UNIVERSITY OF TARTU
Faculty of Science and Technology
Institute of Computer Science
Cybersecurity Curriculum

Andres Jõgi

Quantitative Analysis on Vulnerability to Electronic Business Identity Theft Among Estonian Companies

Master's Thesis (21 ECTS)

Supervisors: Mari Seeba, MSc
Tarmo Oja, MSc
Markko Merzin, MSc

Tartu 2024

Quantitative Analysis on Vulnerability to Electronic Business Identity Theft Among Estonian Companies

Abstract: As 60% of all information security incidents resulting in data breaches involve social engineering, it is essential to understand the extent and motivations behind them. Electronic identity theft is an important component to cyber attacks involving social engineering techniques. While cybercriminals opportunistically exploit the identities of organisations and physical persons, many successful attacks focus on impersonating organisations. Current research focuses on analysing the impact of past cyber attacks and in-depth analysis of attack techniques with vulnerability surface measured across the internet. However, technical vulnerabilities should be tied to the entities responsible for vulnerable assets to predict and prevent potential attacks across organisations. This thesis aims to reduce this gap by providing insights into active organisations' vulnerabilities, from micro-enterprises to large multinational corporations. Data from the Estonian e-Business Register is used to conduct a case study that ties digital assets to a responsible legal entity, allowing for actionable information on vulnerability trends to be analysed and used to improve resilience against impersonation attacks. Discovered vulnerabilities have been forwarded to Estonian Computer Security Incident Response Team (CERT-EE) who used this information to notify affected service providers.

Keywords:
Electronic Identity, Identity Theft, Estonia, e-Business Register, Social Engineering

CERCS: T120 - Systems engineering, computer technology

Kvantitatiivne analüüs Eesti ettevõtete haavatavusest elektroonilise äriidentiteedi vargusele

Lühikokkuvõte:

Tulenevalt suhtlusrünnete olulisest rollist üle 60% andmelekkega lõppenud turvaintsidentides, on rünnete tõkestamiseks vajalik mõista rünnete põhjuseid ja ulatust. Elektrooniline identiteedivargus on oluline komponent suhtlusründel põhinevate küberrünnakute õnnestusmises. Kuigi küberkurjategijad kasutavad vastavalt võimalustele ära nii eraisikutelt kui ettevõtetelt varastatud identiteeti, teeseldakse edukate õngitsusrünnete puhul eelkõige ettevõtte poolset suhtlust. Enamik senistest uurimustest keskendub toimunud rünnete põhjuste analüüsile või kitsaste tehniliste rünnete ja nendele haavatavate teenuste otsimisele üle interneti. Potentsiaalsete rünnete ennetamiseks või veel avastamata sissemurdmiste tuvastamiseks ettevõtete üleselt oleks vaja siduda arvutivõrgust leitud haavatavused nende eest vastutava osapoolega. Käesoleva töö eesmärk on tõsta teadmust, kuidas teaduskirjanduses kirjeldatud tehnilised haavatavused mõjutavad tegutsevaid organisatsioone mikroettevõtetest rahvusvaheliste suurettevõteteni. Uurimus tugineb Eesti äriregistri andmetel, mille alusel seotakse ettevõtete sidevahendid vastava juriidilise isikuga. Töö tulemusel loodud olukorrapilt juhib tähelepanu suurema mõjuga haavatavustele ning aitab suunata vastumeetmete rakendamist. Töö tulemusena tuvastatud turvanõrkused edastati CERT-EE-le, kes teavitas haavatavaid organisatsioone avastatud nõrkustest.

Võtmesõnad:
elektrooniline identiteet, identiteedivargus, Eesti, E-äriregister, suhtlusrünne

CERCS: T120 - Süsteemitehnoloogia, arvutitehnoloogia

Acknowledgments

I thank my supervisors, Mrs Mari Seeba and Mr Tarmo Oja, for their guidance and support. Their insights and consistent drive to strive beyond good enough towards something to be proud of are highly appreciated. I would also like to thank Mr Markko Merzin for validating the initial concept and for his many quotable comments that helped to shape my approach to information security.

I wrote this thesis while working as a part of the fantastic information security team at Bolt; encouragement and advice from my amazing colleagues allowed me to tackle this topic.

I would also like to thank the Estonian Information System Authority's CERT-EE team for validating the findings and providing necessary feedback.

Finally, I would like to thank my wonderful fiancee, Marielle, and our daughter, Madeleen, for their patience and support. Without them, writing this thesis would not have been possible.

Contents

1 Introduction **7**

2 Background **8**
2.1 Identity 8
2.2 Electronic identification 9
2.2.1 Identification of physical persons 9
2.2.2 Identification of legal persons 10
2.3 Electronic identity theft 11
2.3.1 Definition & Motive 11
2.3.2 Methods 12
2.3.3 Opportunities 13
2.4 Related works 14
2.4.1 Data breaches 14
2.4.2 Phishing 15
2.4.3 Cyber squatting 16
2.4.4 Spoofing 17
2.5 Summary 19

3 Research problem **20**

4 Research design **21**
4.1 Constraints and assumptions 21
4.2 Ethical considerations 21
4.3 Data selection 22
4.3.1 Identifying digital assets from the data source 22
4.3.2 Selecting the dataset 23
4.4 Selecting indicators 25
4.4.1 Vulnerability indicators 25
4.4.2 Indicators of Compromise 27
4.5 Instrumentation 28
4.5.1 Preprocessing steps 28
4.5.2 Measuring significant (L1) vulnerabilities 29
4.5.3 Measuring notable (L2) and informative vulnerabilities (L3) 30
4.5.4 Measuring indicators of compromise 31

5 Results **33**
5.1 Vulnerability measurements 33
5.1.1 VULN W.1 Detecting unregistered website domains 33
5.1.2 VULN W.2 Missing or broken transport layer encryption . . 35

5.1.3 VULN W.3 Missing HSTS support . . . 38
5.1.4 VULN E.1 Detecting unregistered email domains . . . 38
5.1.5 VULN E.2.1 Usage of a private email address for business purposes . . . 41
5.1.6 VULN E.2.2 Missing SPF & DMARC records . . . 44
5.1.7 VULN M.1 Detecting unregistered phone numbers . . . 47
5.2 IOC measurements . . . 49
5.2.1 IOC E.1 Leaked credentials . . . 49
5.2.2 IOC W.1 Unexpected redirection . . . 51
5.3 Summary . . . 52

6 Validation **55**
6.1 Issues with domain extraction . . . 55
6.2 Notifying CERT-EE . . . 55

7 Discussion **57**
7.1 Limitations . . . 58

8 Conclusions **60**
8.1 Future work . . . 60

Appendix I. Glossary **70**

Appendix II. Licence **71**

Appendix III. List of leaked data types (IOC E.1) **72**

1 Introduction

Business identity theft is a type of crime in which someone wrongfully impersonates a business for fraud or deception[1]. According to industry authorities, such techniques are often used in social engineering attacks to make the fraudulent request appear legitimate. Due to impersonation's relative simplicity and effectiveness in social engineering, it is among the most popular elements of modern cybercrime. The European Union Agency for Cybersecurity (ENISA) reports that the elements of social engineering have been observed in over 60% of data breaches. [2]

However, as the knowledge of the extent and impact of identity theft relies heavily on the reported cases, the situational picture of the extent of such risks is incomplete [3]. This thesis aims to reduce the knowledge gap by measuring available opportunities for electronic identity theft for impersonation of businesses. To achieve this goal, a case study was conducted to measure the vulnerability to common electronic identity theft methods across active organisations registered in the Republic of Estonia.

The thesis utilises open data from the Estonian e-Business Register[4] to gather digital assets that represent surveyed organisations' electronic identities. The security level of these assets is measured against common impersonation techniques identified in academic literature. The resulting data provides a comprehensive overview of the vulnerabilities affecting businesses of all sizes, from micro-enterprises to large multinational corporations. The impact of the company size and revenue on the level of implemented identity protection measures is analysed, and recommendations for improving the current situation are made.

The thesis is split into eight sections, including introduction and conclusions. Section 2 introduces electronic identity, giving a brief overview of general concepts, regulations, current problems surrounding electronic identity theft, and existing research on the topic. Sections 3 and 4 state the research problem and describe the methods chosen for answering the stated research questions. Section 5 describes the experimentation results. Sections 6 and 7 describe how the findings were validated and how the results contrast and build upon the previous works on the topic. Section 8 discusses and summarises the findings and limitations that could be addressed as part of future works in this domain.

2 Background

This section provides the necessary background on the subject matter to aid in understanding the thesis results. General concepts relating to electronic identity and identity theft are explained, and an overview of current problems in this space is given. Special attention is given to the legal environment in the European Union and one of its member states - Estonia.

2.1 Identity

NIST-SP800-63-3[5] defines identity as "An attribute or set of attributes that uniquely describe a subject within a given context." This definition covers the common usage of this term, at least in the context of information systems, but the implications apply outside the domain of information technology as well. It is important to remember that such identifying attributes are not limited to a single context. In real-world use, utilising situationally available attributes for authenticating entities is common, even if their proof value is less than ideal.

As an example, the value of scientific articles is often measured based on the number of citations or the impact factor of the journal in which they were published. Yet, they are not ideal proxies for measuring meaningful contributions to the field. However, as they are easy to measure, they are used until a more suitable alternative is found. A similar effect can be seen in the information system context as the availability of strong authenticators is limited, and readily available means of authentication are used instead.

Such occurrences occur during most transactions between a private person and an online service. Both parties, the user and the service provider, need to verify the legitimacy of the other party before committing to sensitive transactions. The user needs to confirm the legitimacy of the service, and the service provider needs to verify the authenticity of the user to the extent required by the operational context.

However, there exists a disparity in the level of identity proofing available to each party. The service provider has control over procedural and technological means used to conduct and verify the legitimacy of the transaction[6]. For the user, the situation is quite different. As available technical and non-technical attributes and their interpretation differ case-by-case, it is difficult to authenticate the website before receiving the goods. As an example, fraudulent e-commerce websites have been extensively studied to understand the motivation and techniques used to fool users into buying counterfeit goods from untrustworthy e-shop[7] and to propose countermeasures for their detection[8, 9, 10].

Yet despite extensive research, reliable electronic identification is not a problem that has been able to be solved at a significant scale.

2.2 Electronic identification

There have been efforts to unify legislation around electronic identification, especially for identifying physical or legal persons via electronic means. This section provides background on the commonly used electronic identification methods of physical and legal persons. It overviews the assurance levels and issues associated with electronic identification schemes.

2.2.1 Identification of physical persons

Due to the relatively low access barrier for many online environments, a physical person is not limited to a single online identity. The entities participating in casual online interactions on social media applications, mobile apps and online games can not, with a high degree of certainty, be tied to a specific physical person. The context in which many information systems identify their end-users relies on verification of continuity of the account ownership from initial registration to all the subsequent logins. For the context of such information systems, different entities are identified based on the date entered during self-registration, such as username and password or proof of ownership for a bound social media account, phone number or email address.[11]

This becomes problematic when the context changes, e.g. the user account has been used to conduct fraud or other criminal activities. According to theories on criminal justice, attribution and punishment for misdeeds committed to contribute towards deterrence from future offences [12]. Yet, as the administration of justice takes place in a different context from the "happy path" of benign use of information systems, additional attributes are necessary to identify the guilty party. Due to the large proportion of fraudulent actors targeting victims online, a similar need for stronger authentication exists outside law enforcement, as laypersons need to identify scammers before falling victim to their schemes. Yet internet fraud has evolved alongside the rest of the digital society, with new tools such as artificial intelligence and machine learning finding their way into the toolbox of fraudsters. Thus, identifying legitimate parties from pretenders is becoming more complex over time. [13]

The situation is somewhat better for online interactions where a higher level of confidence has been considered necessary by design, such as e-government services or electronic signatures. The EU eIDAS directive regulates the level of assurances given through electronic identification across the member states[14]. With the member states adopting the EU law into their national legislation, the level of electronic identification in the regulated contexts improves. As an example, requirements for e-government services in the Republic of Estonia specify the appropriate eIDAS-compliant identification assurance level for the service based

on the risks resulting from unauthorized access to it. Such requirements cover the whole life-cycle of compliant authenticators, including technical descriptions and the level of identity verification on issuance. [15]

As eIDAS and similar directives do not cover electronic identification in every context, weaker identification schemes are still extensively used. This leaves an open opportunity for fraudsters to abuse weaker identification schemes for their benefit. With both ENISA[16] and Estonian Information System Authority[13] reporting emails being widely used to send out fraudulent bills or requests for change of banking information.

2.2.2 Identification of legal persons

While all online interactions involve physical persons in some form, most online interactions are done on behalf of legal persons. Yet identifying legal entities online seems more difficult than doing so for physical persons.

Similarly to identifying physical persons online, identifying legitimate communication channels and other digital assets of a legal person is not straightforward. The identity of a corporation tends to be confused with its associated branding and trademarks. However, the presence of recognized branding elements in electronic communications is, at most, a weak proof of the legal entity behind them. To illustrate this, a court dispute between Uzi Nissan and Nissan Motor Co. during the early 2000s over the usage rights of *nissan.com* can be used. In principle, a Domain Name System (DNS) domain can be registered by anyone, even if the domain name represents an established global brand.[17]

Similarly, using branding elements of well-known organisations, such as colour schemes, logos, or other visual attributes, has an equally low barrier of entry. For financial institutions, the principle of Know Your Customer (KYC) and its extension Know Your Business (KYB) mandates strict verification, identification and assessment of the risks resulting from starting or maintaining business relationships with all potential clients. This process is aided by supporting regulations such as the Commercial Register Act of Estonia[18] or the Anti-Money Laundering directive of EU[19]. As such, the legislation mandates the creation of national registers of judicial bodies that help to maintain basic information about companies and their beneficiaries; depending on the regulation itself, the registry can also contain information on official digital communication channels used by the company. For example, the e-Business Register of Estonia accommodates email addresses, telephone numbers and website addresses of the organisations listed in it[18].

In addition to KYC-mandated background checks, this data is used in other domains as a trusted authority for checking the legitimacy of the judicial body in different transactions such as stricter owner verification checks of SSL/TLS certificate issuance [20]. As a result, relying on a central authority to authenticate

online assets' relation to registered corporations results in an authentication scheme similar to the ones used to describe federated identity concepts.

This makes the contents of the e-Business register a de-facto authority binding legal entity to its communication channels (e.g. email domain, website and official phone number). The confidence level provided by this sort of "federated authentication" might be less than expected in many contexts. For example, the extended certificate validation trust chain has been subverted by registering a corporation with the same name as the target in another juristiction[21].

Due to this, there are many cases where well-known organisations are successfully impersonated for fraudulent purposes.

2.3 Electronic identity theft

Identity theft of any kind is a complex topic; for the scope of this thesis, simplification based on the 3-element model commonly used in criminal investigations is used. This subsection introduces modern electronic identity theft's role in cyber attacks by analysing the motives, methods and opportunities that drive these misdeeds. [3]

2.3.1 Definition & Motive

For the scope of this work, electronic identity theft is defined as the misuse of personal data by involving fraud or deception via electronic communication channels. By relying on the highly cited work of Newman et al., we can further dissect the concept of identity theft into separate stages[3]:

1. *Acquisition* of identifying information. This can be done legally (e.g., public information or purchase information) or illegally (e.g., through theft, computer hacking, fraud, trickery, or even physical force).

2. *Use* of acquired identity for either financial gain or for masking one's own identity.

3. *Discovery* of the identity theft, which, depending on the type of misuse, might take considerable time (e.g. months to years). Yet, it can be assumed that there are cases where the discovery stage never happens.

As the target of the acquisition stage is often not the same entity that is deceived in the usage phase, identity theft is considered a "dual crime", with one of the victims having their identity stolen and the other being deceived with a false identity. This means that the burden of successful identity theft influences multiple parties.[3]

Identity theft does not tend to be the end goal of the misdeed itself. Rather, impersonation is often used to aid in achieving the actual goal. In many cases,

the end goal is obtaining financial gain (e.g., hopes of defrauding someone via impersonation of a trusted party) but also acquiring restricted information or concealing the impersonator's real identity to avoid retaliation.[3]

In recent years, electronic identity theft has been noted to be tightly coupled with cyberattacks both as an end goal of intrusion and as one of the tools in the attacker's arsenal[16]. According to ENISA's 2020 threat landscape report on phishing[22], identity theft in phishing attacks is commonly used to gain initial access to the targeted systems. Yet it is not uncommon for impersonation to be used in later stages of intrusion to act on targets, such as is the case with Business Email Compromise (BEC) attacks intended to defraud or extort from victims[16].

2.3.2 Methods

The use of identity theft in different cyberattacks can be mapped to phases in Lockheed Martin's Cyber Kill Chain model of cyberattacks (see Figure 1) where the acquisition stage corresponds only to the very first *reconnaissance* phase of the kill-chain[1] and the usage stage encompassing the remaining six.

