pages 231–246. Springer, 2005.

[196] N. Robinson. Incentives and barriers of the cyber insurance market in europe. 2012.

[197] F. Roesner, T. Kohno, and D. Wetherall. Detecting and defending against third-party tracking on the web. In *Networked Systems Design and Implementation (NSDI)*, 2012.

[198] S. Romanosky. Examining the costs and causes of cyber incidents. *Journal of Cybersecurity*, 2(2):121–135, 2016.

[199] S. Romanosky, L. Ablon, A. Kuehn, and T. Jones. Content analysis of cyber insurance policies: How do carriers write policies and price cyber risk? 2017.

[200] S. Samat, A. Acquisti, and L. Babcock. Raise the curtains: The effect of awareness about targeting on consumer attitudes and purchase intentions. In *Symposium on Usable Privacy and Security (SOUPS)*, 2017.

[201] I. Sanchez-Rola, D. Balzarotti, C. Kruegel, G. Vigna, and I. Santos. Dirty Clicks: a Study of the Usability and Security Implications of Click-related Behaviors on the Web. In *The World Wid Web Conference (WWW)*, 2020.

[202] I. Sanchez-Rola, D. Balzarotti, and I. Santos. BakingTimer: Privacy Analysis of Server-Side Request Processing Time. In *Annual Computer Security Applications Conference (ACSAC)*, 2019.

[203] I. Sanchez-Rola, M. Dell'Amico, D. Balzarotti, P.-A. Vervier, and L. Bilge. Journey to the Center of the Cookie Ecosystem: Unraveling Actors' Roles and Relationships. In *Proceedings of IEEE Symposium on Security and Privacy (Oakland)*, 2021.

[204] I. Sanchez-Rola and I. Santos. Knockin' on trackers' door: Large-scale automatic analysis of web tracking. In *Proceedings of the International Conference on Detection of Intrusions and Malware, and Vulnerability Assessment (DIMVA)*, 2018.

[205] E. Sangaline. Making Chrome Headless Undetectable. https://intoli.com/blog/making-chrome-headless-undetectable/, 2017.

[206] E. Sangaline. It is Not Possible to Detect and Block Chrome Headless. https://intoli.com/blog/not-possible-to-block-chrome-headless/, 2018.

[207] E. Sangaline. Bypassing Headless Chrome Tests, the game goes on... https://www.tenantbase.com/tech/blog/cat-and-mouse/, 2019.

[208] A. Sarabi, P. Naghizadeh, Y. Liu, and M. Liu. Prioritizing security spending: A quantitative analysis of risk distributions for different business profiles.

[209] Sarah Berry. 2020 search market share: 5 hard truths about today's market. https://www.webfx.com/blog/seo/2019-search-market-share/, 2020.

[210] R. Scandariato, J. Walden, A. Hovsepyan, and W. Joosen. Predicting vulnerable software components via text mining. *IEEE Transactions on Software Engineering*, 40(10):993–1006, 2014.

[211] S. E. Schechter. Toward econometric models of the security risk from remote attack. *IEEE security & privacy*, (1):40–44, 2005.

[212] C. Schmittner, T. Gruber, P. Puschner, and E. Schoitsch. Security application of failure mode and effect analysis (fmea). In *International Conference on Computer Safety, Reliability, and Security*, pages 310–325. Springer, 2014.

[213] B. Schneier. The dangers of a software monoculture. *Information Security Magazine*, 2010.

[214] G. Schwartz, N. Shetty, and J. Walrand. Cyber-insurance: Missing market driven by user heterogeneity. *preparation, www. eecs. berkeley. edu/nikhils/SecTypes. pdf*, 2010.

[215] G. A. Schwartz and S. S. Sastry. Cyber-insurance framework for large scale interdependent networks. In *Proceedings of the 3rd international conference on High confidence networked systems*, pages 145–154. ACM, 2014.

[216] M. Sharif, J. Urakawa, N. Christin, A. Kubota, and A. Yamada. Predicting impending exposure to malicious content from user behavior. 2018.

[217] N. Shetty, G. Schwartz, M. Felegyhazi, and J. Walrand. Competitive cyber-insurance and internet security. In *Economics of information security and privacy*, pages 229–247. Springer, 2010.

[218] N. Shetty, G. Schwartz, and J. Walrand. Can competitive insurers improve network security? In *International Conference on Trust and Trustworthy Computing*, pages 308–322. Springer, 2010.

