University of Zurich[UZH]

MASTER THESIS – Communication Systems Group, Prof. Dr. Burkhard Stiller

Strategic Alignment in Cybersecurity Information Sharing: A Multidimensional Approach to Company Similarity Analysis

Christian Omlin
Zurich, Switzerland
Student ID: 14-936-165

Supervisor: Dr. Muriel F. Franco, Jan von der Assen
Date of Submission: October 6, 2023

University of Zurich
Department of Informatics (IFI)
Binzmühlestrasse 14, CH-8050 Zürich, Switzerland

ifi

Master Thesis
Communication Systems Group (CSG)
Department of Informatics (IFI)
University of Zurich
Binzmühlestrasse 14, CH-8050 Zürich, Switzerland
URL: http://www.csg.uzh.ch/

Declaration of Independence

I hereby declare that I have composed this work independently and without the use of any aids other than those declared (including generative AI such as ChatGPT). I am aware that I take full responsibility for the scientific character of the submitted text myself, even if AI aids were used and declared (after written confirmation by the supervising professor). All passages taken verbatim or in sense from published or unpublished writings are identified as such. The work has not yet been submitted in the same or similar form or in excerpts as part of another examination.

Zürich, October 6th, 2023

Signature of student

Abstract

In the prevailing digital era, heightened by an increasing incidence of cyberattacks, cybersecurity stands out as a critical concern for organizations of all sizes. While the necessity to bolster cybersecurity measures is universally acknowledged, determining an optimal strategy presents a complex challenge. This master thesis introduces a novel approach leveraging inter-company cybersecurity data sharing to assist organizations in honing their defensive measures. A tool was developed to discern the relevance of information-sharing entities by classifying companies across three dimensions: business, economic, and technical. Each dimension is defined by distinct factors, allowing for a precise comparison. An accompanying application was devised to represent the similarities among companies using the Euclidean distance and Pearson correlation. Through extensive evaluation, the Euclidean distance proved superior in the business and economic realms. However, for the technical dimension, dominated by integer values, the efficacy of both measures was comparable, suggesting their combined use for holistic insights. This master thesis offers a strategic pathway for organizations aiming to refine their cybersecurity strategies by leveraging shared data insights.

In der vorherrschenden digitalen Ära, die durch eine zunehmende Anzahl von Cyberangriffen gekennzeichnet ist, steht die Cybersicherheit als zentrales Anliegen für Organisationen aller Größen im Vordergrund. Obwohl die Notwendigkeit, Cybersicherheitsmaßnahmen zu verstärken, allgemein anerkannt ist, stellt die Bestimmung einer optimalen Strategie eine komplexe Herausforderung dar. Diese Masterarbeit stellt einen neuartigen Ansatz vor, der den Austausch von Cybersicherheitsdaten zwischen Unternehmen nutzt, um Organisationen bei der Verbesserung ihrer Abwehrmaßnahmen zu unterstützen. Es wurde ein Tool entwickelt, um die Relevanz von informationsaustauschenden Einheiten zu ermitteln, indem Unternehmen in drei Dimensionen klassifiziert werden: geschäftlich, wirtschaftlich und technisch. Jede Dimension wird durch eindeutige Faktoren definiert, was einen präzisen Vergleich ermöglicht. Eine begleitende Anwendung wurde konzipiert, um die Ähnlichkeiten zwischen Unternehmen anhand des euklidischen Abstands und der Pearson-Korrelation darzustellen. Durch eine umfangreiche Bewertung erwies sich der euklidische Abstand in den geschäftlichen und wirtschaftlichen Bereichen als überlegen. Bei der technischen Dimension, die von ganzzahligen Werten dominiert wird, war die Wirksamkeit beider Maße vergleichbar, was auf ihren kombinierten Einsatz für ganzheitliche Einblicke hindeutet. Diese Masterarbeit bietet einen strategischen Weg für Organisationen, die ihre Cybersicherheitsstrategien durch die Nutzung geteilter Dateninformationen verfeinern möchten.

Acknowledgments

First of all I feel the need to thank my supervisor Dr. Muriel Franco for his regular assistance, our in-depth discussions and his very helpful inputs throughout this thesis.

I would also like to thank my co-supervisor, Jan von der Assen for his support.

Finally, I would also like to thank Prof. Dr. Burkhard Stiller, head of the Communication System Research Group (CSG) at University of Zurich, for giving me the possibility to write my master's thesis about such an interesting topic.

Contents

Chapter 1

Introduction

In today's digital age, cybersecurity has assumed a paramount role. With the increasing connectivity of Internet-enabled devices, the vulnerability to cyberattacks has also escalated. The emergence of the Internet of Things (IoT), while enhancing automation and process improvements, has simultaneously presented attackers with new entry points. These cyber threats have swiftly become the norm across both public and private sectors, ranking as the fifth-greatest risk in 2020 [14]. Furthermore, this risky landscape is expected to expand further in 2023, with IoT cyberattacks projected to double by 2025 [14].