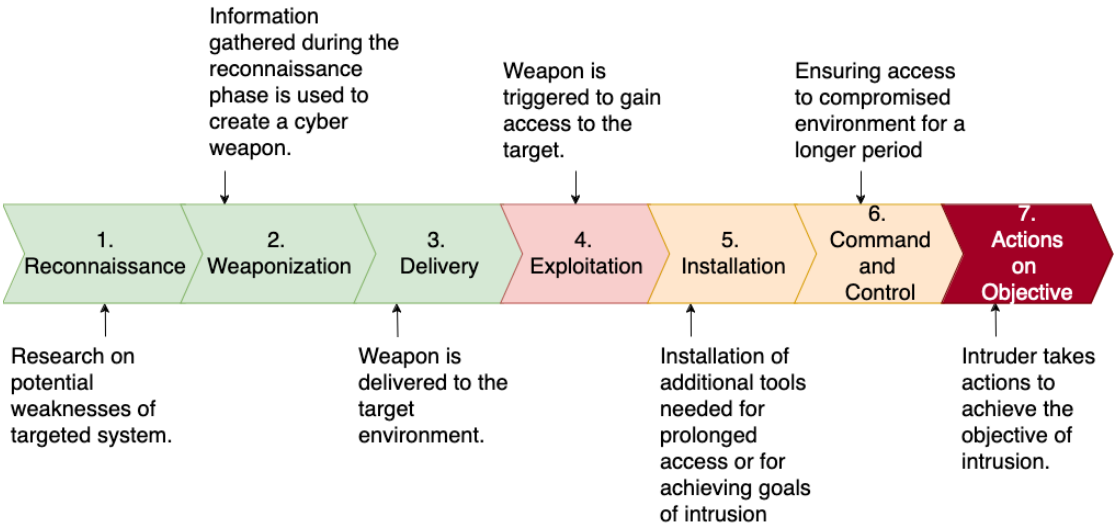


Figure 1. Visual representation of the sequence of phases in the Lockheed Martin's Cyber Kill Chain model of cyberattacks.

The methods used for the reconnaissance phase of the acquisition stage can be relatively low-tech. Information can be acquired via physical means such as

[1]Although it should be noted that the reconnaissance itself might entail full-fledged cyber attacks and attempts of identity theft against secondary targets. This is done to gather information and other resources to support the main effort (such as attacking personal accounts and devices of those working for the targeted organisation).[16]

stealing wallets, purses or mailboxes, searching through residential trashcans or by a personal relationship with the victim [3]. Yet, in the current information age, information can be harvested through the internet at an unprecedented scale.

Publicly available information from the presence of social media platforms and corporate websites can be used to gain the organisational context necessary to impersonate an organisation or its employees. Users often divulge many personal details on these platforms, encompassing their identities, birthdates, residential addresses, and daily routines. Malicious actors can harness this wealth of information to craft convincing impersonations, enabling them to perpetrate identity theft or targeted scams with alarming precision.[23]

In addition, information from previous data leaks and breaches gathered from the internet or purchased from other cybercriminals can be used for further misdeeds [16]. Depending on the contents of leaked data, it can contain login credentials, personal information usable for more crafting detailed phishing e-mails or even credit card information. This, combined with the prevalence of password reuse and easily guessable combinations, facilitates unauthorized access to user accounts across various online platforms. Trying out commonly used passwords or known previous passwords of targeted users can enable easy access to legitimate user accounts for fraudsters.[24]

The weaponization of the information gathered in the reconnaissance phase depends heavily on the end goal of the activities. It can entail leveraging gathered information to craft a believable lure (such as an email message or a lookalike website) for pretexting or phishing attacks. The delivery phase of the attack involves delivering the crafted lure to its intended victim; common mediums for this are unsolicited emails, social media messages, SMS messages, and phone calls. [16]

The exploitation is the pivotal moment for the attacker, as the overall success of their activities depends on it. Depending on the type of the attack, the exploitation might entail attempts to use the stolen credentials to gain access to the target information system, sale or usage attempt of stolen credit card data or execution of actual malware on the victim's computer.

Installation, command and control steps from the kill-chain model do not usually entail identity theft. However, acting on the objectives can be another attack, such as using a compromised account for further intrusion or sending out fraudulent bills under the newly assumed identity.

2.3.3 Opportunities

According to ENISA Threat Landscape Report 2023[16], social engineering remains one of the preferred approaches by threat actors due to its simplicity, low cost and ease of execution. The continued trend of increasing social engineering attacks

powered by identity theft, such as phishing, indicates numerous opportunities to be exploited. Anti-Phishing Working Group (APWG) reporting increasing volume year-on-year for reported phishing attacks, with 2023 being the worst year to date with almost 5 million phishing attacks observed through the year[25].

Increased specialization of cybercrime groups has been observed to lower the barrier of entry for benefiting from fraudulent activities in cyberspace. While insights into cybercrime have noted that specialized knowledge of computers and networks has been required for the criminals to be influential, the emergence of markets selling malware, personal data and tools has made it easier to participate in cyberattacks and fraud [26].

As an example of specialisation, the emergence of Phishing-as-a-Service (PhaaS) groups has been noted. These phishing kits are offered to would-be scammers for as little as $15 per day or a flat $40. Such kits come with all the capabilities and tooling, such as email and website templates, lists of potential targets and detailed instructions on launching an attack, making it easy for scammers with lower skill sets to conduct convincing phishing attacks. In addition, the groups specialising in providing services lower the risk for themselves by not directly participating in the fraud. [27]

Due to the simplicity of such attacks, social engineering attacks via electronic channels are expected to remain amongst the top cyber threats for the foreseeable future[16]. An in-depth overview of prevalent technical weaknesses abused for such purposes, as highlighted by academia, is introduced in the following subsection.

2.4 Related works

This subsection gives an overview of standard impersonation techniques used to deceive potential victims, as identified by previous works.

2.4.1 Data breaches

Data breaches consistently rank amongst the most impactful cyber threats in recent ENISA Threat Landscape reports [2, 16, 28]. The total number of accounts breached since 2004 surpassed 16.7 billion, with 6.3 billion unique emails found amongst the known breaches[29].

Due to their relevance, data breaches and leaks have been thoroughly studied in academic circles across multiple disciplines. According to the review by Schlackl et al., most of the research on the subject has focused on the preconditions and consequences for the breached organisations. They note that further research towards the consequences to data subjects whose data was disclosed due to leaks could be beneficial.[30]

KD. Martin et al. also point out that wide-scale usage and resale of customer data in marketing increases customers' vulnerability to falling victim to a data breach or identity theft. They also mention that customer identity theft resulting from a corporate data breach might be abused to defraud the corporation itself. While not all victims of data breaches experience identity theft, the uncertainty and lack of control make the most significant impact of the data breaches, the feeling of violation and loss of trust, not the financial losses from the misuse itself.[31]

Since the emergence of General Data Protection Regulation (GDPR)[32] and other data breach disclosure laws, there has been discussion on their impact. Often claimed goals of such legislation are to reduce the opportunities for identity theft by notifying the potential victims of the incident and allowing them to take appropriate actions to defend themselves. The aim is also to encourage companies to resolve their information security gaps to reduce the likelihood of falling victim to data breaches. Yet while there is some empirical research done on the actual positive impact[33, 34], both quantity and quality of available data on such incidents limits the generalisations.[34]

2.4.2 Phishing

While identity theft does not necessitate the use of technology, there is a subset of social engineering attacks utilizing digital means and impersonation known as phishing. Phishing attacks rely upon social engineering techniques to coerce their target to reveal confidential information via unsolicited digital communications (e.g. SMS, voice, social media, e-mail or website) pretending to be trustworthy. [35]

Arguably, the best-known phishing attacks could be classified as digital fraud. The variation of the advance-fee scam known as the Nigerian Prince scam is a well-known example of this. Such a scam involves the perpetrator contacting the victim via email and promising them a large sum in return for a smaller up-front payment. After the victim makes a payment, the attacker either ceases all communications or invents further fees for the victim to pay. This type of attack does not have substantial technological components, with similar scams conducted via postal service as early as the 16th century (e.g. Spanish Prisoner scam). [36]

Phishing attacks are also known to be used as malware delivery vectors for initial access. ENISA Threat Landscape 2022 reports Phishing as the most common initial vector for ransomware attacks [2]. In the case of malware distribution attacks, targets are enticed to open attached files containing malicious documents and scripts that result in the downloading and running of malware. In some cases, phishing covers reconnaissance, weaponisation and delivery stages of a wider cyber attack. [37]

In addition to fraud and malware distribution, the primary aim of a large percentage of phishing attacks is data theft. Themes for data theft vary from

credential harvesting operations and stealing confidential or personal information to trying to gain access to financial accounts [25]. For these types of attacks, the common technique the attackers use is to lure their target to follow a fraudulent Uniform Resource Locator (URL), which will be used for collecting personal information or account credentials.[2]

Highly targeted phishing attacks are known as spear phishing. Spear phishing relies upon comprehensive reconnaissance of the target before the attack. These preparations enable perpetrators to craft personalized phishing lures for each target. Due to these steps, spearphishing attacks are reported to be one of the most dangerous types of phishing attacks since targeted attacks have a higher chance of successfully baiting their target.[2]

One of the financially most impactful uses of spearphishing techniques is known as Business Email Compromise (BEC). In case of a BEC attack, attackers craft a seemingly legitimate email message from a known source asking for a funds transfer. Due to the technical simplicity of this type of attack, ENISA estimates that this type of phishing attack will remain popular in the near future.[2]

2.4.3 Cyber squatting

While data breaches and phishing attacks are well-known attacks frequently listed among top cyber threats, domain squatting is not always considered a standalone threat. In its essence, squatting is registering the domain of a well-known trademark or brand, the primary motivation for which is the expectation of monetary gain. Common monetization schemes include domain parking, affiliate- and trademark abuse, and phishing. [38]

The earliest form of domain squatting that emerged is called cybersquatting, which entails speculative registration of a trademark or brand name as a domain name before the legitimate owner can do it. In these cases, the squatters aim to target financial gain by eventually selling the domain to a legitimate trademark holder with a high markup and leveraging its popular name for ad revenue. [38]

Even the squatting domain itself does not have to be identical to the targeted trademark or brand; academic literature tends to differentiate six additional types of squatting:

- *typosquatting* involves registration of DNS domains that are similar to the targeted popular domain or a brand name but contain small spelling errors that might be made to be hard to notice by users[39].

- *bitsquatting* relies on random bit flips occurring in computer memory due to hardware errors and, as such, registering acrshortdns domains with 1-bit difference from popular domains[40].

- *homograph squatting* uses visual similarity of different characters to register domains that are similar to targeted popular domains but are written with different symbols (such as replacing "l" with uppercase "i" or using international characters)[41]. (example, *googIe.com* instead of *google.com*)
- *combosquatting* adds additional words to the targeted domain name in a way that leaves the target brand intact [38]. (example *authentication-google.com*)
- *wrong Top-Level Domain (TLD)* domains register an identical second-level domain to the targeted one but under a different top-level domain. (example, *google.kz* instead of *google.com*) [42]
- *soundsquatting* relies on registering similar sounding (homophone) domains to the targeted domain name (for example, *utube.com* instead of *youtube.com*)[43].

Yet misuse of famous brand names and trademarks inside the Domain Name System is not limited to ill-intended registration of previously unregistered domains. Liu et al.[44] highlight abuse of subdomains of legitimate domains for malicious purposes, known as domain shadowing. In cases where the credentials of the domain registrant are compromised, malicious parties can access their account to register additional subdomains to leverage the reputation of the parent domain to avoid detection [44]. Cybercriminals have also been known to take over stale subdomains by taking over the resource subdomain pointing at [45].

As domain registrations expire, any interested party can re-register said domains on a first-come-first-serve basis. While re-registration of domains is usually done for monetisation and speculation purposes to leverage traffic still reaching them, it can also be used to subvert existing domain-based trust mechanisms. [46]

In addition to cybersquatting on previously registered domains, similar abuse vectors exist for phone numbers on mobile networks. Due to phone numbers being recycled by telecommunication companies, the trust put towards phone numbers being an identifying attribute can be subverted in various ways. For targeted attacks, fraudsters have been known to convince so-called sim-swapping attacks where the attacker convinces telecommunication providers to bind the victim's phone number to a SIM card under the attacker's control. The more wide-scale issue comes from the routine recycling of unused phone numbers by telecommunication service providers. If the previous owner had associated their number with any of their digital accounts, the new owner can authenticate themselves to their accounts.[47]

2.4.4 Spoofing

Spoofing attacks entail impersonating a legitimate system by falsifying information about its identity. In information security, spoofing is mostly associated with

network security topics such as WiFi, DNS and Address Resolution Protocol (ARP) spoofing attacks. However, depending on the targeted system and protocol, additional considerations should be taken into account for the other parts of the trust chain as well (e.g. Secure Sockets Layer and Transport Layer Security (SSL/TLS) certificates in case of Hypertext Transfer Protocol Secure (HTTPS)).[48]

In many cases, spoofing enables man-in-the-middle (MiTM) type of attacks, wherein malicious party covertly surveils or even proxies communications between two legitimate parties [48]. Well-known spoofing methods in the domain of network security are focused on the lower layers of the protocol stack, with ARP spoofing on the same segment, allowing for impersonation of another computer already on the same network[48] and wireless spoofing attacks mimicking legitimate access point on the physical layer[49]. While there are also application layer spoofing attacks such as Dynamic Host Configuration Protocol (DHCP) and DNS spoofing (and DNS cache poisoning), they tend to serve as a method for achieving MiTM situation at a limited scale[48].

However, end-goals for spoofing are not limited to eavesdropping on and tampering with legitimate communications. With the increasing popularity of biometric authentication (fingerprinting, facial, iris verification), they are becoming targets for spoofing attacks by malicious actors [50, 51, 52]. There is also a noted usage of location spoofing to break assumptions made by mobile and web applications to break security assumptions of location checks[53] or to mislead people on social media[54].

Potential for spoofing also exists in the context of phishing attacks, with email address spoofing highlighted as one of the dominant impersonation techniques used for phishing already 20 years ago[55]. SMTP (Simple Mail Transport Protocol) used for email transmission does not natively offer any protections against spoofing, instead relying on protocol extensions and optional cryptographic end-to-end protection measures for security[56]. Yet according to recent works, the adoption rate of security-focused protocol extensions such as Sender Policy Framework (SPF) and Domain-based Message Authentication, Reporting & Conformance (DMARC) is far from 100% [57]. As estimated by Czybik et al., the adoption rate for email security extensions can be around 13.6% (DMARC) and 60.2% (SPF), with the measurements based on the measurement of the top 1M most visited website domains[58].

For attackers that can execute network-level attacks against the victims, impersonating websites leveraging modern SSL/TLS encryption at the transport layer is more complicated than it was 20 years ago[59]. Modern browsers such as Mozilla Firefox[60] and Google Chrome[61] attempt to initiate the connection to a new website over an encrypted connection first before falling back to an unencrypted Hypertext Transfer Protocol (HTTP) connection. In addition, implementation

of HTTP Strict Transport Security (HSTS) headers by website owners helps to protect their sites against such impersonation attacks[62], doubly so if their domain has been submitted for HSTS preload lists built into browsers[63]. Yet neither transport layer encryption itself nor HSTS headers are nearing 100% coverage across all the public websites, remaining at 85% and 27% adoption [64, 65].

2.5 Summary

Due to the evident popularity of electronic identity theft, there is significant interest in its research by academia, government and industry stakeholders. Much of the academic research into impersonation of legal entities has been focused on specific types of attacks as described in Section 2.4. Studies have attempted to detect malicious infrastructure mimicking the branding of known legal entities. Such as Tian et al.[42] focusing on detecting targeted phishing websites based on the similarity of the Fully Qualified Domain Name (FQDN) to known brands.

The cybersecurity industry is also actively measuring and reporting the extent of different cyber intrusions, tools & techniques used, as well as the resulting losses for companies and individuals. Service providers release yearly and quarterly reports on data breaches, social engineering, and more.[16]

On the legislative side, there are EU initiatives to improve the current state of affairs on overall Cybersecurity, with the Directive (EU) 2022/2555 (widely known as the NIS2 Directive) set to take effect on 18th October 2024. The directive focuses on important sectors that maintain critical infrastructure, with a size threshold established for those subject to requirements. With this in mind, the current level of resilience of legal entities to intrusions (including but not limited to electronic identity theft) is expected to increase.[66]

However, despite significant work on the topic, the level of identity protection maintained by active organisations remains unknown. Much of the current research tends to focus on deep exploration of specific technical vulnerabilities (as seen in Section 2.4) or tracking down the number of impersonators of popular brands[42]. Yet, there does not seem to be much work done to measure the vulnerability surface of organisations and the effect that available resources have on the outcomes.

3 Research problem

This thesis aims to bridge the gap between technical research focusing on finding vulnerabilities and providing actionable context around them. As such, vulnerability to common electronic identity theft is measured across many organisations. The chosen approach provides additional insights into existing gaps between industry recommendations, best practices and the situation faced by organisations that might find themselves on the receiving end of electronic identity theft.

All the findings relate to registered legal entities in Estonia, with publically available contact details, allowing for actionable findings to be reported directly or through the local authorities. Improving cyber hygiene amongst the sample population is an additional benefit.