[219] O. Sheyner, J. Haines, S. Jha, R. Lippmann, and J. M. Wing. Automated generation and analysis of attack graphs. In *Security and privacy, 2002. Proceedings. 2002 IEEE Symposium on*, pages 273–284. IEEE, 2002.

[220] O. Sheyner and J. Wing. Tools for generating and analyzing attack graphs. In *International Symposium on Formal Methods for Components and Objects*, pages 344–371. Springer, 2003.

[221] W. Shim. An analysis of information security management strategies in the presence of interdependent security risk. *Asia Pacific Journal of Information Systems*, 22(1):79–101, 2012.

[222] Y. Shin and L. Williams. An empirical model to predict security vulnerabilities using code complexity metrics. In *Proceedings of the Second ACM-IEEE international symposium on Empirical software engineering and measurement*, pages 315–317. ACM, 2008.

[223] M. Simeonovski, G. Pellegrino, C. Rossow, and M. Backes. Who controls the internet?: Analyzing global threats using property graph traversals. In *Proceedings of the 26th International Conference on World Wide Web*, WWW '17, pages 647–656, Republic and Canton of Geneva, Switzerland, 2017. International World Wide Web Conferences Steering Committee.

[224] K. Solomos, P. Ilia, S. Ioannidis, and N. Kourtellis. Clash of the trackers: measuring the evolution of the online tracking ecosystem. *arXiv preprint arXiv:1907.12860*, 2019.

[225] M. Stamp. Risks of monoculture. *Communications of the ACM*, 47(3):120, 2004.

[226] G. Stoneburner, A. Y. Goguen, and A. Feringa. Sp 800-30. risk management guide for information technology systems. 2002.

[227] D. W. Straub Jr. Effective is security: An empirical study. *Information Systems Research*, 1(3):255–276, 1990.

[228] B. Sussman. 5 reasons cyber insurance market will hit $23 billion. https://www.secureworldexpo.com/industry-news/5-reasons-cyber-insurance-market-will-hit-23-billion. Accessed: 2019-07-31.

[229] Symantec. The need for threat risk levels in secure web gateways. https://docs.broadcom.com/doc/need-for-threat-risk-levels-in-secure-web-gateways-en, 2017.

[230] Symantec. Webpulse. https:www.symantec.com/content/dam/symantec/docs/white-papers/webpulse-en.pdf, 2017.

[231] S. Szczypinski. What is financial profiling? https://www.lexingtonlaw.com/blog/finance/financial-profiling.html, 2021.

[232] I. Tarlowska and A. Zebrowska. Customer profiling for credit decisions made easy for the financial industry under new polish legislation. https://www.jdsupra.com/legalnews/customer-profiling-for-credit-decisions-85940/, 2019.

[233] C.-W. Ten, C.-C. Liu, and M. Govindarasu. Vulnerability assessment of cybersecurity for scada systems using attack trees. In *Power Engineering Society General Meeting, 2007. IEEE*, pages 1–8. IEEE, 2007.

[234] The World Wide Web Consortium. Same origin policy. https://www.w3.org/Security/wiki, 2020.

[235] O. Thonnard, L. Bilge, A. Kashyap, and M. Lee. Are you at risk? profiling organizations and individuals subject to targeted attacks. In *International Conference on Financial Cryptography and Data Security*, pages 13–31. Springer, 2015.

[236] A. R. Tomas Girnius, Scott Stransky. Aggregated cyber risk: The nightmare scenarios. http://www.air-worldwide.com/Publications/AIR-Currents/2015/Aggregated-Cyber-Risk--The-Nightmare-Scenarios/. Accessed: 2019-07-31.

[237] I. A. Tøndel, P. H. Meland, A. Omerovic, E. A. Gjære, and B. Solhaug. Using cyber-insurance as a risk management strategy: Knowledge gaps and recommendations for further research. 2015.

[238] C. Toregas and N. Zahn. Insurance for cyber attacks: The issue of setting premiums in context. *George Washington University*, 2014.

[239] H. ulisi Ogut and S. Raghunathan. Cyber insurance and it security investment: Impact of interdependent risk.

[240] G. Uuganbayar, A. Yautsiukhin, F. Martinelli, and F. Massacci. Optimisation of cyber insurance coverage with selection of cost effective security controls. *Computers & Security*, 101:102121, 2021.