Based on data from 2022, it is important to highlight the significant impact of cyberattacks and their widespread occurrence. Globally, a substantial number of websites, approximately 30,000, are targeted and hacked each day, underscoring the scale of this threat. Furthermore, an alarming 64% of companies worldwide have experienced at least one form of cyberattack, indicating the pervasive nature of these incidents. Email serves as a prominent channel for malware distribution, with approximately 94% of all malicious software being distributed through this medium. The frequency of cyberattacks is relentless, with an average of one attack occurring every 39 seconds, posing continuous challenges for web security. In 2021 alone, a staggering 22 billion records were breached, highlighting the magnitude of data compromises that organizations face [8]. It is crucial to note that successful cyberattacks often result in substantial financial costs, as exemplified by the average expense of $ 4.35 million per data breach in 2022 [28].

To ensure protection against these cyberattacks, there is a wide array of available solutions, encompassing firewalls, antivirus programs, encryption, security monitoring, physical security measures, and backups. Nevertheless, navigating through the multitude of security systems to make the optimal choice is not a straightforward task. Complicating matters further, Small- and Medium-sized Enterprises (SMEs) often operate within constrained cybersecurity budgets, necessitating judicious and efficient allocation of resources for cybersecurity investments. A technology that specifically addresses this concern is the SECAdvisor Tool [22]. However, for smaller businesses, effectively harnessing the tool's potential proves to be challenging due to the prerequisite cybersecurity knowledge required for interpreting numerous input parameters. Consequently, SMEs would greatly benefit

from the ability to benchmark their operations against similar companies and draw insights from shared experiences and information. This very topic is being addressed within the scope of this thesis.

1.1 Motivation

Companies are dedicating significant resources to safeguarding against the escalating threats by investing in robust cybersecurity measures. Projections indicate that the cybersecurity sector will witness a substantial investment of $188.3 billion in 2023, marking an impressive 11.3% growth compared to the previous year [10]. Merely increasing investments in cybersecurity is insufficient for a company to effectively safeguard itself against the expanding cyber threat landscape. It is crucial that these investments are meticulously planned and strategically executed [19, 17].

To derive optimal security measures, it is essential to first assess and validate the risks involved. This entails conducting a thorough analysis to identify the vulnerable components within a company that are prone to successful cyberattacks, as well as determining the areas that should be safeguarded to the highest degree possible. An integral part of this analysis involves evaluating the value and susceptibility of each business component, enabling informed decisions regarding targeted investments.

In the domain of cybersecurity, a range of cybersecurity economics and tools have emerged, each with the overarching goal of aiding companies in making judicious and effective investments in safeguarding their digital infrastructure. However, a critical challenge persists, as there is a dearth of comprehensive tools and metrics that enable companies to make informed cybersecurity investment decisions by leveraging shared information and insights derived from other organizations [44].

By embracing the practice of sharing cybersecurity information, companies can tap into the valuable experiences and insights of their peers, enabling them to make more calculated and effective cybersecurity investments. Furthermore, the ability to identify and learn from similar companies allows organizations to adapt their security measures based on the shared information, fostering a proactive and adaptive cybersecurity approach. It is important to emphasize that the benefits of shared information extend beyond the immediate recipient. The company sharing its cybersecurity knowledge also gains from a larger and more robust database, which in turn provides more accurate and comprehensive data for informed decision-making. Thus, fostering a culture of cybersecurity information sharing is crucial in enabling organizations to make more precise and informed investments in their security measures.

1.2 Description of Work

The goal of this Master's Thesis is to enable information sharing between companies. This begins by understanding and mapping the diverse types of information required to

accurately characterize data segments within businesses for the appropriate application of economic models. In subsequent steps, factors specific to different dimensions (*e.g.*, technical, economic, and business) are defined, aiding in the comparison of companies across these dimensions. The ability to compare businesses is crucial, as it leads to the identification and data utilization from comparable companies. Correlation measures, algorithmic methods designed for such comparisons and the identification of similar companies across various sectors, are established for this process. This information is then stored and shared in a structured format. The developed application permits the sharing of information about the economic impacts of cyberattacks and the configurations of cybersecurity economic models between companies. Moreover, it employs anonymization techniques, ensuring that information can be shared without revealing sensitive company-specific data. This application is subsequently integrated into the SECAdvisor tool. This integration allows for the import and export of company-related information, facilitating sharing of information between companies to be used for their customized models, including the sharing of the custom breach probability function, a crucial factor in calculating the optimal investment amount.

Following the creation of the prototype, an evaluation, embodying real-world scenarios, is conducted. This assessment showcases the utility and precision of the designed application, illustrating its role in facilitating efficient information sharing between companies. By doing so, it enables companies to make more accurate investments in cybersecurity, thereby reinforcing its practical relevance and effectiveness.

1.3 Thesis Outline

The structure of the remaining work unfolds as follows. Chapter 2 furnishes the theoretical foundation for this study. This is followed by Chapter 3, which illuminates related works. Chapter 4 then discloses the approach, emphasizing the methodology and metrics implemented for the prototype. Subsequently, Chapter 5 provides a detailed insight into the technical elements and technologies engaged during the creation of the prototype. Chapter 6 offers a comprehensive evaluation of the work. Chapter 7 concludes the thesis and provides suggestions for future work.