To reach the stated goals, this thesis seeks answers to the following questions:

- **PRQ**: To what extent are organisations of different sizes vulnerable to identity theft?
  (a) How does organisations' adoption rate of digital identity protection measures compare to global statistics?
  (b) How much do the technical corporate identity protection measures depend on the company size?
- **SRQ**: To what extent have digital identities or active organisations been compromised?

The research setup used for finding the answers to the stated questions is described in detail within Section 4, and the results themselves are presented in Section 5 of the thesis.

4 Research design

A case study was chosen as the research method to answer the proposed research questions. Due to the relatively high level of digitization and favourable legislation[2], organisations registered in the Republic of Estonia were chosen as the subjects for the case study.

4.1 Constraints and assumptions

The primary objective of the research project was to quantify the extent to which companies are vulnerable to identity theft. To fulfil that objective, the following constraints were used for scoping the study itself:

- *Need for a sufficiently large sample size* - To generalise the findings, the sample size must represent a large percentage of the total population.
- *Work done must be reproducible and measurements should be automated* - Data used for the study must be publicly available to ensure transparency and repeatability.
- *Selected samples should represent active organisations* - Inclusion of inactive companies is assumed to bias the survey population towards more vulnerable findings. If the organisation is inactive, it is not expected to put much effort towards protecting its digital identity.

4.2 Ethical considerations

Data relating to weaknesses in impersonation attacks is sensitive information. Even though the study aims are benign, malicious actors might abuse publicised vulnerabilities. As such, the study's ethical aspects were weighted carefully to avoid causing collateral damage to the surveyed organisations.

As a result, the following principles were defined and followed throughout the work:

1. *Notify national cyber security authority of the relevant findings.* - Actionable findings from the study should be shared with the Estonian Computer Security Incident Response Team (CERT-EE).
2. *Do not reveal the identities of survey population members.* - To restrict the potential misuse of the thesis results, the identities of surveyed companies should not be disclosed in this work.

[2]Commercial Register Act[18] makes contact details and revenue of companies available as public information through the e-Business Register.

3. *Notify relevant authorities on queries with potentially detrimental impact to the target system* - As collection of data for vulnerability measurements can cause considerable strain on queried services or be misidentified as an attack, maintainers of queried services should be notified beforehand.

4. *Prefer non-intrusive methods for indicators of vulnerability and compromise* - Due to the sensitive nature of security vulnerabilities in information systems, all indicators for vulnerability measurements should be considered non-intrusive. None of the digital assets subjected to the study should be attempted to be compromised or impersonated.

Following these principles constitutes *due diligence* on the author's behalf to minimize the potential negative impact of this work. While these measures might not prevent the future misuse of the study results, the benefits are expected to outweigh the negatives.

4.3 Data selection

Open data from the e-Business Register of the Republic of Estonia[4] was selected as the primary data source due to the high quantity of publicly available relevant data. General Data Records for registered companies were downloaded along with supplemental data concerning the key indicators on the annual reports for the most recent fiscal year at the time of the writing (2022).

4.3.1 Identifying digital assets from the data source

According to § 26 of the Commercial Register Act of The Republic of Estonia[18], some digital communication channels used by registered companies are made publically available[18]. Table 1 lists communication channels as they are listed in the registry (Data retrieved on 4th November 2023) [4].

Table 1. Digital communication channels for companies listed in the e-Business Register.

Listed digital communication channel	**Number of companies (% total)**
Email address	348,375 (99.2%)
Mobile	207,560 (59.1%)
Landline	57,530 (16.3%)
Website	22,656 (6.4%)
Fax	13,685 (3.9%)
Other	899 (0.2%)

Based on the vulnerabilities highlighted in the previous works (see Section 2.4 and the relatively low adoption rate of some of the listed communication channels, the following three types of digital assets were chosen to be included in the study:

- Email address
- Mobile[3]
- Website

4.3.2 Selecting the dataset

Of the 358,509 registered organisations (as of 4th of November 2023) with general data records, 193,133 companies had submitted their financial reports for the fiscal year 2022. Amongst these, only companies that had listed either a mobile phone number, an e-mail address or a website were included in the study (see Figure 2.).

[3]While a significant proportion of companies also have a landline number listed, the value of including them in the study was not considered worth a significant increase in the scope that would have been required to do so.

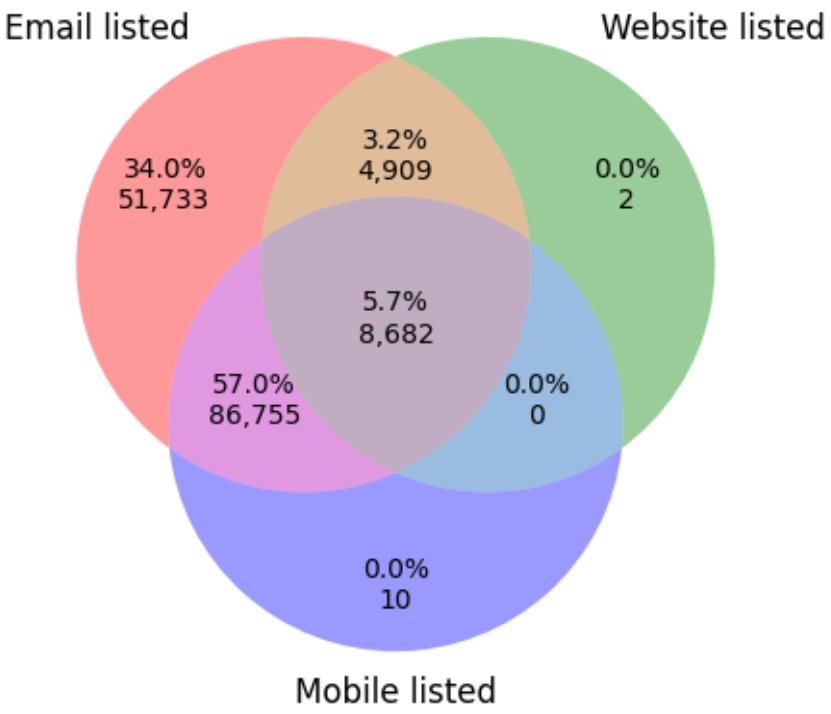


Figure 2. Distribution of types of digital assets listed in the e-Business Register for companies that had submitted their fiscal reports for 2022.

This process yielded a total of 152,091 judicial bodies representing over 42% of organisations included in the e-Business Register and over 78% of the active companies[4]. To generalize the findings of the results based on the Small and medium-sized enterprise (SME) classification recommendations provided by the European Union in the Recommendation 2003/361[67]. The distribution of the companies based on the classification can be seen in Table 8.

[4]Active company in this context meaning company with a submitted report for the fiscal year of 2022.

Table 2. Distribution of the companies in the dataset based on the size classification, as recommended by European Union Recommendation 2003/361[67].

Classification	**Definition**	**Proportion (count)**
micro-enterprise	Number of employees <10, revenue and annual balance sheet total < €2 m	93.8% (142,666)
small enterprise	Number of employees <50, revenue and annual balance sheet total < €10 m	4.8% (7,243)
medium enterprise	Number of employees <250, revenue < €50 m and annual balance sheet total < €43 m	1.1% (1,695)
large enterprise	employees > 250, revenue > €50 m or annual balance sheet total >= €43 m	0.3% (487)

4.4 Selecting indicators

To cover vulnerability estimates against common methods for identity theft described in Section 2.4, two classes of indicators were defined:

- *Vulnerability indicators (VULN)* - Indicators for measuring an organisation's digital assets' vulnerability to impersonation.
- *Indicators of compromise (IOC)* - Indicators for measuring whether the digital assets have already been compromised.

After the base types for the measurement indicators had been specified, concrete indicators had to be chosen to cover common techniques covered in Section 2.4. Indicators were selected based on the constraints and assumptions stated in Sections 4.1 and 4.2.

4.4.1 Vulnerability indicators

Selecting indicators for detecting vulnerabilities related to *mobile phone numbers* was relatively straightforward. The weaknesses of mobile networks to spoofing are that they are not straightforward to abuse without specialized equipment, and phishing attacks against telecommunications providers to measure vulnerability to sim-swapping would be neither scalable nor ethical. However, the Republic of Estonia's Consumer Protection and Technical Regulatory Authority of the Republic of Estonia provides a public service for checking phone number availability[68]. It

was considered practical to run Estonian phone numbers in the registry through the service to see how many numbers attributed to a judicial body could be registered by squatters. This yielded the following indicator:

- **VULN M.1** Mobile number is available for re-registration.

There are already well-known mechanisms used to detect vulnerabilities relating to email addresses to combat unsolicited email, phishing, and spoofing attempts. Some examples of these are email reputation services and email protocol security extensions, namely SPF and DMARC[5]. In addition, based on the preliminary data exploration, many email addresses listed in the registry seemed to be hosted by public email service providers (Such as Google's Gmail, Microsoft's Outlook and VK Group's Mail.Ru). Intuitively, such addresses seem to entail a greater risk of account compromise due to the greater risk of cross-usage for private and business purposes. This process yielded the following vulnerability indicators for checking email infrastructure:

- **VULN E.1** Email domain is not registered.
- **VULN E.2.1** Private email address is used for business purposes.
- **VULN E.2.2** Recommended email security extension (SPF and/or DMARC) is not used.

A domain registration check and two new indicators were chosen to measure the vulnerability of company websites. As man-in-the-middle attacks against websites depend on misconfigured Secure Sockets Layer and Transport Layer Security (SSL/TLS), the SSL/TLS configuration of the website is a good indicator of the website's protection against network-level spoofing attacks. This yielded the following indicators for checking the vulnerabilities of websites:

- **VULN W.1** Website domain is not registered.
- **VULN W.2** SSL/TLS is not supported by the website.
- **VULN W.3** HSTS is not supported by the website.

It should be noted that the indicators defined here do not have the same level of importance. Figure 3 illustrates the hierarchy of the selected indicators. An unregistered domain allows hackers to register the domain for themselves and redirect any resulting traffic however they want. In addition, unregistered domains would not be expected to point to any servers, and HSTS headers would be ignored over plain-text HTTP connection.

[5]Third email security extension, DomainKeys Identified Mail (DKIM), was not included in the indicators due to difficulty of obtaining necessary selectors for verifying existence of DKIM all the companies included in the study.

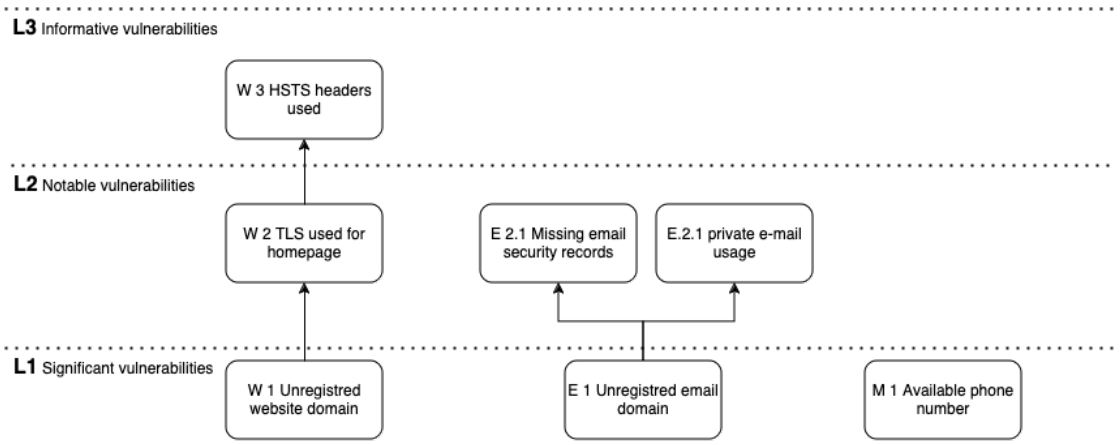


Figure 3. Dependency graph describing the relationships between different vulnerability indicators. Arrows represent the preconditions on which vulnerability checks must succeed to check for the next level of vulnerabilities.

4.4.2 Indicators of Compromise

Indicators of compromise were chosen to collect metrics indicating with reasonable certainty that the digital asset has already been compromised. As was the case with vulnerability indicators, indicators of compromise were selected per each digital asset included in the study, except for mobile numbers for which indicators of compromise were defined[6].

Indicators for compromised email addresses were relatively easy to select. In case an email address has been included in data leaks containing credential-related sensitive information, the email will be considered potentially compromised:

- **IOC E.1** Credentials associated with the organisation's listed email address have been published as a result of a leak.

There are numerous techniques and vulnerabilities through which websites can be compromised. However, as the focus of this work is on impersonation, detecting cases of cybersquatting was considered to be the most appropriate. However, in many cases, squatting detection would involve parsing the site's contents for signs of domain parking or trademark abuse. Building these types of checks required disproportionately more resources than those needed for other indicators. Due to

[6]Due to the nature of mobile communications and relatively scarcely available ownership information of the mobile telephone numbers, the author was unable to find any scalable and non-intrusive method for verifying if the current owner of the phone number is its original owner or not.

these concerns, only a single predictor for detecting cybersquatting was included in this work:

- **IOC W.1** Website redirects visitors to an unrelated website (as a form of affiliate squatting).

4.5 Instrumentation

An experiment was set up to measure the vulnerability of the survey population via selected indicators. Jupyter notebook[70] with the Python language utilising Pandas[71] library was chosen as the core tool.

4.5.1 Preprocessing steps

As the general data on registered organisations came in JavaScript Object Notation (JSON) format, Bash script was written to extract the necessary parameters (company name, registry code, website, email, phone number)[7] from the file and output them into Pandas-friendly file in Comma-Separated Values (CSV) format.

As the data about financial reports was already formatted as CSV files, reformatting was unnecessary. However, to ease the debugging of the analysis pipeline, unnecessary fields were removed from them, leaving only information on report IDs, registry codes, annual balance sheet total, revenue, and employee count to be included.

After extracting the necessary elements, all the resulting CSV files were ingested into Pandas DataFrame, where financial report data was merged with the general data. Additional processing steps were undertaken to extract domain information from the email addresses and website URL to measure the vulnerability indicators that rely on checking the DNS records. The following steps were taken to prepare data for vulnerability analysis:

1. Remove protocol-specific information
    - Protocol prefixes (e.g http://, https://) and URL path related information
    - Anything before @ symbol for email address
2. Breaking extracted domain into top-level-domain and second-level-domain[8]

[7]It should be noted that for companies that had multiple phone numbers, email addresses or websites listed, only the first match for each of these parameters was included in the study.

[8]The technique used for it turned out to be rather naive, as not all top-level domains are entirely on the right side from the last "." symbol (see Section 6 for more information).

After this process was concluded, extracted top- and second-layer domain components were concatenated to get a registered domain, which was used for checking for vulnerabilities.

4.5.2 Measuring significant (L1) vulnerabilities

While both indicators were defined based on existing works, multiple approaches can be used to measure many of them. For the context of this thesis, the tradeoffs were guided by a preference for low complexity and resource usage and high speed for carrying out experiments. The chosen approach allowed more resources to be put towards analysing results and limited the time spent building and fine-tuning the toolkit.

The toolkit for measuring the vulnerability of surveyed organisations was written into a single Jupyter notebook using the Python programming language with additional libraries used by the author to reduce manual labour.

A relatively straightforward process was chosen to measure significant vulnerabilities (L1). VULN W.1 and VULN E.1 measurements relied on measuring domain availability for the registered domain, for which using the WHOIS protocol would be preferable. However, it is restricted to a limited number of queries per day[69]. Due to this, relying on it to query tens of thousands of domains was considered infeasible. Since the protocol specification[70] requires Start of Authority (SOA) records to exist for all DNS zones, querying for the SOA record was relied upon instead. WHOIS queries were only done to validate the registration status of domains that didn't return an SOA record.

For mobile number availability, all the phone numbers were checked against the number availability service provided by the Estonian Consumer Protection And Technical Regulatory Authority (as mentioned in Section 4.4.1). As some of the phone numbers listed in the registry are registered outside Estonia, these were omitted from the availability checks due to the complexity of finding suitable authorities for checking all of them. A comprehensive summary of the significant (L1) vulnerability measurement methods is listed in Table 3.

Table 3. Measurement methods chosen to measure significant (L1) vulnerability indicators.

Indicator	**Measurement method**	**Python packages**
VULN W.1 & VULN E.1	1. Run a DNS query against the second-level domain to check for the existence of an SOA record for the domain. 2. Check the registration status of domains that failed the first step via WHOIS protocol	*pydig*, *python-whois*
VULN M.1	Check the phone numbers against number availability checking service[68].	*beautifulsoup4*

4.5.3 Measuring notable (L2) and informative vulnerabilities (L3)

Detection of notable vulnerability VULN W.2 for websites relies on attempting an SSL/TLS handshake on the web server's 443/TCP port. If the handshake is successful and the certificate offered by the web server is trusted (e.g., it is not expired nor self-signed), then the website passes the check.

Information queried via DNS is enough for the second-level vulnerabilities of listed email addresses. For checking the usage of email security extensions (VULN E.2.2), only the presence of DNS records was checked. As such, email domains were queried for DMARC and SPF records, but the documents themselves were not validated against the schema or checked for vulnerable configuration.