[241] N. Vallina-Rodriguez, J. Shah, A. Finamore, Y. Grunenberger, K. Papagiannaki, H. Haddadi, and J. Crowcroft. Breaking for commercials: characterizing mobile advertising. In *Internet Measurement Conference (IMC)*, 2012.

[242] S. Vanden Broucke and B. Baesens. *Practical Web scraping for data science*. Springer, 2018.

[243] E. J. Vaughan and T. Vaughan. *Fundamentals of risk and insurance*. John Wiley & Sons, 2007.

[244] V. Verendel. Quantified security is a weak hypothesis: a critical survey of results and assumptions. In *Proceedings of the 2009 workshop on New security paradigms workshop*, pages 37–50. ACM, 2009.

[245] T. Vissers, N. Nikiforakis, N. Bielova, and W. Joosen. Crying wolf? on the price discrimination of online airline tickets. In *7th Workshop on Hot Topics in Privacy Enhancing Technologies (HotPETs 2014)*, 2014.

[246] J. Walden, J. Stuckman, and R. Scandariato. Predicting vulnerable components: Software metrics vs text mining. In *Software Reliability Engineering (ISSRE), 2014 IEEE 25th International Symposium on*, pages 23–33. IEEE, 2014.

[247] B. Weinshel, M. Wei, M. Mondal, E. Choi, S. Shan, C. Dolin, M. L. Mazurek, and B. Ur. Oh, the places you've been! user reactions to longitudinal transparency about third-party web tracking and inferencing. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, pages 149–166, 2019.

[248] S. Wheatley, T. Maillart, and D. Sornette. The extreme risk of personal data breaches and the erosion of privacy. *The European Physical Journal B*, 89(1):7, 2016.

[249] S. White. What is GDPR's impact on Cyber Insurance? https://gdpr.report/news/2019/05/15/what-is-the-impact-of-gdprs-on-cyber-insurance. Accessed: 2019-07-31.

[250] Wikipedia. 2016 dyn cyberattack: Timeline and impact. https://en.wikipedia.org/wiki/2016_Dyn_cyberattack. Accessed: 2019-07-31.

[251] Wikipedia. Filter bubble. https://en.wikipedia.org/wiki/Filter_bubble, 2021.

[252] J. Williams. Java serialization vulnerability threatens millions of applications. https:

//www.contrastsecurity.com/security-influencers/java-serialization-vulnerability-threatens-millions-of-applications. Accessed: 2019-07-31.

[253] J. Wolff. Cyberinsurance tackles the wildly unpredictable world of hacks. https://www.wired.com/story/cyberinsurance-tackles-the-wildly-unpredictable-world-of-hacks/. Accessed: 2019-07-31.

[254] D. Woods, I. Agrafiotis, J. R. Nurse, and S. Creese. Mapping the coverage of security controls in cyber insurance proposal forms. *Journal of Internet Services and Applications*, 8(1):8, 2017.

[255] D. W. Woods and R. Böhme. Systematization of knowledge: Quantifying cyber risk. In *IEEE Symposium on Security & Privacy*, 2021.

[256] C. Wright. Understanding kaminsky's dns bug. http://www.linuxjournal.com/content/understanding-kaminskys-dns-bug. Accessed: 2019-07-31.

[257] Z. Yang and J. C. Lui. Security adoption and influence of cyber-insurance markets in heterogeneous networks. *Performance Evaluation*, 74:1–17, 2014.

[258] A. A. Yayla and Q. Hu. The impact of information security events on the stock value of firms: The effect of contingency factors. *Journal of Information Technology*, 26(1):60–77, 2011.

[259] T.-F. Yen, V. Heorhiadi, A. Oprea, M. K. Reiter, and A. Juels. An epidemiological study of malware encounters in a large enterprise. In *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*, pages 1117–1130. ACM, 2014.

[260] Zenith. Adspend forecast live. http://adforecast.zenithmedia.com/, 2020.

[261] X. Zhao, L. Xue, and A. B. Whinston. Managing interdependent information security risks: A study of cyberinsurance, managed security service and risk pooling. *ICIS 2009 Proceedings*, page 49, 2009.

[262] T. Zimmermann, N. Nagappan, and L. Williams. Searching for a needle in a haystack: Predicting security vulnerabilities for windows vista. In *Software Testing, Verification and Validation (ICST), 2010 Third International Conference on*, pages 421–428. IEEE, 2010.