Chapter 2

Background

This section offers fundamental information crucial for understanding this work. It initially introduces the concept of correlation measures, particularly concentrating on two frequently used types. Subsequently, it delves into the theme of information sharing, defining what it entails and the process involved. Comprehending these notions is essential to fully grasp the broader implications of the work.

2.1 Correlation Measures

Correlation measures serve as valuable statistical techniques employed to quantify and assess the relationship or association between two variables. Their purpose is to provide a numerical value that effectively captures the strength and direction of the relationship exhibited by the variables in question. Within the realm of correlation analysis, a multitude of diverse measures exist, including but not limited to the Euclidean Distance, Pearson Correlation, Manhattan Distance, Cosine Similarity, Hamming Distance, and Minkowski Distance. Despite their shared objective of determining similarities between variables, each correlation measure adopts a distinct approach in accomplishing this goal. In the scope of this work, particular emphasis will be placed on explaining and elucidating the characteristics and properties of the Euclidean Distance and the Pearson Correlation [7, 25].

2.1.1 Pearson Correlation

The Pearson correlation coefficient (r) is a widely used measure of linear correlation between two variables. It ranges from -1 to 1 and indicates both the strength and direction of the relationship. It is a descriptive statistic that summarizes the characteristics of a dataset by quantifying the strength and direction of the linear relationship between two quantitative variables [52].

Table 2.1 offers insights into the connection between the values of the Pearson correlation coefficient and their corresponding strength and direction. It elucidates that a coefficient

of 0.5 or greater signifies a strong and positive correlation, indicative of a pronounced relationship. As the coefficient progressively diminishes, the strength of the relationship gradually wanes, ultimately reaching a coefficient of 0, which implies the absence of any discernible positive or negative correlation. Subsequently, as the coefficient ventures into the negative range, it exposes a negative correlation between the compared values, suggesting an inverse association. Therefore, when the correlation coefficient approximates 1 (Figure 2.1 left), the comparison values demonstrate a remarkable degree of similarity, while values hovering near -1 reflect the utmost dissimilarity (Figure 2.1 right), highlighting the presence of disparate characteristics or traits.

Table 2.1: Pearson Correlation Strenghts [52]

Pearson correlation coefficient (r) value	**Strength**	**Direction**
Greater than 0.5	Strong	Postive
Between 0.3 and 0.5	Moderate	Positive
Between .3 and .5	Weak	Positive
0	None	None
Between 0 and –0.3	Weak	Negative
Between –0.3 and –0.5	Moderate	Negative
Less than –0.5	Strong	Negative

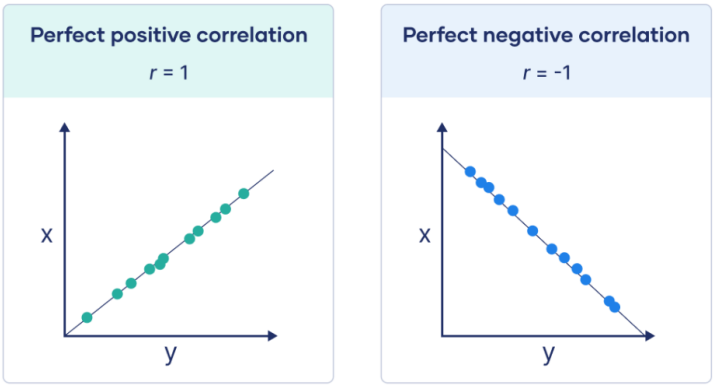


Figure 2.1: Pearson Extreme Correlations [52]

Equation 2.1 illustrates the calculation process for the Pearson correlation coefficient (r). Here, n stands for the aggregate number of data points. The individual data points for variables x and y are signified by x_i and y_i, respectively. Furthermore, $\bar{x}$ and $\bar{y}$ are representations of the mean values corresponding to variables x and y, respectively.

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{2.1}$$

The Pearson correlation presents certain benefits, such as its straightforward computation and interpretation, and the fact that it doesn't necessitate any alterations to the variables.

It facilitates the exploration of relationships in their inherent context and signifies the strength and direction of correlation between variables. However, this method also has its shortcomings. It is limited to assessing linear relationships and is particularly sensitive to outliers, implying that extreme values can drastically distort the outcomes. It presupposes the normal distribution of data and only indicates association, not causation. The Pearson correlation might also perform poorly when the span of observations is restricted [23, 52, 32, 47].

2.1.2 Euclidean Distance

The Euclidean distance is a measure of the straight-line distance between two points in Euclidean space and is derived from the Pythagorean theorem. The Euclidean distance between two points can be calculated by considering their Cartesian coordinates and using the Pythagorean theorem. In simpler terms, it represents the length of a line segment connecting two points in space [9].