Measuring the cross-usage of email accounts for personal/business purposes (VULN E.2.1) is difficult to measure directly. To bypass this limitation, proxy measurement of common email provider usage was chosen instead. For this, mail domains that occurred amongst email addresses of at least 100 organisations were extracted, and their background and usage were investigated further to highlight possible risks. A summary of the methods chosen for measuring notable vulnerability indicators can be seen in Table 4.

Table 4. Measurement methods chosen to measure notable (L2) vulnerability indicators.

Indicator	**Measurement method**	**Python packages**
VULN W.2	1. Attempt to establish SSL/TLS connection on a well-known 443/TCP port with the webserver that the listed website URL points to. 2. Check if the certificate offered by the server is valid.	*sslyze*
VULN E.2.1	Extract popular email domains used by a large number (at least 100) of the surveyed organisations.	-
VULN E.2.2	Check the email domain for the existence of TXT records containing DMARC and SPF information.	*pydig*

The measurement methodology for the only informative vulnerability (VULN W.3) relied on checking for the presence of HSTS headers for the websites that passed the preceding VULN W.2 check (see Table 5).

Table 5. Measurement methods chosen to measure informative (L3) vulnerability indicators.

Indicator	**Measurement method**	**Python packages**
VULN W.3	Check for the existence of HTTP Strict Transport Security headers for websites that support SSL/TLS connections.	*sslyze*

4.5.4 Measuring indicators of compromise

To find an answer to several potentially compromised corporate email addresses, the listed surveyed companies were queried through the API of the *Have I Been Pwned* website[9]. The website hosts an extensive public collection of breached user account data and information on the data type disclosed due to these breaches. While most of the information from these breaches can aid in impersonation, not everything can be used to take over the associated account itself.[71]

[9]https://haveibeenpwned.com/

Since this analysis focuses on measuring the vulnerability to online business identity theft and the extent of compromised communication channels, an additional analysis step was considered necessary to rule the email address as potentially compromised with reasonable certainty. To achieve this, only companies affected by breaches that included secrets (as defined in Table 6) were chosen for the subsequent analysis step.

Table 6. Selected data types used for filtering more impactful data leaks for further analysis.

Classification	**Data type**	**Explanation**
Leaked secrets	Passwords Historical Passwords PINs	High risk of password reuse.

To measure if any of the websites had been compromised for affiliate scam (IOC W.1), following the HTTP redirection chain to a predefined depth (five redirects) and manually checking for any patterns in the resulting URLs that would indicate maliciousness (such as lengthy or reoccurring URLs). See Table 7 for a summary of the measurement strategy for indicators of compromise.

Table 7. Measurement methods are chosen to measure indicators of compromise.

Indicator	**Measurement method**
IOC E.1	1. Query email addresses through "Have I Been Pwned" API[71] 2. Check matched leaks and filter the matches down to leaks containing sensitive information.
IOC W.1	1. Follow HTTP redirection chain to a reasonable depth (5) 2. Manually review patterns in redirect URLs for signs of maliciousness.

An overview of the measurement results and their interpretation and analysis is given in the next section.

5 Results

This section gives an overview of the survey results as defined in Section 4. In addition to a numeric summary of the findings, additional analysis is provided into possible causes and risk patterns identified during the study.

5.1 Vulnerability measurements

Experimentation was carried out to measure the chosen indicators in order of severity based on the digital asset targeted. The website vulnerability measurements from VULN W.1 through VULN W.3 were sequential, with only organisations passing the first tests included in the subsequent ones. For email measurements, VULN E.1 and VULN E.2.2 are sequential, with VULN E.2.1 including all included email addresses. For mobile numbers, only one indicator was defined. The findings from these measurements are shown and analysed in this subsection.

5.1.1 VULN W.1 Detecting unregistered website domains

The measurements for detecting unregistered website domains were conducted on the 20th of November, 2023. During the process, it was discovered that among 13,593 websites included in the study, 1% (145) addresses were instead email addresses mistakenly entered into the website field on the registry. As a result, these organisations were removed from the study population used for website vulnerability checks.

By measuring the website domain registration status for the remaining 13,448 companies that had their website information listed in the dataset, it was discovered that up to 8% (1060) of the domains for the listed websites could be available for re-registration (based on the SOA record checks).

But as a missing SOA record does not necessarily mean that the domain is available for re-registration, follow-up checks were done through WHOIS queries. This provided additional granularity for further analysis. The results of the WHOIS query and their interpretation in the context of domain registration status can be seen in Table 8.

Table 8. Results of the website domain registration status check

Vulnerable to squatting	**Response to WHOIS**	**Number of Domains (%)**
VULNERABLE	N/A (No Match)	805 (6%)
	Available	2 (<1%)
	Expired	21 (<1%)
	AtAuction	1 (<1%)
	Exception (No Match)	118 (1%)
	Total:	**947 (7%)**
NOT VULNERABLE	Ok	113 (1%)
	SOA record present	12,388 (92%)
	Total:	**12,501 (93%)**

The numeric data alone does not offer sufficient insights into the data to help answer the research questions. A manual review of sampled organisations from each size classification was done to gather and analyse the nature and possible causes that made these vulnerabilities manifest (see Figure 4 for distribution).

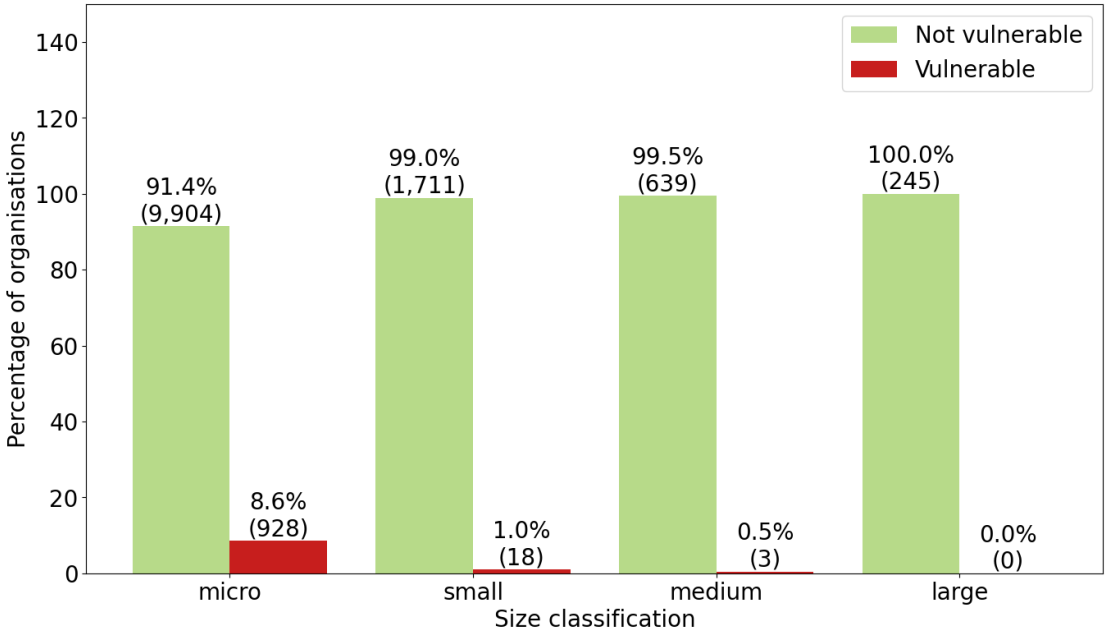


Figure 4. Percentage of vulnerable companies per size classification based on the findings for the indicator - Website domain is not registered (VULN W.1). Absolute numbers for vulnerability status of surveyed organisations can be seen in parentheses.

A deep dive into the vulnerable medium-sized companies shows that all three have a different domain used for their official email address. None of the email domains seems to be susceptible to re-registration. It should be noted that for 66% of them, the email domain is used for hosting the company website. For one particular organisation, the issue with the faulty domain seemed to have been caused by a typo in the website URL entered into the registry (the character "i" was substituted by "l").

The proportion of smaller companies shown as vulnerable according to the chosen methodology is consistent with the findings for medium-sized companies. A total of 18 small companies with vulnerabilities were detected, with 44% (8) being identified as expired domains of previously functional websites[10]. 50% (9) seemed to have a typo in the website name, with the website seeming to have never existed (In one case, the email address had been entered into the website field with the "#" symbol used instead of "@"). The remaining smaller organisations had two websites entered into a single website field, which the automation script couldn't parse into a legitimate domain name. After manual checking, both domains were deemed registered, but one of the domains seems to have changed ownership.

For micro-sized companies, the proportion of vulnerable companies is noticeably higher, with 10% of surveyed companies' websites being considered susceptible to domain squatting. As the manual review of all the vulnerable websites was deemed unfeasible to the scope of this work, ten vulnerable samples with the highest reported revenue were chosen as a sample for deep dive. The distribution for probable causes for vulnerability seems to follow the distribution also seen for small and medium counterparts with 30% (3) false positives, 30% (3) websites seemingly never existing and 40% (4) having existed but have expired ever since.

The results from the analysis of vulnerability surface for the surveyed companies show that larger companies seem to put more effort into protecting their digital assets. None of the large companies had let their website domain expire. The proportion of vulnerable sites appears to be similar when comparing small and medium-sized companies, with the proportion of vulnerable companies increasing tenfold for micro companies compared to small and medium-sized ones. This could be due to micro-organisations letting their website domain expire if it is not considered to impact the revenue for the organisation positively.

5.1.2 VULN W.2 Missing or broken transport layer encryption

Relevant data was collected on the 20th of April 2024 and yielded a total of 7% (881) of the organisations with SSL/TLS-related problems[11]. The distribution per

[10]Checked via https://archive.org/

[11]Only 12,388 organisations that had passed the SOA check were checked for SSL/TLS configuration issues

company size does give insights that the extent of vulnerable Secure Sockets Layer and Transport Layer Security configurations is more common than domain-related issues described previously (see Figure 5).

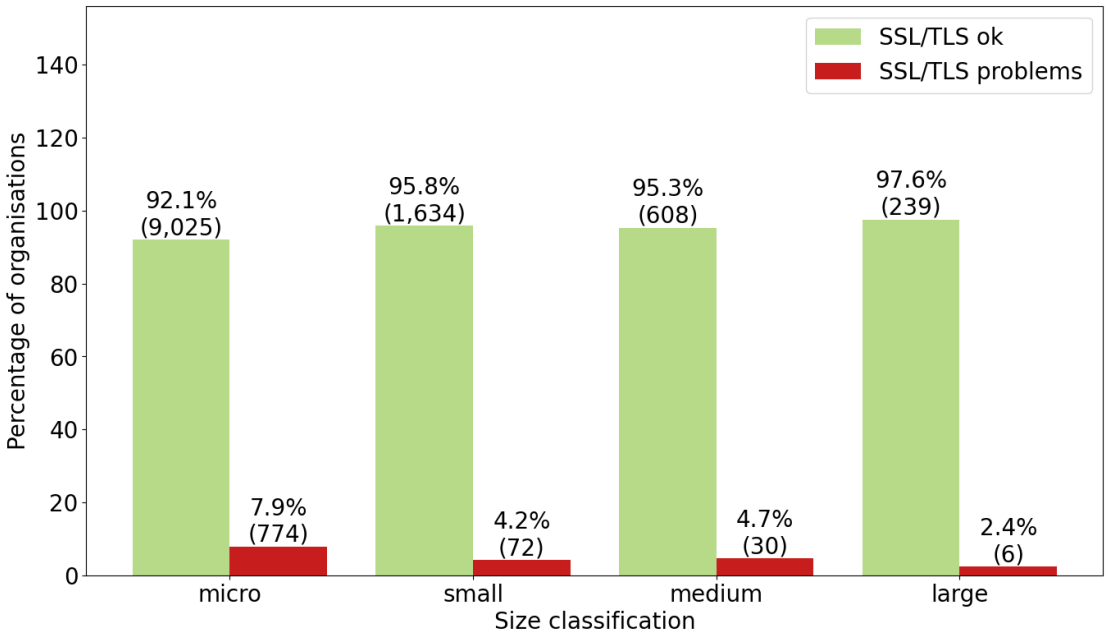


Figure 5. Percentage of vulnerable companies per size classification based on the findings for the indicator - SSL/TLS is not supported by the website (VULN W.2). Absolute numbers for vulnerability status of surveyed organisations can be seen in parentheses.

A deeper inspection of the discovered SSL/TLS usage-related issues highlights temporary issues with the SSL/TLS certificate renewal process and vulnerable redirections after probable branding-related changes. In addition, exceptions in measuring SSL/TLS support seem to have been caused by A or CNAME records not existing for the website domain as listed in the registry. Table 9 gives an overview of the related issues found during the testing.

Table 9. Distribution of measurement results for the indicator -SSL/TLS is not supported by the website (VULN W.2)

Vulnerable to squatting	**SSL/TLS configuration issue**	**Number of Domains (%)**
VULN W.1 (SOA fail)	Not measured	1060 (8%)
SSL/TLS related issues	N/A (No SSL/TLS connection)	391 (3%)
	Exception	258 (2%)
	certificate not trusted	195 (1.4%)
	No certificate information	38 (<1%)
	Total:	**882 (6.5%)**
SSL/TLS ok	**Total:**	**11,506 (85.5%)**

A deeper dive based on the company sizes revealed that out of the six larger companies flagged as vulnerable during the first test, 16% (1) had issues with expired certificates. Among the rest, 66% (4) websites negotiated for NULL cipher, and one had no server listening on the HTTPS port (443/TCP). Of the problematic medium-sized companies, 10 with the highest revenue were selected for manual review. Among these companies, 50% (5) did not have a listener on the HTTP port, 30% (3) used untrusted certificates (2 were expired, one was missing listed domain from certificate CN and SAN fields), and 20% (2) did not have a website on listed domain.

Organisations classified as small- and micro-sized were investigated similarly, with ten samples with related issues and the highest revenue selected for manual investigation. Among them, 40% (4) does not have a listener on the HTTP port, 30% (3) have certificate trust issues (2 websites did not have a valid certificate for the listed domains, and 1 offers an incomplete certificate chain), and another 30% (3) do not have a website on listed domain. For The manually reviewed micro-sized organisations, 20% (2) did not have a server on the HTTP port, and 40% offered null cipher and no valid server certificate. 20% (2) did not have a website on the listed domain, and 20% (2) had a misconfigured server offering an incomplete certificate chain.

The findings amongst the SSL/TLS misconfigurations seem similar across all four size classifications, with null cipher offerings being a surprisingly common type of misconfiguration amongst expired certificates and deprecated websites.

5.1.3 VULN W.3 Missing HSTS support

From amongst the 11.506 companies that passed SSL/TLS configurations checks in Section 5.1.2, only 18.7% (2,147) had HSTS headers configured according to measurements done on 20th April 2024, the rest of the 81.3% (9,359) of the companies that passed the SSL/TLS configuration check, did not have HSTS headers configured. As Figure 6 shows, most surveyed companies do not implement HSTS headers to protect their websites.

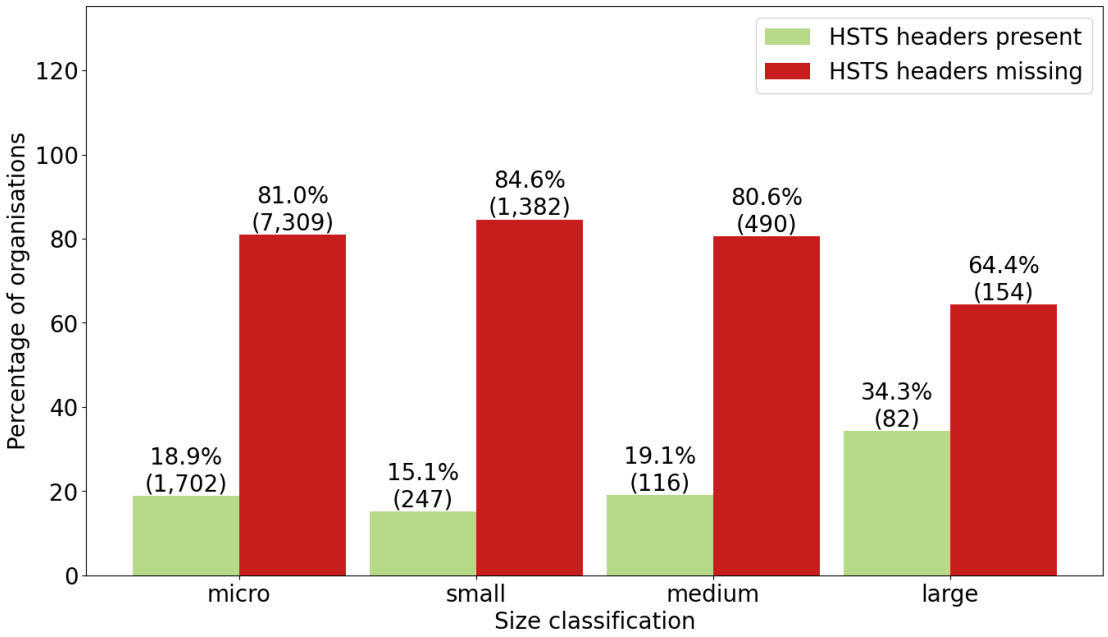


Figure 6. Percentage of vulnerable companies per size classification based on the findings for the indicator - HSTS is not supported by the website (VULN W.3). Absolute numbers for the vulnerability status of surveyed organisations can be seen in parentheses.