Equation 2.2 and Figure 2.2 represent the Euclidean distance calculation between two points. In this equation, the coordinates of the first point are denoted as *(x1, y1)*, while the coordinates of the second point are represented as *(x2, y2)*. The resulting distance between these two points is represented by the variable *d*. By substituting the respective coordinate values into the equation, the Euclidean distance between the points can be computed, providing a measure of their straight-line separation in Euclidean space.

$$d = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} \tag{2.2}$$

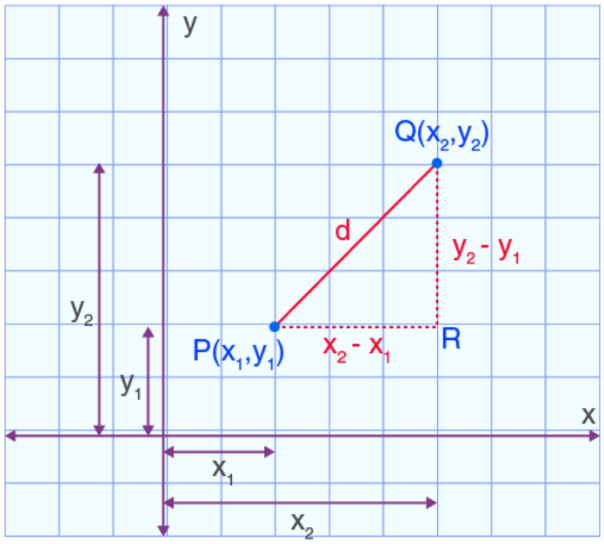


Figure 2.2: Euclidean Distance Calculation [9]

The Euclidean distance has several advantages and disadvantages. On the positive side, it offers simplicity, making it easy to understand and implement. It can be applied to various data types and is widely used in clustering, k-nearest neighbors, and distance-based

classification. Euclidean distance also has an intuitive interpretation as the straight-line distance between two points, aligning with our understanding of physical space. Additionally, it is computationally efficient, especially in low-dimensional spaces. However, there are limitations to consider. Euclidean distance is sensitive to variable scale and may require standardization or normalization. It is less effective in high-dimensional spaces due to the curse of dimensionality, where distances become less informative [15, 25].

2.2 Information Sharing

Information sharing represents a critical process in today's interconnected society, encompassing the exchange of data among a myriad of entities such as individuals, organizations, and technologies. This process has seen tremendous acceleration due to advancements in technology, which have enabled the widespread distribution of networks, intranets, cross-platform compatibility, application porting, and standardization of IP protocols. The content of this exchange can vary significantly, spanning from personal videos to organizational feeds and software interactions, all of which are carried out under the safeguard of appropriate permissions [50].

The effectiveness of information sharing finds significant importance across a broad range of sectors, with different sectors deploying different scopes and applications as per their requirements. In the realm of cybersecurity, information sharing becomes a vital strategy for Information Sharing and Analysis Organizations (ISAOs), aiding in risk management and mitigation of cybersecurity threats [30]. In healthcare, it promotes the security, safety, and resilience of the sector by providing stakeholders with an understanding of current threats and incidents, thereby facilitating effective mitigation strategies [27]. Likewise, in organizational communication, information sharing enhances performance, cultivates a collaborative culture, and assists in achieving organizational goals [26]. Thus, the establishment of effective information sharing practices holds paramount importance in the successful operation of various sectors and institutions in today's digital era.

2.2.1 Data Processing

In today's interconnected business landscape, data has emerged as a valuable asset, and sharing it with other companies can unlock numerous benefits and opportunities. However, the process of sharing data requires careful consideration to ensure its quality and privacy. This subsection explores the essential steps involved in processing data that will be shared with other companies, highlighting the importance of data assessment, cleaning, anonymization, transformation and its security.

The initial step in preparing data for distribution involves a thorough evaluation of its quality and structure. This means understanding the origin of the data, its format, and identifying any potential issues that could affect its usability. This understanding aids in pinpointing areas that require enhancements or modifications. The process of data cleaning and preprocessing plays a pivotal role in augmenting the reliability and

compatibility of the data set. This involves meticulously removing duplicates, rectifying any errors, and addressing gaps in data through suitable techniques. These steps are vital in maintaining the integrity of the data. As a result, the receiving companies can derive accurate insights and make informed decisions based on this cleaned data [45].

Respecting privacy is paramount when sharing data. Anonymization or de-identification techniques are employed to protect personally identifiable information (PII) and sensitive data. By removing or obfuscating direct identifiers, such as names or social security numbers, the shared data retains its value while safeguarding individuals' privacy. Striking a balance between data utility and privacy is crucial to build trust among companies and ensure compliance with relevant data protection regulations [29].

As an additional step, it is necessary to package the data into suitable formats, which may include Excel files, CSV files, or TAB-delimited text files. When it comes to encoding date and time variables, it's recommended to follow the ISO 8601 guidelines. It's also vital to incorporate a comprehensive codebook that clearly defines variables, their respective units, selection summaries, and the experimental study design. This codebook enhances understanding by elaborating how each variable is coded, aligning with its data type [13].

After deciding on the structure in which the information will be saved, it's crucial to identify the location for data storage. This step should involve ensuring that the stored data is safeguarded from unauthorized intrusion or inappropriate usage. The security of storage is as critical as the manner in which the data is transported. The incorporation of security strategies such as encryption, access controls, and safe data transfer protocols is fundamental in providing comprehensive protection to the data throughout its exchange procedure [40].

In summary, the appropriate organization and preparation of data destined for dissemination is incredibly crucial. The process involves guaranteeing that only the selected data for release is circulated, and that this data is stored in a precise manner, adhering to a certain standard or format. Furthermore, it's necessary to verify that this data is free of any sensitive or confidential company-specific information, to protect the firm's privacy and security. Lastly, the usage of a safe and secure protocol for the data transfer process is also an essential requirement, ensuring the entire process's integrity and security.

2.2.2 Information Sharing Incentives

Incentivizing companies to share data requires a multifaceted approach that addresses both technical and trust-related barriers. First, it is crucial to invest in technological solutions that allow for secure and regulated data sharing. The evolving data ecosystem has led to the development of various tools and platforms that facilitate this process, ranging from data exchanges to APIs and secure data vaults [16, 36]. By offering these solutions, we can mitigate some of the risks associated with data sharing and help organizations feel more comfortable engaging in this practice.

Second, data sharing should be framed as an opportunity for mutual benefit, leading to unrestricted innovation and increased market value. Companies, dubbed as "data

masters", that understand the potential of data ecosystems tend to make significantly better decisions, leading to the delivery of smarter products and services. By emphasizing these potential rewards and presenting successful case studies of data sharing, companies might be more willing to participate in the data economy. However, the challenges such as regulatory compliance, trust, and privacy concerns must be addressed in tandem [34].

Another key aspect to incentivize data sharing is implementing the appropriate incentives. This can include financial benefits, access to shared resources, or public recognition. For example, in the healthcare industry, some organizations have adopted pay-for-performance programs that reward data sharing [12]. Similar incentives could be utilized in other sectors. Further, companies might be encouraged to share their knowledge when there is uncertainty about market opportunities, as sharing such information could lead to more buy-in from other industry players [38].

Transparency and trust-building are other essential components of incentivizing data sharing. Clear communication around how data will be used, who will have access to it, and what measures are in place to protect it can help alleviate concerns. For instance, in the wake of the General Data Protection Regulation (GDPR), companies are being encouraged to be as transparent as possible about how customer data is handled [54]. The same principle applies to inter-company data sharing. A strong framework for data rights and governance, possibly supported by legislation or industry standards, can also increase trust and willingness to share data [49].

Lastly, data sharing should be incorporated into broader business and societal goals. Aligning data sharing with overarching objectives like sustainability, innovation, or solving complex societal challenges can provide a strong motivation for companies to participate [16]. Data sharing is essential in addressing many of the UN's sustainable development goals (SDGs) [50]. In this sense, companies that contribute their data towards solving these complex problems can benefit from the positive public image, potential for new partnerships, and alignment with their corporate social responsibility (CSR) goals. All these considerations together can help create a conducive environment for companies to share data with each other.

Chapter 3

Related Work

In the forthcoming chapter, an extensive overview of related works in the domain of distance measurement is presented, which serves as a pivotal tool for discerning and exploring the similarities among various objects. Moreover, a examination of the related works conducted in the field of information sharing is undertaken to to gain insights into how valuable shared information can be.

3.1 Application of Correlation Measures

The realm of distance measurement encompasses a plethora of distinct algorithms that facilitate the comparison of distances among diverse objects, allowing for informed assessments regarding their relative similarities. Within the scope of this thesis, two specific algorithms are examined, which have often been utilized in related studies. The following section presents a comprehensive compilation of relevant works that have employed either the Euclidean distance or Pearson correlation to unravel insights and establish connections within their respective domains.

In [2], the authors utilized the Pearson correlation and explores the application of machine learning algorithms in intrusion detection systems for cybersecurity. The paper recognizes the effectiveness of machine learning models in detecting anomalies and enhancing security using comprehensive datasets with various attack types. However, the high dimensionality of these datasets poses challenges in extracting relevant information due to time and space complexity. To address these challenges, the paper proposes the Dynamic Feature Selector, which selects pertinent features from the dataset to improve the prediction potential of machine learning models in cybersecurity. The Pearson correlation is employed to express the correlation between features. This solution aims to reduce the dimensionality of the dataset and enhance the efficiency and effectiveness of machine learning algorithms in intrusion detection systems.

The work [4] emphasizes the significance of the Euclidean algorithm within the context of RSA cryptography. The algorithm plays a crucial role in efficiently calculating the Greatest Common Divisor (GCD) of two numbers. By repeatedly dividing the larger

number by the smaller number and considering the remainders, the Euclidean algorithm enables the determination of the GCD. This approach proves to be more efficient than direct prime factorization, especially for large numbers. The Euclidean algorithm's effectiveness is particularly vital in the RSA algorithm, where it contributes to key generation, encryption, and decryption processes. The work underscores efficiency of the Euclidean algorithm and highlighting its integral role in ensuring the security and effectiveness of cryptographic operations in RSA.