Only large organisations preceded the global adoption rate of 27% mentioned in 2.4.4, with the 30% adoption rate of HSTS headers. The rest of the categories fall behind, with the adoption rate for small- to medium-sized companies hovering around 15-20% adoption rate.

5.1.4 VULN E.1 Detecting unregistered email domains

Measurements for identifying unregistered email domains were conducted on 30th January 2024 (Step 1.) and 11th May (Step 2.). Initial detection relying on SOA

record checks yielded 0.8% (1,319) potentially vulnerable companies out of 152,079 with email addresses listed.

Table 10. Results of the email domain registration status check

Vulnerable to squatting	**Response to WHOIS**	**Number of Domains (%)**
VULNERABLE	N/A (No Match)	876 (<1%)
	Available	15 (<1%)
	Expired	0 (<1%)
	AtAuction	0 (0%)
	Exception (No Match)	161 (<1%)
	Total:	**1052 (0.7%)**
NOT VULNERABLE	Ok	267 (<1%)
	SOA record present	150,760 (99.2%)
	Total:	**151,027 (99.3%)**

The WHOIS status of the domains found vulnerable per SOA query can be seen in Table 10. In general, most companies are not susceptible to email domain takeover. Most vulnerable companies are on the smaller side of the spectrum, consistent with similar findings for the website domains analysed in the previous section (See Figure 7 for the distribution by company size).

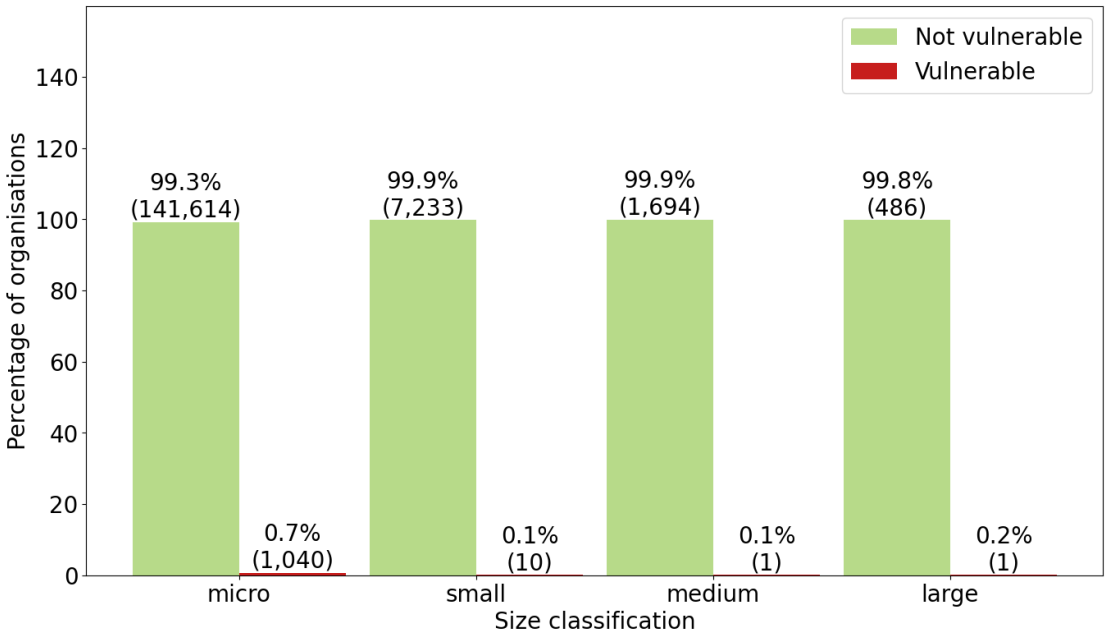


Figure 7. Percentage of vulnerable companies per size classification based on the findings for the indicator - Email domain is not registered (VULN E.1). Absolute numbers for the vulnerability status of surveyed organisations can be seen in parentheses.

There is one large organisation whose domain seems to be available for registration. However, the company appears inactive, with no reported employees or revenue flow for the past few years, even though its total assets exceed the EU threshold to be considered a large enterprise.

The single medium-sized company with a vulnerable domain seemed connected to the large one, as the email domain was the same. The remaining three medium-sized companies were listed among vulnerable organisations due to the same parsing error that caused problems for the website checks.

All ten vulnerable companies were confirmed true positives and registrable for the small companies. For the potentially vulnerable 1040 micro-organisations, a deeper dive was done on the ten companies with higher revenue for the last fiscal year. Amongst them, one of the organisations had input a short text after the email, which broke the measuring script, expressing their wish not to receive any automated emails. The rest of the nine micro-organisations had let their email domain expire; however, the reasons why these email domains are no longer used could not be identified.

It is worth mentioning that most surveyed companies appear to be using well-known email service provider domains, such as *gmail.com* or *outlook.com*, instead

of their own registered domain. Using a well-known provider makes it unlikely for these companies to be vulnerable to subdomain takeover or technical weaknesses concerning email security-related DNS misconfiguration. However, they might be more susceptible to credential leaks (in case a personal email account is cross-used for business purposes) or phishing (as it is relatively easy to acquire a similarly named free email account with a known provider). Such concerns will be further explored in Sections 5.1.5, 5.2.1.

5.1.5 VULN E.2.1 Usage of a private email address for business purposes

Over half of the organisations in the data set utilised public mail service providers, relying on the free default option of using the provider-owned email domain. This section gives a more in-depth overview of the usage of typical mail providers amongst the surveyed organisations.

Of 152,079 companies in the dataset, only 20% (30,698) had a unique email domain. The rest of the 80% (121,381) used 6,552 domains, with 46% (69,733) of the surveyed organisations being hosted at Gmail. The rest of the common mail providers are much less popular with hot.ee domain ranking as second most popular covering 5% of the organisations(see Figure 8 for reference).

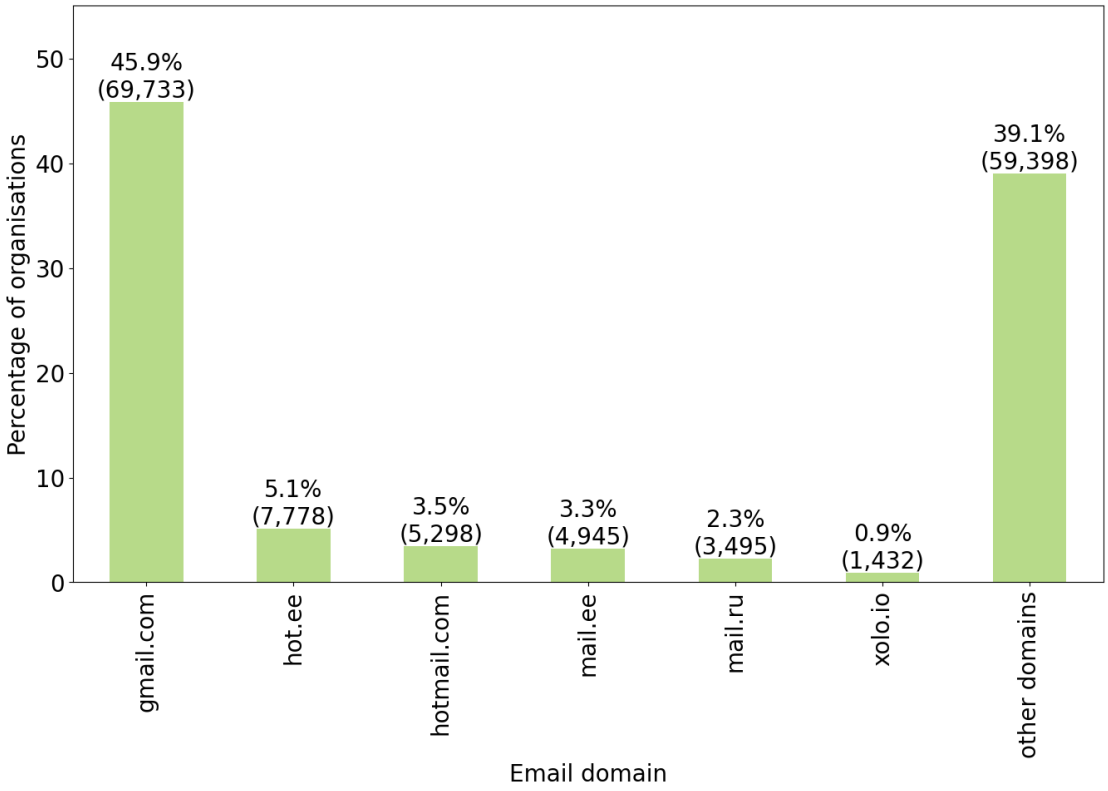


Figure 8. Proportion of six most popular email domains used by surveyed organisations compared to other email domains. Absolute numbers for organisations relying on listed domains can be seen in parentheses.

A deeper dive into usage of common mail provider accounts was limited to 26 email domains that appeared in the email addresses of more than 100 organisations included in the survey. These mail domains are used by more than 65% (99,560) of the surveyed companies with their email addresses listed in the registry. These 25 Email domains were manually reviewed and grouped into six categories as shown in Table 11.

Table 11. Groupings of popular email domains

Category	**Company/Description (Domain(s)**	**Number of companies (%)**
US mail providers	Google (gmail.com)	69,733 (46%)
	Microsoft (hotmail.com, outlook.com, msn.com, live.com)	6,439 (4.2%)
	Yahoo (yahoo.com)	787 (<1%)
	Mail.com	125 (<1%)
	Total:	**77,084 (50.7%)**
Russian mail providers	mail.ru	3,495 (2.3%)
	list.ru	399 (<1%)
	bk.ru	319 (<1%)
	yandex.ru	279 (<1%)
	inbox.ru	223 (<1%)
	rambler.ru	175 (<1%)
	Total:	**4,890 (3.2%)**
Online.ee	hot.ee	7,778 (5.1%)
	online.ee	595 (0.4%)
	neti.ee	400 (0.3%)
	Total:	**8,773 (5.8%)**
Inbokss Ltd	mail.ee	4,945 (3.3%)
	inbox.lv	108 (0.1%)
	Total:	**5053 (3.3%)**
Misc	National email service(eesti.ee)	1032 (0.7%)
	Business Administration Software (xolo.io)	1,432 (0.9%)
	Local telecommunications company (tt.ee)	196 (<1%)
	Shelf company provider (wasp.ee)	232 (<1%)
	Proton AG (protonmail.com)	234 (<1%)
	Total:	**1837** (1.2%)

Based on the categories, potential security risks become apparent. Support for Estonian National email service eesti.ee email formats *business-name@eesti.ee*, *firstname.lastname@eesti.ee* first ended on 1st November 2023, with support remaining for *registry-code@eesti.ee* for businesses and *personal-code@eesti.ee* for physical persons. Only 36% (374) of the listed *eesti.ee* emails are still supported; the remaining 56% (658) use the unsupported format.

The use of email services hosted outside the European Union for business

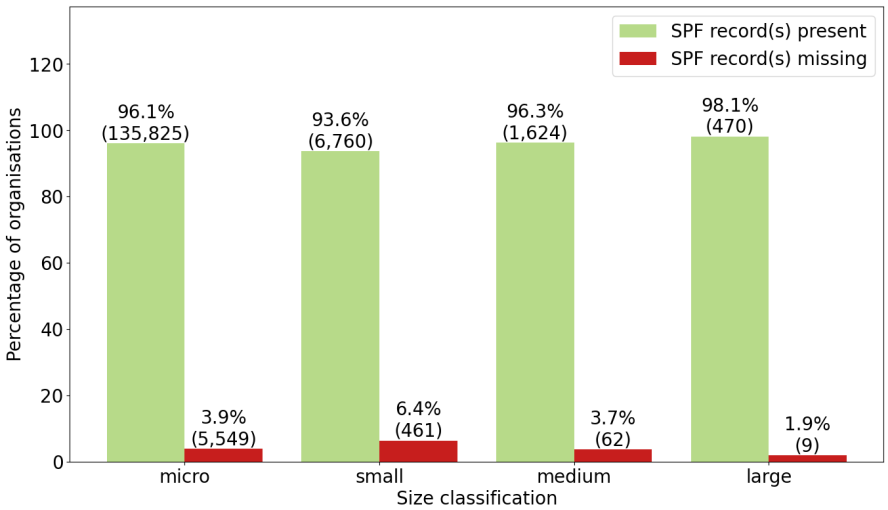
communications, such as email services hosted in Russia (or, to an extent, the USA), might cause liabilities in the context of GDPR, with *mail.ru* explicitly limiting the GDPR-mandated protections available to its user based on the local laws of its country of operations[72].

In addition, Services under Online.ee umbrella was sold by Telia Estonia to Fjordmail Technologies AS on 8th August 2023, after which the mail service became a paid service[73]. Due to such change, it is possible that many of the previous Online. E-mail service users are migrating off the platform. While it does not seem likely to register new accounts on the webpage, if previous customers' addresses are not locked for registration, it might open up new possibilities for impersonation.

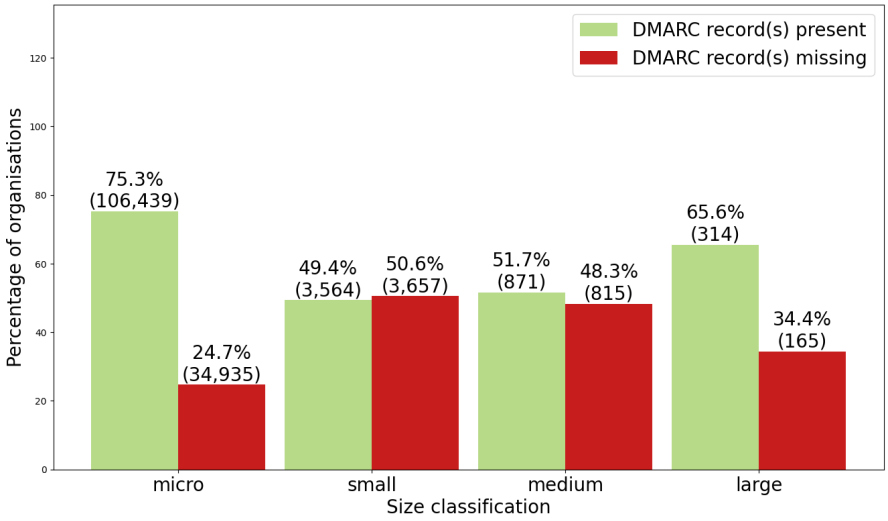
The size distribution of companies that use the common mail provider domain includes companies from all four size classifications. With 2% (9), 7% (120), 26% (1,870) and 68% (97,561) for large, medium, small and micro companies using mail services from amongst providers listed in Table 11.

5.1.6 VULN E.2.2 Missing SPF & DMARC records

Data on the configuration of SPF and DMARC records were collected on the 10th of April 2024 from amongst 150,760 organisations that passed the SOA step of VULN E.1. In general, the adoption rate of SPF records for maintaining email security seems to be adopted by around 95% of the companies in all surveyed size categories. DMARC adoption does not offer a similarly uniform picture, with around half of the small and medium-sized companies having adopted DMARC. Adoption rates for micro- and large-sized companies were respectively 75% and 66% of the companies with registered email domains. Results for these checks can be seen in Figure 9a (SPF) and Figure 9b (DMARC).



(a) Results of the SPF checks component of the VULN E.2.2 measurements.

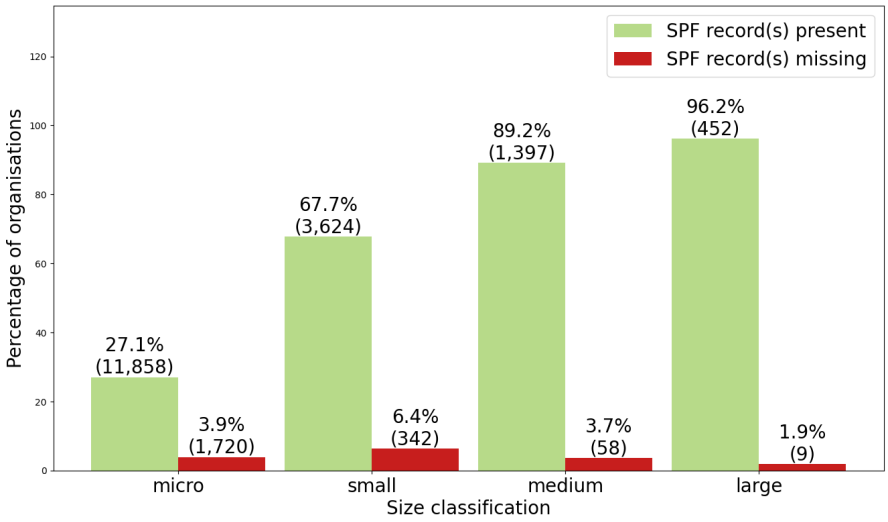


(b) Results of the DMARC checks component of the VULN E.2.2 measurements.

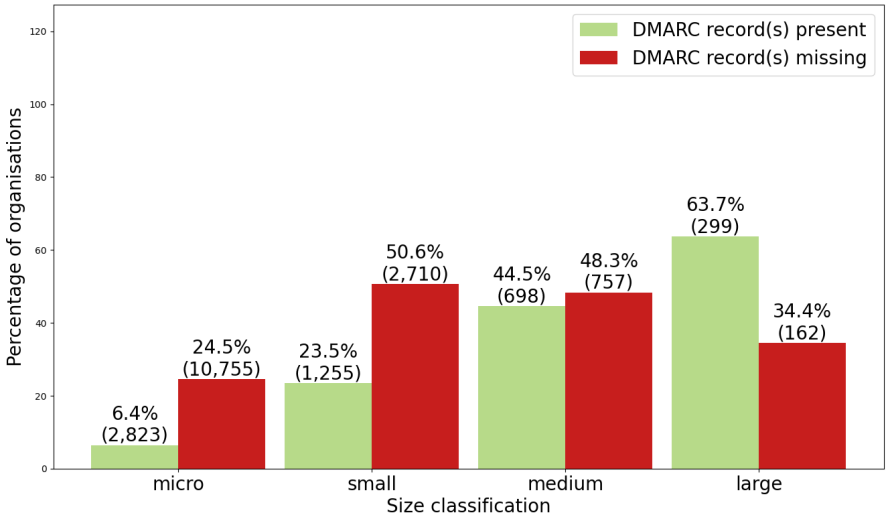
Figure 9. Percentage of vulnerable companies per size classification based on the findings for the indicator - Recommended email security extension (SPF and/or DMARC) is not used (VULN E.2.2). Absolute numbers for the vulnerability status of surveyed organisations can be seen in parentheses.

The high adoption rate of DMARC for micro companies can be explained by the high proportion of well-known mail providers used by them, as shown in Section 5.1.5. All 25 popular mail domains analysed in Section 5.1.5 had SPF record set. For missing DMARC records, the popular shelf company email domain *wasp.ee* was also missing the records. As such, after removing the organisations relying on

these popular mail providers, the distribution changed (see Figure 10a for results for SPF and 10b for DMARC).



(a) Results of the SPF component of the VULN E.2.2 measurements (with popular service provider domains removed).



(b) Results of the DMARC component of the VULN E.2.2 measurements (with popular service provider domains removed).

Figure 10. Percentage of vulnerable companies per size classification based on the findings, with popular service provider domains removed, for the indicator - Recommended email security extension (SPF and/or DMARC) is not used (VULN E.2.2). Absolute numbers for vulnerability status of surveyed organisations can be seen in parentheses.

Based on these findings, it appears that the utilisation of DMARC records for email security with custom company domains is relatively low. Yet SPF adoption rate remains around 90% for custom domains as well. Due to the relatively low adoption rate of DMARC by surveyed companies, only samples of companies that failed SPF checks were picked for more detailed manual analysis.

All nine large companies with missing SPF records were selected for manual review, and it was observed that all of these companies also failed corresponding DMARC checks. Manual verification also revealed 27% (3) false positives amongst these companies, which seemed to have been caused by DNS caching in the local network when the checks were run. Yet the other tested large companies didn't have SPF records configured for their mail domain.

For the medium companies that had failed the SPF check, 10 with the highest revenue were selected for further checks. Similarly to the large organisations, all of the chosen organisations also lacked DMARC records. However, in contrast to findings for large organisations, no false positives were found, and in fact, SPF records were missing for all sampled organisations. Except for two outliers, the same patterns were accurate for small- and micro-sized companies. A single micro organisation had defined an SPF policy *"v=spf1 -all"* that disallowed all emails from being sent, and there was one small company amongst the ten reviewed ones that had a DMARC record but no SPF one.

The adoption rate for SPF records among surveyed organisations is significantly higher than that of DMARC adoption. In total, 74.5% (106,327), 49% (3,551), 51.4% (871) and 64.5% (314) of, respectively, micro-, small-, medium and large-sized organisations have adopted both SPF and DMARC records to protect their email domains. However, it should be noted that for smaller organisations, this is most likely due to high reliance on free email service provider accounts as mentioned in Section 5.1.5.

5.1.7 VULN M.1 Detecting unregistered phone numbers

As the service responded with service provider details for known numbers and the majority of unknown numbers and with numbers unknown for the small subset of unknown numbers, the experiment was set up assuming that the response number unknown means the phone number is available.

The experiment was carried out on 22nd November 2023 after notifying the Estonian Computer Security Incident Response Team and the Consumer Protection and Technical Regulatory Authority of the increased traffic against the service. The results from querying for availability for 95,899 mobile numbers associated with Estonian companies can be seen in Figure 11. Notice that there is a third category in addition to registered and dangling phone numbers called foreign, which is used as a catchall to filter out numbers with non-Estonian prefixes considered

out-of-scope for this study.

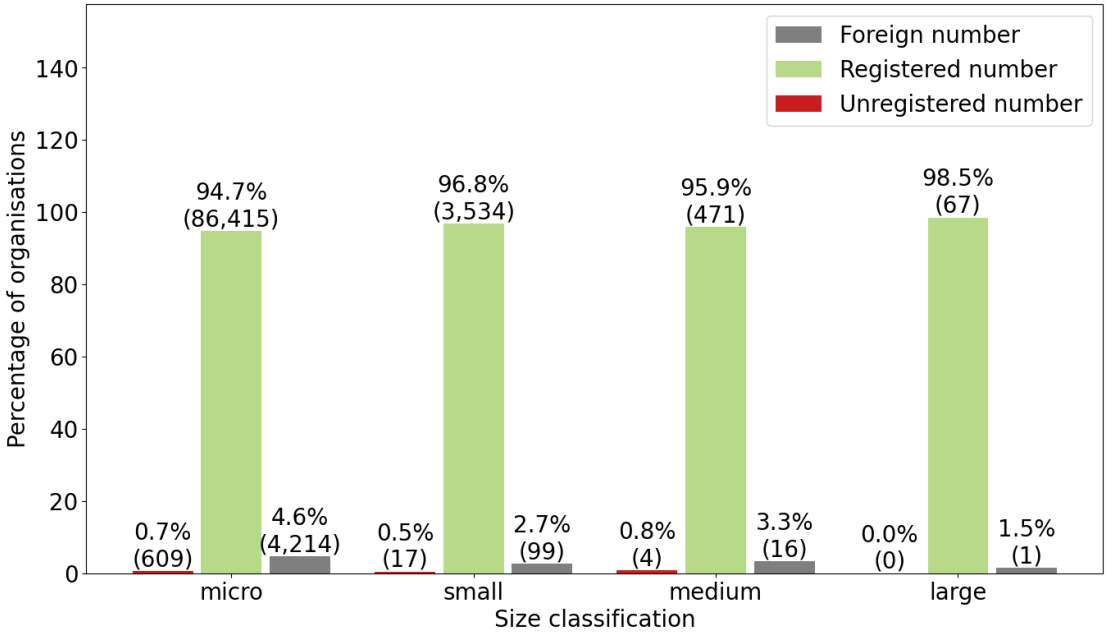


Figure 11. Percentage of vulnerable companies per size classification based on the findings for the indicator - Mobile number is available for re-registration (VULN M.1). Absolute numbers for vulnerability status of surveyed organisations can be seen in parentheses.

Number verification yielded 632 companies whose associated mobile numbers did not get a match from the service. However, analysis of the findings and comparison of the results with the Estonian numeration plan[74], as well as checking the number of registration opportunities with the local telecommunication companies, does not seem to support the initial assumptions. Since new numbers offered on the telecommunication service provider websites are shown as booked on the technical regulatory authority's website.

As such, assumptions about the possibility of identifying available phone numbers through the number availability service offered by the Consumer Protection and Technical Regulatory Authority were proven false. Numbers that didn't get a match can not be re-registered by any parties according to the local numeration plan[74]. Yet the author of this work believes that it is still possible to register phone numbers associated with companies. This process would most likely involve querying available numbers from telecommunication provider websites.

5.2 IOC measurements

For measuring IOC indicators, IOC E.1 was calculated separately from the vulnerability indicators defined for email addresses. IOC W.1 relied on findings from Section 5.1.1, as domains without SOA records were assumed not to host any websites. The results from these measurements are described throughout this subsection.

5.2.1 IOC E.1 Leaked credentials

The queries were run between 12-16 February 2024, and as a result, out of the 152,079 e-mail addresses, only 25.8% (39,271) were not reported as being part of any data leaks reported to "Have I Been Pwned". The rest of the email addresses had been associated with 566 data leaks (see Figure 12 for distribution of companies that had their email associated with data leaks).

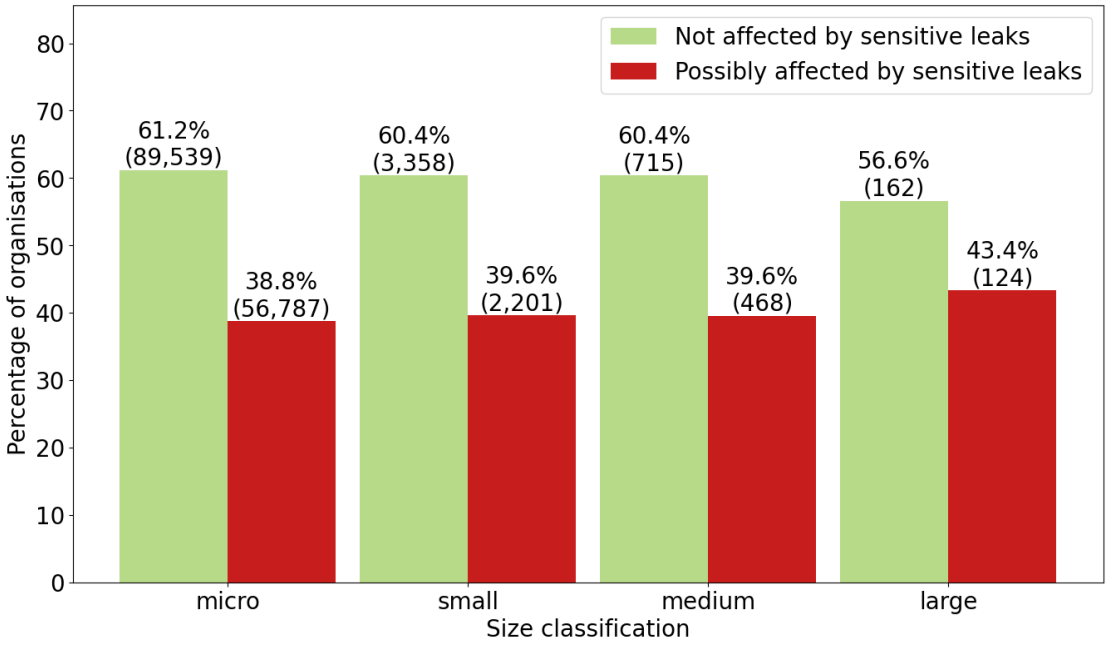


Figure 12. Percentage of vulnerable companies per size classification based on if their contact email had been exposed in at least one data leak. Absolute numbers for the vulnerability status of surveyed organisations can be seen in parentheses.

However, not all data leaks have similar severity and security implications. To provide in-depth insights into other data types associated with impacted business e-mail addresses, the author queried the relevant API again to acquire additional information on all 566 data breaches affecting the studied companies. The query

revealed that 112 different classes of data were leaked due to these breaches. Amongst them were sensitive data such as passwords, credit card data, contents of email messages, as well as already public information such as homepage URLs, phone numbers and dates of birth (see Appendix 8.1 for a complete list of data types found from the leaks).

As a result of filtering for secrets, 418 data breaches were classified as containing leaked secrets (as defined in Table 6). Companies from all four classifications have had their email associated with secrets from one or more data leaks, with a total of 27.3% (41,944) companies impacted. It is worth mentioning that accounts relating to email addresses of some organisations had been disclosed in more than 20 breaches.

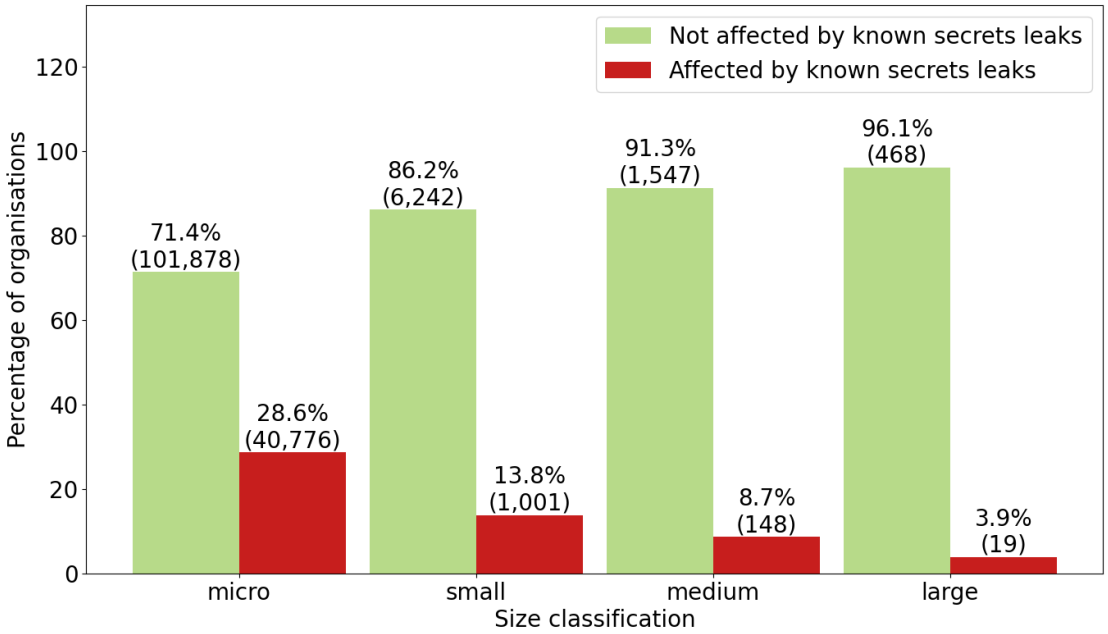


Figure 13. Percentage of vulnerable companies per size classification based on the findings for the indicator - Credentials associated with the organisation's listed email address have been published as a result of a leak (IOC E.1). Absolute numbers for the vulnerability status of surveyed organisations can be seen in parentheses.

For a deeper analysis, the nature and composition of leaks affecting surveyed companies were analysed separately for each size classification. For the large companies, this meant a total of 10 leaks affecting 6.6% (19) of the tested large companies, with some companies affected by up to 4 leaks. All the leaks seem to originate from data breaches of websites used for professional purposes (such as

LinkedIn[12], NitroPDF[13], and Dropbox[14]).

The volume and distribution of breaches for medium-sized companies are significantly different from those noted for large organisations, with 12.5% (148) of the medium organisations being affected by 42 secret leaks. The source of the leaks is not limited to leaks from sites primarily associated with professional use but also originate from breaches of sites such as VKontakte[15], MyHeritage[16] and Myfitnesspal[17]. This indicates email account cross-usage for both business and private purposes.

A total of 18% (1,001) and 27.9% (40,776) of respectively small- and micro-sized organisations were impacted by sensitive data breaches involving secrets disclosure, with micro-sized being represented in 99.7% (417) of the tested breaches involving secrets disclosure, with the only exclusion of no emails relating to surveyed micro-companies being leaked from the Malwarebytes breach[18].

5.2.2 IOC W.1 Unexpected redirection

To gather necessary metrics for unexpected redirections, website addresses were visited with a script that followed all the redirects up to the depth of five. The script was run on the 12,388 companies' websites that passed the first layer of domain checks mentioned in Section 5.1.1 (website domain had SOA record).

The testing did not yield any significant results, with 23% (2,951) of the queries resulting in an exception and <1% (12) of the websites redirecting above the configured red line of five redirects. During manual testing, three of the 12 redirection issues did not load in the modern browser. Reviewing raw HTTP data revealed an endless redirection loop from the website root for all three sites (all belonging to micro-sized organisations). From the rest of the potentially redirect looped sites, eight opened fine on a modern browser, and one was unreachable after redirecting to HTTPS connection.

The domains From each redirect were extracted from the sites that successfully followed the redirect chain. Analysis of the sites located on domains that were found amongst the highest number of redirects revealed a site providing shelf companies 1% (171) and judicial bodies created for individual ships belonging to a shipping corporation <1% (36). Yet, from amongst any of the more populated domains, no indicators of malicious redirects were identified

[12]https://monitor.mozilla.org/breach-details/linkedin

[13]https://www.bleepingcomputer.com/news/security/hacker-leaks-full-database-of-77-million-nitro-pdf-user-records/

[14]https://monitor.mozilla.org/breach-details/Dropbox

[15]https://monitor.mozilla.org/breach-details/VK

[16]https://monitor.mozilla.org/breach-details/MyHeritage

[17]https://monitor.mozilla.org/breach-details/MyFitnessPal

[18]https://monitor.mozilla.org/breach-details/Malwarebytes

The presence of affiliate links was also investigated by listing redirection URLs with longer lengths (>50 characters). This yielded 53 matches, most of which were redirects to social media pages or website login forms for small and micro companies, with the single large company included in the query redirecting to a closed sign-up page.