In another work, [33] focuses on cancelable biometric schemes, aiming to address privacy concerns in biometric systems. The paper acknowledges the importance of biometric systems in authentication and identification, while also recognizing the privacy issues associated with storing and handling sensitive biometric data. To overcome these concerns, the paper proposes transforming the biometric data into cancelable templates. Specifically concentrating on the Euclidean metric, which measures the distance between points in a multi-dimensional space, the authors present a method for generating cancelable templates that preserve the Euclidean metric. This ensures that the distance between two cancelable templates accurately reflects the similarity of the original biometric data. By leveraging the Euclidean metric, the proposed scheme offers a reliable means of assessing similarity in a privacy-preserving manner, contributing to the development of secure and privacy-enhanced biometric systems.

Roy et al. [5] present their research on the development of a specialized intrusion detection system (IDS) for Internet of Things (IoT) networks in the paper titled "Anomaly-Based Intrusion Detection System for IoT Application". This work aims to propose an anomaly-based IDS that focuses on real-time identification of network traffic behavior to mitigate cyber attacks in IoT networks. The authors highlight the utilization of the Pearson correlation coefficient in a feature selection algorithm to enhance the efficiency of intrusion detection. By analyzing the correlation between features extracted from network traffic data, the algorithm aims to identify relevant features that contribute to effective anomaly detection. The use of the Pearson correlation coefficient enables the IDS to detect patterns and deviations from normal behavior in IoT network traffic by identifying relationships and dependencies between variables. The proposed IDS, with its anomaly-based approach and integration of the Pearson correlation coefficient, is recognized as an effective means of securing IoT networks by promptly detecting and responding to potential cyber threats, thereby ensuring network security and integrity.

The work [31] which made use of the Pearson correlation is "Correlation-Based Anomaly Detection in Industrial Control Systems". This paper introduces a correlation-based approach for detecting anomalies in Industrial Control Systems (ICS), which are crucial for critical infrastructure sectors. The work propose leveraging the Pearson correlation coefficient to measure the linear relationship between variables within the ICS environment, aiming to identify abnormal patterns and deviations from expected behavior that may indicate security breaches or system malfunctions. By calculating correlations between different process variables, the approach can effectively detect known and unknown anomalies, including complex attack scenarios that traditional methods might overlook. The paper highlights the implementation of a correlation matrix that represents the correlation values between process variables, serving as the foundation for anomaly detection. Deviations from normal correlation patterns are flagged as potential anomalies. The

correlation-based approach offers advantages such as capturing coordinated attacks involving multiple variables, adaptability to changing system behavior, and proactive defense through early detection and mitigation of potential threats.

Hiroki Miyahara [42] explores how Euclidean distance can be used to measure the similarity between company exposures, particularly in analyzing and comparing their risk profiles. The objective is to propose a method for quantifying the similarity between companies based on their exposure to different risk factors. Each company is represented as a vector of exposures, and Euclidean distance is used to calculate the dissimilarity or similarity between two companies exposure vectors. A smaller distance indicates a more similar risk profile, while a larger distance suggests greater dissimilarity. The work emphasize the significance of Euclidean distance for similarity analysis due to its straightforward interpretation and wide applicability. It's also discuss its utilization in clustering analysis, which helps group companies with similar risk profiles together, enabling the identification of patterns and aiding in portfolio management and risk assessment.

An further work [41] explores the use of Euclidean distance as a measure for comparing complex networks, highlighting its simplicity and effectiveness compared to more complex methods. It acknowledges the existence of advanced techniques but emphasizes their computational complexity and limitations for large-scale network analysis. In contrast, Euclidean distance offers a straightforward and efficient approach by representing networks as vectors or matrices. The Euclidean distance provides an intuitive and interpretable measure, easily computed and applicable to networks of varying sizes and topologies. The authors present a case study comparing real-world networks, demonstrating that Euclidean distance effectively captures network dissimilarity and reveals meaningful differences.

Jong-Ho Lee [39] introduces a DNN-based approach for efficiently evaluating the minimum Euclidean distance between patterns. He highlight the widespread use of the Euclidean distance metric in quantifying pattern similarity but acknowledge its computational complexity, particularly for large datasets. To address this challenge, the work propose a deep neural network architecture that learns to approximate the minimum distance. By training the network on known distances, it develops an understanding of patterns and relationships. The DNN consists of input, hidden, and output layers, with various activation functions and optimization techniques employed for effective training. Once trained, the DNN can efficiently evaluate the minimum distance between new patterns, reducing computational complexity compared to traditional methods. Experimental results confirm the effectiveness of the proposed approach, showcasing accurate evaluations and computational efficiency. The DNN-based approach provides an effective solution for distance evaluation tasks in pattern recognition, classification, and clustering.

The paper "Euclidean Distance-based Machine Learning Scheme to Detect Vehicle Hacking Cyber-Attacks" [1] presents an innovative approach to tackle the rising concern of cyber-attacks on vehicles. With the objective of enhancing vehicle security and safeguarding driver safety and privacy, this work introduce a machine learning scheme that utilizes the Euclidean distance metric. By collecting and preprocessing a dataset of vehicle sensor readings, relevant features are extracted to capture the vehicle's behavior and characteristics. These features serve as inputs for a supervised learning algorithm, which

calculates the Euclidean distances between data points to distinguish between normal behavior and cyber-attack instances. The model optimizes its internal parameters during the training phase to minimize classification errors and learn patterns associated with cyber-attacks.