In conclusion, check for squatters by checking HTTP redirects of active organisations' websites did not yield malicious redirects to affiliate sites or prominent phishing sites. It can be assumed that such indications might be more populous if the survey included inactive organisations and deprecated domains. As parsing redirect URLs did not indicate the website's contents, several checked website hosts may squat content on the same domain without redirects for parking and ad revenue.

5.3 Summary

In summary, the digital assets of 152,091 organisations were tested for common security issues and misconfigurations, enabling impersonation attacks against the company, its clients or unrelated 3rd parties. These measurements revealed several event findings on the available surface that can be abused to impersonate the vulnerable company. In addition, contact email addresses of surveyed judicial persons were checked against known data breaches for a hint on the scale of ongoing abuse.

It was found that 7% (947) out of 13,448 websites listed in the e-Business register had their website domain expired and opened for registration. As such, the results show the likelihood of smaller companies being vulnerable to website domain takeover for cybersquatting is higher than that of larger companies. Almost 10% of the domains associated with websites of micro-sized companies were open for re-registration, while none of the website domains belonging to large organisations were unregistered.

Table 12. Summary of pass rate of the tested website identity protection measures by organisation size.

Size	**Sample size**	**passed VULN W.1**	**passed VULN W.2**	**passed VULN W.3**
micro	100% (10,832)	85.8% (9,906)	83.3% (9,025)	15.7% (1,702)
small	100% (1,729)	98.5%(1,711)	94.3% (1,634)	14.3 % (247)
medium	100% (642)	99.2% (639)	94.6% (608)	18.2% (116)
large	100% (245)	100% (245)	97.6% (239)	33.5% (82)
Total	**13,448**	**93% (12,501)**	**85.6% (11,506)**	**16% (2,147)**

The second level of vulnerability measurements was conducted for websites, where transport layer encryption for websites was tested. In conjunction with SSL/TLS measurements, a third level of vulnerability checks was done for websites to check whether HSTS headers were configured to defend against MiTM attacks. Results show that adoption rates for SSL/TLS precede global adoption rates, with 92% (11,470) of the surveyed websites offering encrypted HTTPS with valid certificates. Yet the adoption rate of HSTS falls short of the global adoption rate of 27%, especially for smaller companies websites. Table 12 shows summarised results for the test conducted on the websites of studied companies.

Table 13. Summary of pass rate of the tested email security measures by organisation size.

Size	**Sample size**	**passed VULN E.1**	**passed VULN E.2.2**	**passed VULN E.2.1**
micro	100% (142,654)	99.3% (141,614)	74.5% (106,327)	31.6% (45,093)
small	100% (7,243)	99.9%(7,233)	49.0% (3,551)	74.2% (5373))
medium	100% (1,695)	99.9% (1,694)	51.4% (871)	92.9% (1575)
large	100% (487)	99.8% (486)	64.5% (314)	98.2% (478)
Total	**100% (152,079)**	**99.3% (151,027)**	**73.0% (111,063)**	**34.5% (52519)**

For email, only 0.7% (1052) out of the 152,079 had their listed email domain open for re-registration. Organisations that passed the first vulnerability checks for the corresponding asset type were checked for additional weaknesses. The second level of tests for email addresses involved reviewing the mail domain for missing DNS SPF and DMARC records, as well as mapping the usage of common mail providers and associated risks. These checks revealed that 57% (69,747) of the companies had their email hosted with Gmail (using their free domain gmail.com). Additionally, a sizable proportion of other common regional providers was used, covering a total of 65% of the surveyed email addresses with 26 email domains. The proportion of such mail providers was especially prominent for smaller companies. SPF and DMARC usage measurements revealed a relatively high adoption rate of SPF records across all the size classes as well as low adoption of DMARC records across companies not using common mail providers. Comprehensive results of vulnerability checks for email can be seen in Table 13.

Table 14. Summary of pass rate of the business email-related credential leaks.

Size	**Sample size**	**passed IOC E.1**
micro	100% (142,654)	71.4% (101,878)
small	100% (7,243)	86.2%(6,242)
medium	100% (1,695)	91.3% (1,547)
large	100% (487)	96.1% (468)
Total	**100% (152,079)**	**72.4% (110,135)**

As for measuring compromise indicators, corporate email addresses' presence in datasets from known data breaches was checked. This revealed that only 25.6% of the surveyed companies were not a part of any of the 566 data leaks checked against in the scope of this thesis. Yet not all of the data in the leak significantly impacts the success of impersonation (e.g., BVD leak[19] contains similar information that is already available from the e-Business register). However, 27.6% of the surveyed companies had their listed email address and associated passwords (see Table 14) listed in one of the leaks. The proportion of companies affected by the secret leaks was more significant for smaller companies than for larger ones. This seems to have been caused by the cross-usage of business email for personal purposes, which occurs more often in smaller companies.

Phone number availability for re-registration (VULN M.1) and unexpected redirection from websites (IOC W.1) were also measured. However, the findings from these tests did not provide sufficient data to answer the research questions.

[19]https://monitor.mozilla.org/breach-details/BVD

6 Validation

The validity of the research was verified through the manual reviewing of the results for selected organisations from each size classification after each measurement step (see Section 5 for details). As a result, an error in the domain extraction step was identified and fixed. Additionally, Estonian Computer Security Incident Response Team (CERT-EE) was notified of the vulnerabilities discovered due to this work and asked to validate the findings.

6.1 Issues with domain extraction

During the manual analysis of results for the VULN W.1 the author discovered that the original method used for domain extraction was naive due to the existence of valid TLDs that consist of multiple parts (such as.*com.ee* and *.co.uk*). Similar issues were also noted when measuring VULN E.1. Since email service is not limited to being hosted on a registered domain, it can be hosted on subdomains. In these cases, original domain extraction techniques combined with SOA record checks and WHOIS requests did not work (Total impact, along with transitive errors across other checks, can be seen in Table 15).

Table 15. Overview of identified errors resulting from issues with chosen domain extraction strategy and re-measurement time for correction.

Indicator	**Error percentage (number of errors)**	**Re-measurement time**
VULN W.1	0.3% (39)	11th May 2024
VULN W.2	0.1% (14)	12th May 2024
VULN W.3	0.1% (14)	12th May 2024
VULN E.1	0.06% (93)	11th May 2024
VULN E.2.1	-	-
VULN E.2.2	0.3% (94)	12th May 2024

The issue was fixed by redoing by introducing the Python library *tldextract* for extracting domain components and redoing the vulnerability measurements for all the impacted organisations and indicators. Re-measurements were done between 11-12th May 2024, and none of the indicators of compromise were found to be significantly affected.

6.2 Notifying CERT-EE

The measurement results were offered to Estonian Computer Security Incident Response Team (CERT-EE) for analysis on the 11th of April 2024. They responded

to the author on 18th April and agreed to review the findings. As a result, the author compiled and submitted a list of discovered vulnerabilities along with the organisations impacted to CERT-EE on 21st April 2024.

The authority had reviewed the findings by the 2nd of May. CERT-EE confirmed that eight companies were subjected to NIS2 requirements among vulnerable organisations. These organisations were informed of the weaknesses they discovered. As a result, at least two of the discovered weaknesses have already been resolved.

CERT-EE is unaware of any significant incidents caused by the re-registration of expired domains by 3rd parties. The techniques favoured by cyber criminals rely on similar domain names and content from legitimate websites. Password leaks from data breaches remain a significant exception amongst the reported issues as they do not have the resources to investigate the causal relationship between incidents and data leaks involving micro companies. Cross- and re-use passwords significantly increase intrusion risks for these incidents. Awareness-raising campaigns are being done in hopes of lowering said risks.

In addition, the Estonian Computer Security Incident Response Team (CERT-EE) expressed their commitment to notify organisations subject to the National Cyber Security Act of Estonia in cases where they become aware of potential vulnerabilities that could affect said companies. This is expected to lower the number of similar issues in the future.

7 Discussion

After measuring the vulnerability of legal entities to common weaknesses discussed in previous works, a contribution has been made towards an improved understanding of cyber security issues active organisations face. This section answers the stated research questions, discusses the findings within the broader context of the domain and describes known limitations of the methods used.

Answers to the research questions

1. PRQ: To what extent are organisations of different sizes vulnerable to identity theft?

    - **Answer:** Most studied organisations at any size classification are not directly vulnerable to significant electronic identity theft such as email or website domain hijacking. The risk to other notable risks is somewhat higher, with a quarter of studied organisations not having implemented DNS records for email security and 15% of listed company websites not supporting transport layer encryption.

    (a) How does organisations' adoption rate of digital identity protection measures compare to global statistics?

    – **Answer:** According to findings in Section 5.1, the adoption rate of email protocol extensions SPF and DMARC exceeds global adoption rates. The adoption rate of SSL/TLS amongst the survey population also exceeds global statistics, with the HSTS adoption rate being slightly lower when compared to global trends.

    (b) How much do the technical corporate identity protection measures depend on the company size?

    – **Answer:** The findings in Section 5 of this work indicate that, on average, larger companies have better protected their digital identity than those falling inside the SME classifications.

2. SRQ2: To what extent have digital identities or active organisations been compromised?

    - **Answer:** According to findings in Section 5.2.1 passwords connected to email addresses listed as official contacts of 27.3% (41,944) of the surveyed companies have been disclosed due to verified data breaches. However, there is some nuance to this:

    – The password disclosed as part of the data breach is generally used for the breached site and does not automatically mean that the

same password has been used elsewhere (e.g. for the email account itself). Yet, due to the prevalence of password reuse across different sites, the general magnitude of compromised e-mail addresses can still be inferred from them.

The measurement results support critique towards weakness of extended validation of SSL/TLS certificates[21]. It was found that a sizable proportion (7%) of websites associated with Estonian registered organisations are inactive and could be re-registered, potentially enabling abuse of registry information to acquire website certificates with higher claim of authenticity[20].

In addition, the upper boundary for the impact of known data breaches on organisations was quantified. This revealed that accounts associated with more than 27% of the active organisations are known to have been breached along with associated secrets such as passwords. This gives insights on the extent to which data breaches can affect breached organisations as had been proposed by Schlackl et al. in their work[30].

The extent of identity protection measures applied by organisations was compared to global statistics as reported by [64, 65, 58]. The results differed from previous findings, showing that the adoption rate for popular technical identity protection measures for websites (SSL/TLS & HSTS) and email addresses (SPF & DMARC) tends to be higher than the global average. Thus signifying the added granularity of measuring vulnerability associated with entities instead of measuring adoption based on a selection of the most popular website domains.

Work contributes towards a better understanding of available opportunities that contribute towards malicious actors going through with criminal activities [3]. The findings indicate opportunities for speculative cybersquatting on domains previously used by legitimate companies by re-registering expired DNS domains, as observed by Lauinger et al. [46].

In addition, domain re-registration or passwords from public data leaks were noted to be usable to exploit registry information for more believable e-mail phishing, such as BEC attacks. Thus reinforcing observations made by industry authorities on the low technical difficulty of such attacks [2, 16, 28].

7.1 Limitations

The numbers and proportions reflect the tested assets' security when measured. The assessment only included one of the digital assets of each type, as they were listed in the registry. As such, the scope of the work did not include all the digital assets of the surveyed organisations.

In addition, the vulnerability measurements done during this work were non-intrusive. Due to this, none of the discovered weaknesses were attempted to be

used for impersonation, which means that the implications of the discovered issues might differ from the thesis results.

8 Conclusions

This section concludes the results of this thesis. In addition, it provides input for further exploration in this domain. The thesis contributes towards a better understanding of the vulnerability surface of organisations to common types of electronic identity theft. This goal was achieved by measuring vulnerability to common methods for digital identity theft across companies registered in the Republic of Estonia. The thesis results give additional insights into the extent to which organisations of different sizes are vulnerable to digital identity theft.

As a result of this work, it has become apparent that the contact information in the e-Business Register is out-of-date for many active companies. Out-of-date information could be used to leverage implicit trust in the contents of the register to aid in identity theft. To reduce opportunities for such abuse, the author recommends the introduction of a flow to update or verify contact details as a part of sending annual reports.

The findings also raise the question of whether weak authenticators, such as possession of a domain or an email address, should be used for authenticating legal entities. The need for more robust electronic identification for business has been identified based on the vulnerabilities discovered. Further efforts towards providing regulatory support and technical means to implement electronic seals for identifying digital assets belonging to legal entities are recommended across the EU.

Proposals for future works are described in the following subsection.

8.1 Future work

Exploration of information security of assets tied to registered legal entities would be beneficial to get further insights into practical issues faced by corporations inside the EU (and beyond). With the implementation of the EU NIS2 directive on the horizon for member states, the feedback loop between practical issues and legislative goals is necessary. Deeper insights into potential disparities in the industry could be gained by separating surveyed companies into categories according to their field of economic activities.

Additionally, the indicators for security issues used in this work are only a small (albeit popular) subset of impersonation vulnerabilities used for malicious purposes. More intrusive tests in collaboration with a smaller sub-selection of willing organisations could offer additional insights into the apparent and actual surface of vulnerability. Registering several expired email domains and setting a catch-all alias could yield further insights into how malicious actors could abuse these expired email domains.

Methods used in this thesis can be used and extended to analyse the level of electronic identity protection of legal persons registered in other jurisdictions.

References

[1] F. Risk-Based, "Multifactor authentication for e-commerce," *NIST SPECIAL PUBLICATION*, p. 17B, 1800.

[2] "ENISA Threat Landscape 2022, url=https://www.enisa.europa.eu/publications/enisa-threat-landscape-2022, journal=ENISA, year=2022, month=Nov."

[3] G. R. Newman, M. M. McNally *et al.*, "Identity theft literature review," 2005.

[4] C. of Registers and I. Systems, "E-business register open data," 2023. [Online]. Available: https://avaandmed.ariregister.rik.ee/en/downloading-open-data

[5] P. Grassi, M. Garcia, and J. Fenton, "Digital identity guidelines," National Institute of Standards and Technology, Tech. Rep., 2020.

[6] L. Fernandes, "Fraud in electronic payment transactions: Threats and countermeasures," *Asia Pacific Journal of Marketing & Management Review ISSN*, vol. 2319, p. 2836, 2013.

[7] J. Wadleigh, J. Drew, and T. Moore, "The e-commerce market for "lemons": Identification and analysis of websites selling counterfeit goods," in *Proceedings of the 24th International Conference on World Wide Web*, ser. WWW '15. Republic and Canton of Geneva, CHE: International World Wide Web Conferences Steering Committee, 2015, p. 1188–1197. [Online]. Available: https://doi.org/10.1145/2736277.2741658

[8] M. Bitaab, H. Cho, A. Oest, Z. Lyu, W. Wang, J. Abraham, R. Wang, T. Bao, Y. Shoshitaishvili, and A. Doupé, "Beyond phish: Toward detecting fraudulent e-commerce websites at scale," in *2023 IEEE Symposium on Security and Privacy (SP)*. IEEE Computer Society, 2023, pp. 2566–2583.

[9] T. Daengsi, P. Chomchuen, P. Klamklomchit, P. Pornpongtechavanich, K. Saribua, W. Thimthong, and N. Sukniyom, "Chaladohn: Website for avoiding of online shopping scams in thailand," in *2022 IEEE 12th Symposium on Computer Applications Industrial Electronics (ISCAIE)*, 2022, pp. 149–152.

[10] C. Carpineto and G. Romano, "Learning to detect and measure fake ecommerce websites in search-engine results," in *Proceedings of the International Conference on Web Intelligence*, ser. WI '17. New York, NY, USA: Association for Computing Machinery, 2017, p. 403–410. [Online]. Available: https://doi.org/10.1145/3106426.3106441

[11] A. M. Marshall and B. C. Tompsett, "Identity theft in an online world," *Computer Law & Security Review*, vol. 21, no. 2, pp. 128–137, 2005.

[12] C. Guitton, "Criminals and cyber attacks: The missing link between attribution and deterrence." *International Journal of Cyber Criminology*, vol. 6, no. 2, 2012.

[13] E. I. S. Authority, "Cyber security in estonia 2024," 2024.

[14] Council of European Union, "Directive (eu) 2014/910," 2014, https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=uriserv:OJ.L_.2014.257.01.0073.01.ENG.

[15] Riigi Infosüsteemi Amet, "eidas autentimistasemed," 2017, https://www.ria.ee/sites/default/files/documents/2022-11/eIDAS-autentimistasemed.pdf.

[16] I. Lella, C. Ciobanu, E. Tsekmezoglou, M. Theocharidou, E. Magonara, A. Malatras, R. Svetozarov Naydenov *et al.*, "ENISA threat landscape 2023: July 2022 to June 2023," 2023.

[17] "Nissan motor co., ltd. v. nissan computer corp." p. 1154, 2000.

[18] "Commercial register act1." [Online]. Available: https://www.riigiteataja.ee/en/eli/530112022003/consolide

[19] Council of European Union, "Directive (eu) 2015/849," 2015, https://eur-lex.europa.eu/eli/dir/2015/849/oj/eng.

[20] Digicert, "What's the difference between DV, OV EV SSL certificates?" [Online]. Available: https://www.digicert.com/difference-between-dv-ov-and-ev-ssl-certificates#Compare

[21] A. T. Dan Goodin, "Nope, this isn't the HTTPS-validated Stripe website you think it is." [Online]. Available: https://arstechnica.com/information-technology/2017/12/nope-this-isnt-the-https-validated-stripe-website-you-think-it-is/

[22] "Phishing most common Cyber Incident faced by SMEs — ENISA." [Online]. Available: https://www.enisa.europa.eu/news/enisa-news/phishing-most-common-cyber-incidents-faced-by-smes

[23] M. Bossetta, "The weaponization of social media: Spear phishing and cyber-attacks on democracy," *Journal of international affairs*, vol. 71, no. 1.5, pp. 97–106, 2018.

[24] M. H. Nguyen Ba, J. Bennett, M. Gallagher, and S. Bhunia, "A case study of credential stuffing attack: Canva data breach," in *2021 International Conference on Computational Science and Computational Intelligence (CSCI)*, 2021, pp. 735–740.

[25] APWG, "Phishing activity trends report 4rd quarter 2022," 2024. [Online]. Available: https://docs.apwg.org/reports/apwg_trends_report_q4_2023.pdf

[26] E. R. Leukfeldt and T. J. Holt, "Cybercrime on the menu? examining cafeteria-style offending among financially motivated cybercriminals," *Computers in Human Behavior*, vol. 126, p. 106979, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0747563221003022

[27] T. Jon Clay, "Phishing as a Service Stimulates Cybercrime." [Online]. Available: https://www.trendmicro.com/en_us/ciso/23/c/phishing-as-a-service-phaas.html

[28] "ENISA Threat Landscape 2021, url=https://www.enisa.europa.eu/publications/enisa-threat-landscape-2021," Oct 2022.

[29] Surfshark B.V, "Global data breach statistics," 2024, https://surfshark.com/research/data-breach-monitoring.

[30] F. Schlackl, N. Link, and H. Hoehle, "Antecedents and consequences of data breaches: A systematic review," *Information & Management*, vol. 59, no. 4, p. 103638, 2022.

[31] K. D. Martin, A. Borah, and R. W. Palmatier, "Data privacy: Effects on customer and firm performance," *Journal of Marketing*, vol. 81, no. 1, pp. 36–58, 2017. [Online]. Available: https://doi.org/10.1509/jm.15.0497

[32] G. Johnson, "Economic research on privacy regulation: Lessons from the gdpr and beyond," 2022.

[33] R. Sen and S. Borle, "Estimating the contextual risk of data breach: An empirical approach," *Journal of Management Information Systems*, vol. 32, no. 2, pp. 314–341, 2015.

[34] S. Romanosky, R. Telang, and A. Acquisti, "Do data breach disclosure laws reduce identity theft?" *Journal of Policy Analysis and Management*, vol. 30, no. 2, pp. 256–286, 2011.

[35] C. Paulsen, "Glossary of key information security terms," National Institute of Standards and Technology, Tech. Rep., 2018.

[36] A. A. Gillespie, "The electronic spanish prisoner: romance frauds on the internet," *The Journal of Criminal Law*, vol. 81, no. 3, pp. 217–231, 2017.

[37] "ENISA Threat Landscape 2020 - phishing, url=https://www.enisa.europa.eu/publications/phishing, journal=ENISA, year=2020, month=Oct."

[38] Y. Zeng, T. Zang, Y. Zhang, X. Chen, and Y. Wang, "A comprehensive measurement study of domain-squatting abuse," in *ICC 2019 - 2019 IEEE International Conference on Communications (ICC)*, 2019, pp. 1–6.

[39] J. Spaulding, D. Nyang, and A. Mohaisen, "Understanding the effectiveness of typosquatting techniques," in *Proceedings of the Fifth ACM/IEEE Workshop on Hot Topics in Web Systems and Technologies*, ser. HotWeb '17. New York, NY, USA: Association for Computing Machinery, 2017. [Online]. Available: https://doi.org/10.1145/3132465.3132467

[40] A. Dinaburg, "Bitsquatting: Dns hijacking without exploitation," *Proceedings of BlackHat Security*, 2011.

[41] F. Quinkert, T. Lauinger, W. Robertson, E. Kirda, and T. Holz, "It's not what it looks like: Measuring attacks and defensive registrations of homograph domains," in *2019 IEEE Conference on Communications and Network Security (CNS)*, 2019, pp. 259–267.

[42] K. Tian, S. T. Jan, H. Hu, D. Yao, and G. Wang, "Needle in a haystack: Tracking down elite phishing domains in the wild," in *Proceedings of the Internet Measurement Conference 2018*, 2018, pp. 429–442.

[43] N. Nikiforakis, M. Balduzzi, L. Desmet, F. Piessens, and W. Joosen, "Soundsquatting: Uncovering the use of homophones in domain squatting," in *Information Security: 17th International Conference, ISC 2014, Hong Kong, China, October 12-14, 2014. Proceedings 17*. Springer, 2014, pp. 291–308.

[44] D. Liu, Z. Li, K. Du, H. Wang, B. Liu, and H. Duan, "Don't let one rotten apple spoil the whole barrel: Towards automated detection of shadowed domains," in *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, ser. CCS '17. New York, NY, USA: Association for Computing Machinery, 2017, p. 537–552. [Online]. Available: https://doi.org/10.1145/3133956.3134049

[45] S. M. Z. U. Rashid, M. I. Kamrul, and A. Islam, "Understanding the security threats of esoteric subdomain takeover and prevention scheme," in *2019 International Conference on Electrical, Computer and Communication Engineering (ECCE)*, 2019, pp. 1–4.

[46] T. Lauinger, A. Chaabane, A. S. Buyukkayhan, K. Onarlioglu, and W. Robertson, "Game of registrars: An empirical analysis of Post-Expiration domain name takeovers," in *26th USENIX Security Symposium (USENIX Security 17)*. Vancouver, BC: USENIX Association, Aug. 2017, pp. 865–880. [Online]. Available: https://www.usenix.org/conference/usenixsecurity17/technical-sessions/presentation/lauinger

[47] A. McDonald, C. Sugatan, T. Guberek, and F. Schaub, "The annoying, the disturbing, and the weird: Challenges with phone numbers as identifiers and phone number recycling," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021, pp. 1–14.

[48] B. Bhushan, G. Sahoo, and A. K. Rai, "Man-in-the-middle attack in wireless and computer networking — a review," in *2017 3rd International Conference on Advances in Computing,Communication Automation (ICACCA) (Fall)*, 2017, pp. 1–6.

[49] Y. Chen, W. Trappe, and R. P. Martin, "Detecting and localizing wireless spoofing attacks," in *2007 4th Annual IEEE Communications Society Conference on Sensor, Mesh and Ad Hoc Communications and Networks*, 2007, pp. 193–202.

[50] Z. Wu, N. Evans, T. Kinnunen, J. Yamagishi, F. Alegre, and H. Li, "Spoofing and countermeasures for speaker verification: A survey," *Speech Communication*, vol. 66, pp. 130–153, 2015. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167639314000788

[51] A. Hadid, N. Evans, S. Marcel, and J. Fierrez, "Biometrics systems under spoofing attack: An evaluation methodology and lessons learned," *IEEE Signal Processing Magazine*, vol. 32, no. 5, pp. 20–30, 2015.

[52] A. Hadid, "Face biometrics under spoofing attacks: Vulnerabilities, countermeasures, open issues, and research directions," in *2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2014, pp. 113–118.

[53] S. K. Wong and S.-M. Yiu, "Location spoofing attack detection with preinstalled sensors in mobile devices." *J. Wirel. Mob. Networks Ubiquitous Comput. Dependable Appl.*, vol. 11, no. 4, pp. 16–30, 2020.

[54] X. Ye, B. Zhao, T. H. Nguyen, and S. Wang, "Social media and social awareness," *Manual of digital earth*, pp. 425–440, 2020.

[55] B. Adida, S. Hohenberger, and R. L. Rivest, "Fighting phishing attacks: A lightweight trust architecture for detecting spoofed emails," *DIMACS Wkshp on Theft in E-Commerce, April 2005*, 2005.

[56] D. J. C. Klensin, "Simple Mail Transfer Protocol," RFC 5321, Oct. 2008. [Online]. Available: https://www.rfc-editor.org/info/rfc5321

[57] S. Maroofi, M. Korczyński, A. Hölzel, and A. Duda, "Adoption of email anti-spoofing schemes: A large scale analysis," *IEEE Transactions on Network and Service Management*, vol. 18, no. 3, pp. 3184–3196, 2021.

[58] S. Czybik, M. Horlboge, and K. Rieck, "Lazy Gatekeepers: A Large-Scale Study on SPF Configuration in the Wild," in *Proceedings of the 2023 ACM on Internet Measurement Conference*, 2023, pp. 344–355.

[59] J. Selvi, "Bypassing http strict transport security," *Black Hat Europe*, vol. 54, 2014.

[60] J. G. Christoph Kerschbaumer, Tomer Yavor and A. Edelstein, "Firefox 91 introduces https by default in private browsing," 2021. [Online]. Available: https://blog.mozilla.org/security/2021/08/10/firefox-91-introduces-https-by-default-in-private-browsing/

[61] G. C. team, "Increasing https adoption," 2021. [Online]. Available: https://blog.chromium.org/2021/07/increasing-https-adoption.html

[62] J. Hodges, C. Jackson, and A. Barth, "HTTP Strict Transport Security (HSTS)," RFC 6797, Nov. 2012. [Online]. Available: https://www.rfc-editor.org/info/rfc6797

[63] "Hsts preload list submission." [Online]. Available: https://hstspreload.org/

[64] W3techs, "Increasing https adoption," 2024. [Online]. Available: https://w3techs.com/technologies/details/ce-hsts

[65] ——, "Usage statistics of default protocol https for websites," 2024. [Online]. Available: https://blog.chromium.org/2021/07/increasing-https-adoption.html

[66] Council of European Union, "Directive (eu) 2022/2555," 2022, https://eur-lex.europa.eu/eli/dir/2022/2555.

[67] E. Commission *et al.*, "Recommendation 2003/361," 2003.

[68] C. Protection and T. R. A. of the Republic of Estonia. [Online]. Available: https://nba.ttja.ee/numbriparing.aspx

[69] R. Oliemans, "WHOIS versus GDPR," July 2019. [Online]. Available: http://essay.utwente.nl/78753/

[70] P. Mockapetris, Tech. Rep., 1987. [Online]. Available: https://www.ietf.org/rfc/rfc1035.txt

[71] T. Hunt, "Have i been pwned," *URL: https://haveibeenpwned.com*, 2019.

[72] VK, "Information Services Agreement." [Online]. Available: https://help.mail.ru/legal/terms/mytracker/eng/agreement

[73] I. AS, "Teave uue Online.ee kohta." [Online]. Available: https://www.fjordmail.no/online.ee/

[74] Minister of Economic Affairs and Information Technology, "Eesti numeratsiooniplaan," 21.05.2018, https://www.riigiteataja.ee/akt/126022019009.

Appendix I. Glossary

APWG Anti-Phishing Working Group. 14

ARP Address Resolution Protocol. 18

BEC Business Email Compromise. 12, 16, 58

CERT-EE Estonian Computer Security Incident Response Team. 2, 6, 21, 47, 55, 56

CNAME Canonical Name. 36

CSV Comma-Separated Values. 28

DHCP Dynamic Host Configuration Protocol. 18

DKIM DomainKeys Identified Mail. 26

DMARC Domain-based Message Authentication, Reporting & Conformance. 6, 18, 26, 30, 31, 44–47, 53, 57, 58, 70

DNS Domain Name System. 10, 16–18, 28–30, 41, 47, 53, 57

ENISA The European Union Agency for Cybersecurity. 7, 10, 12–16

FQDN Fully Qualified Domain Name. 19

GDPR General Data Protection Regulation. 15, 44

HSTS HTTP Strict Transport Security. 6, 19, 26, 31, 38, 53, 58, 70

HTTP Hypertext Transfer Protocol. 18, 26, 32, 37, 51, 52

HTTPS Hypertext Transfer Protocol Secure. 18, 37, 51, 53

IOC Indicator of Compromise). 25, 49

IOC E.1 Credentials associated with the organisation's listed email address have been published as a result of a leak. 27, 32, 49, 50

IOC W.1 Website redirects visitors to an unrelated website (as a form of affiliate squatting). 28, 32, 49, 54

JSON JavaScript Object Notation. 28

KYB Know Your Business. 10

KYC Know Your Customer. 10

MiTM man-in-the-middle. 18, 26, 53

PhaaS Phishing-as-a-Service. 14

SME Small and medium-sized enterprise. 24

SOA Start of Authority. 29, 30, 33, 34, 39, 51, 55

SPF Sender Policy Framework. 6, 18, 26, 30, 31, 44–47, 53, 57, 58, 70

SSL/TLS Secure Sockets Layer and Transport Layer Security. 18, 26, 30, 31, 35–38, 53, 57, 58, 70

TLD Top-Level Domain. 17, 55

URL Uniform Resource Locator. 16, 28, 31, 32, 50, 52

VULN E.1 Email domain is not registered. 26, 29, 30, 33, 40, 44, 55

VULN E.2.1 Private email address is used for business purposes.. 26, 30, 31, 33, 53, 55

VULN E.2.2 Recommended email security extension (SPF and/or DMARC) is not used. 26, 30, 31, 33, 45, 46, 53, 55

VULN M.1 Mobile number is available for re-registration. 26, 30, 48, 54

VULN W.1 Website domain is not registered. 26, 29, 30, 33, 34, 55

VULN W.2 SSL/TLS is not supported by the website. 26, 30, 31, 36, 37, 55

VULN W.3 HSTS is not supported by the website. 26, 31, 33, 38, 55

II. Licence

Non-exclusive licence to reproduce thesis and make thesis public

I, **Andres Jõgi**,

1. I herewith grant the University of Tartu a free permit (non-exclusive licence) to

   reproduce for the purpose of preservation, including for adding to the DSpace digital archives until the expiry of the term of copyright,

   Quantitative Analysis on Vulnerability to Electronic Business Identity Theft Among Estonian Companies,

   supervised by Mari Seeba, Tarmo Oja and Markko Merzin.

2. I grant the University of Tartu a permit to make the work specified in p. 1 available to the public via the web environment of the University of Tartu, including via the DSpace digital archives, under the Creative Commons licence CC BY NC ND 3.0, which allows, by giving appropriate credit to the author, to reproduce, distribute the work and communicate it to the public, and prohibits the creation of derivative works and any commercial use of the work until the expiry of the term of copyright.

3. I am aware that the author retains the rights specified in p. 1 and 2.

4. I certify that granting the non-exclusive licence does not infringe on other persons' intellectual property rights or rights arising from the personal data protection legislation.

Andres Jõgi
15/05/2024

Appendix III. List of leaked data types (IOC E.1)

1. Driver's licenses
2. Partial credit card data
3. Salutations
4. Instant messenger identities
5. Avatars
6. IMSI numbers
7. Professional skills
8. Social media profiles
9. Flights taken
10. Appointments
11. Email messages
12. Cellular network names
13. Address book contacts
14. IMEI numbers
15. Device serial numbers
16. Employment statuses
17. Home ownership statuses
18. Partial dates of birth
19. Password strengths
20. Phone numbers
21. Relationship statuses
22. Physical attributes
23. Survey results
24. Passport numbers
25. Ages
26. Social connections
27. Clothing sizes
28. Payment methods
29. Vehicle details
30. Customer feedback
31. Purchases
32. Login histories
33. Customer interactions
34. Purchasing habits
35. Historical passwords
36. Government issued IDs
37. Living costs
38. Age groups
39. Political donations
40. PINs
41. Deceased statuses
42. Years of professional experience
43. Nationalities
44. Education levels
45. Security questions and answers
46. Mothers maiden names
47. Credit cards
48. Browser user agent details
49. Occupations
50. Social security numbers
51. Charitable donations
52. IP addresses
53. Job titles
54. Employers
55. Account balances
56. Net worths
57. Eating habits
58. Personal health data
59. Financial transactions
60. MAC addresses

61. User website URLs
62. Taxation records
63. Health insurance information
64. Time zones
65. Support tickets
66. Marital statuses
67. Browsing histories
68. Financial investments
69. Device usage tracking data
70. Apps installed on devices
71. Photos
72. Loyalty program details
73. Personal interests
74. Private messages
75. SMS messages
76. Email addresses
77. Spouses names
78. Job applications
79. Password hints
80. Passwords
81. Website activity
82. Genders
83. Homepage URLs
84. Audio recordings
85. Payment histories
86. Encrypted keys
87. Mnemonic phrases
88. Spoken languages
89. Places of birth
90. Buying preferences
91. Recovery email addresses
92. Profile photos
93. Bios
94. Nicknames
95. Smoking habits
96. Income levels
97. Religions
98. Chat logs
99. Auth tokens
100. Names
101. Credit status information
102. Geographic locations
103. Dates of birth
104. Bank account numbers
105. Ethnicities
106. Device information
107. Partial phone numbers
108. Licence plates
109. Family structure
110. Usernames
111. Physical addresses
112. Telecommunications carrier