MENTOR [18] is a cybersecurity tool developed by IFI. Its purpose is to combat the swiftly growing menace of cyberattacks and the corresponding difficulties in choosing efficient security measures. MENTOR's primary function is to assist network managers and end-users in choosing the most suitable security service, tailored to their specific needs and scenarios. It achieves this by employing four distinct measurement methods - Euclidean distance, Manhattan distance, Cosine similarity, and Pearson correlation. These metrics allow MENTOR to accurately determine and recommend the best security service. MENTOR consists of several components, including the Service Requestor, Extractor, Classifier, Service Aggregator, and Retriever, which collectively contribute to its robust functionality. An essential feature of MENTOR is its role in simplifying the adoption of advanced cybersecurity solutions, making it a valuable asset in the ever-evolving landscape of cyber threats. MENTOR has been incorporated into the ProtectDDoS[20] tool, which suggests DDoS protection measures. Moreover, MENTOR is also integrated into SECAdvisor[44], which assists companies in determining the extent of their cybersecurity investments and subsequently recommends cybersecurity protections.

Table 3.1 offers a comprehensive summary of the related works, categorizing them based on several parameters. These include the particular correlation measure that was employed, the specific field or area in which the correlation measure was applied, and the dataset utilized in each research study. Upon examining the table, it is abundantly clear that the two correlation measures, Euclidean distance and Pearson correlation, are particularly prevalent as tools for identifying and quantifying similarities amongst diverse datasets. The use of these correlation measures is not confined to a singular field of study or application. On the contrary, they have been implemented across a wide spectrum of areas, underlining their versatile nature and broad applicability. One area where the application of these correlation measures is especially pronounced is in the field of cybersecurity. A careful analysis of the table reveals that both the Euclidean distance and Pearson correlation measures have been extensively used in numerous cybersecurity studies, demonstrating their significant role in enhancing the security measures within this increasingly important domain.

Table 3.1: Correlation Measures Comparison

Work	**Description**	**Correlation Measure**	**Area**	**Dataset**
[2]	Applies machine learning algorithms in cybersecurity for intrusion detection.	Pearson	Cybersecurity & Machine Learning	NSL-KDD & KDD'99
[4]	Shows the importance of the Euclidean algorithm in RSA cryptography for efficiently calculating the Greatest Common Divisor (GCD) of two numbers.	Euclidean	Cybersecurity	GCD & Prime numbers
[33]	Cancelable biometric schemes to address privacy concerns in biometric systems.	Euclidean	Cybersecurity & Business	Biometric data
[5]	Present a study on a specialized intrusion detection system (IDS) for IoT networks that uses real-time analysis of network traffic.	Pearson	Cybersecurity & Computer Network	NSL-KDD, CICIDS-2017 and IOTID20
[31]	A correlation-based approach is proposed for identifying anomalies in critical Industrial Control Systems (ICS).	Pearson	Cybersecurity & Computer Network	ICSs
[42]	Quantifies the similarity between companies' risk exposures using Euclidean distance.	Euclidean	Business	Company Informations
[41]	Demonstrates the use of Euclidean distance as an efficient and intuitive measure for comparing complex networks.	Euclidean	Computer Network	Brain connectivity & Controls subjects
[39]	Presents a deep neural network (DNN) based method to efficiently approximate the minimum Euclidean distance between patterns.	Euclidean	AI	-
[1]	Enhances vehicle security by identifying cyber-attacks.	Euclidean	Business	Vehicle sensor readings
[18]	Recommends cybersecurity solutions.	Euclidean & Pearson	Cybersecurity	Cybersecurity services

3.2 Information Sharing Efforts

This subsection provides an overview of literature available on the subject of data sharing practices among companies. It highlights the prevailing patterns and trends regarding the types of data that are frequently exchanged between these entities. Additionally, the subsection examines a range of research studies that explore the diverse challenges and benefits associated with information sharing within the corporate landscape.

In the area of cybersecurity, companies exchange data on identified or potential cyber threats, enabling other organizations to anticipate and thwart similar attacks [53, 48]. They also share specifics of past breaches, encompassing the nature of the attack, the affected systems or networks, and the countermeasures undertaken. This collaborative exchange facilitates learning from past incidents, helping to bolster defenses across the board. This sharing ethos extends to the implementation of programs and services designed to protect critical infrastructures and advance cybersecurity, such as the Automated Indicator Sharing (AIS) [11]. AIS facilitates real-time sharing of machine-readable cyber threat indicators and defense strategies across public and private sector entities, fostering an ecosystem of cyber protection. By promoting information sharing and providing participants with immediate insights, AIS helps to mitigate the impact of cyberattacks. The AIS community is diverse, featuring private sector entities, different levels of government, information sharing organizations, and international collaborators.

A research study [46] focused on the challenges in sharing information across various sectors including extractives, fintech, transport, healthcare, and environmental industries. The study highlighted that while data is vital for new businesses and technologies, many face difficulties in determining the appropriate methods to access and utilize data. Key challenges discovered were related to data discovery due to limited and unstructured data available, issues around data control and access due to the presence of multiple stakeholders, problems regarding trust in the data's quality and its source, resistance to transparency fearing negative implications, lack of understanding of business models that encourage data sharing, and the struggle to use personally identifiable data legally and effectively.

A survey of 1,000 British businesses found that a majority (68%) were not prepared to open access to their own non-personal datasets, despite acknowledging that such data sharing could yield commercial benefits. The key barriers to data sharing were rooted in concerns about corporate privacy, with 43% of respondents citing this as a primary reason. The need to protect intellectual property was another major obstacle, with 32% of businesses highlighting this. The risk of online data mismanagement, which could lead to poor quality information and a subsequent loss of data value, was also a significant concern for 29% of the businesses surveyed. Despite these companies' advocacy for open public sector data, they perceived substantial risks in opening up their own datasets, demonstrating a disconnect between their expectations of public data transparency and their readiness to participate in similar initiatives [3].

The paper "Information sharing in the context of business cooperation – as a source of competitive advantage [43] underscores the role of information sharing between companies as a catalyst for competitive advantage. By sharing timely and relevant data, businesses

can streamline decision-making processes, increase operational efficiency, and gain critical insights into market trends and customer preferences. This paves the way for rapid and effective responses to market fluctuations, recognition of new business opportunities, and development of innovative solutions. Furthermore, information sharing fosters a culture of trust and mutual understanding, while safeguards must be implemented to ensure information security during the exchange process.

Gordon, Loeb et al. examine the link between information sharing and cybersecurity investment decisions and present a clear case for the benefits of information sharing. The study illuminates how information sharing fosters a shift from a reactive to a proactive stance in cybersecurity investments for firms. It mitigates the inclination to defer such crucial investments, a behavior often intensified by the uncertainties surrounding cyber threats. Through an analytical real options framework, the research reveals how information sharing can actually enhance the value of the embedded option to invest in cybersecurity, especially in view of the increasing uncertainty about the costs of attacks. The authors also highlight the business case for sharing security information, noting that it is a strategic complement to investment in security technologies [24].

In another approach, decision making based on averages is investigated. While averages serve as an easy tool to comprehend complex data by providing a central trend or measure, their usage often presents significant downsides. They risk disregarding the data's variability, distribution skewness, and the influence of outliers, thus misrepresenting the true nature of a dataset. Averages are especially misleading in non-linear models where the average output doesn't align with the output of average inputs, and they tend to underestimate risk in uncertain situations. They also neglect the actual distribution of uncertain variables, potentially leading to incorrect or suboptimal decisions. Inferring from this work, it could be asserted that sharing information serves as a more effective substitute for employing averages [37].

In reviewing relevant literature, it becomes evident that an array of information is consistently shared between corporations. More specifically, through such information exchange in the realm of cybersecurity, firms are equipped to bolster their defense against cyber threats, thereby ensuring cybersecurity investments aren't merely grounded on mean values. Nonetheless, this process of information sharing isn't devoid of challenges, with privacy concerns being a major stumbling block. The apprehension stems from the potential exposure of sensitive data to unintended parties through this sharing. Consequently, it is crucial to maintain a strategic approach when sharing information between companies, ensuring that only data intended for release is shared, and that company-specific information is as anonymized as possible, thereby mitigating the risk of sensitive data breaches.

Chapter 4

Approach

This Master Thesis proposes an structured and collaborative approach to address the problem of lack of data in cybersecurity field. The approach is proposed to facilitate the identification of comparable companies from different perspectives (*e.g.*, technical, economic, and business) and enables data sharing among them. This approach allows companies to actively seek out companies exhibiting similar characteristics, thus, being able to have a tailored understanding of cybersecurity-related information and investments without relying only on average companies. Once such a company is found, the tool provides the capability to access and utilize the data shared by this similar entity. By absorbing and analyzing this shared data, a company can modify and enhance its cybersecurity investments, measurements, and cybersecurity metrics, using insights derived from similar organizations to adapt their cybersecurity strategy and operations effectively. For that, different correlation measurements (*e.g.*, Pearson and Euclidean distance) have been explored to suggest companies that have more similarities according to specific perspectives and needs.

A full-fledged prototype was designed and systematically developed as a Proof-of-Concept (PoC), reinforcing the feasibility of the proposed scheme. The development of the PoC, including its design rationale and developmental process, is described in detail in Chapter 5. In addition to this, an exhaustive evaluation was conducted to critically assess the utility and performance of the developed prototype, the results and analysis of which are presented and explored within Chapter 6.

Figure 4.1 shows the process from raw corporate information to distinct clusters of comparable companies and the process of consuming shared information from a specific company. The foundation upon which both processes are based is a collection of various companies, as described in *Point 1.* A dataset, featuring a variety of businesses and their properties, is utilized. This dataset is extracted from the Real Cyber Value at Risk (RCVaR) model [21]. However, these data do not fully cover all necessary corporate attributes. As a solution, the *Data Generator* expands these data with additional features, as portrayed in *Point 1.1.* This enhanced dataset with corporate information now provides the foundation for finding similar companies and consuming shared corporate information.

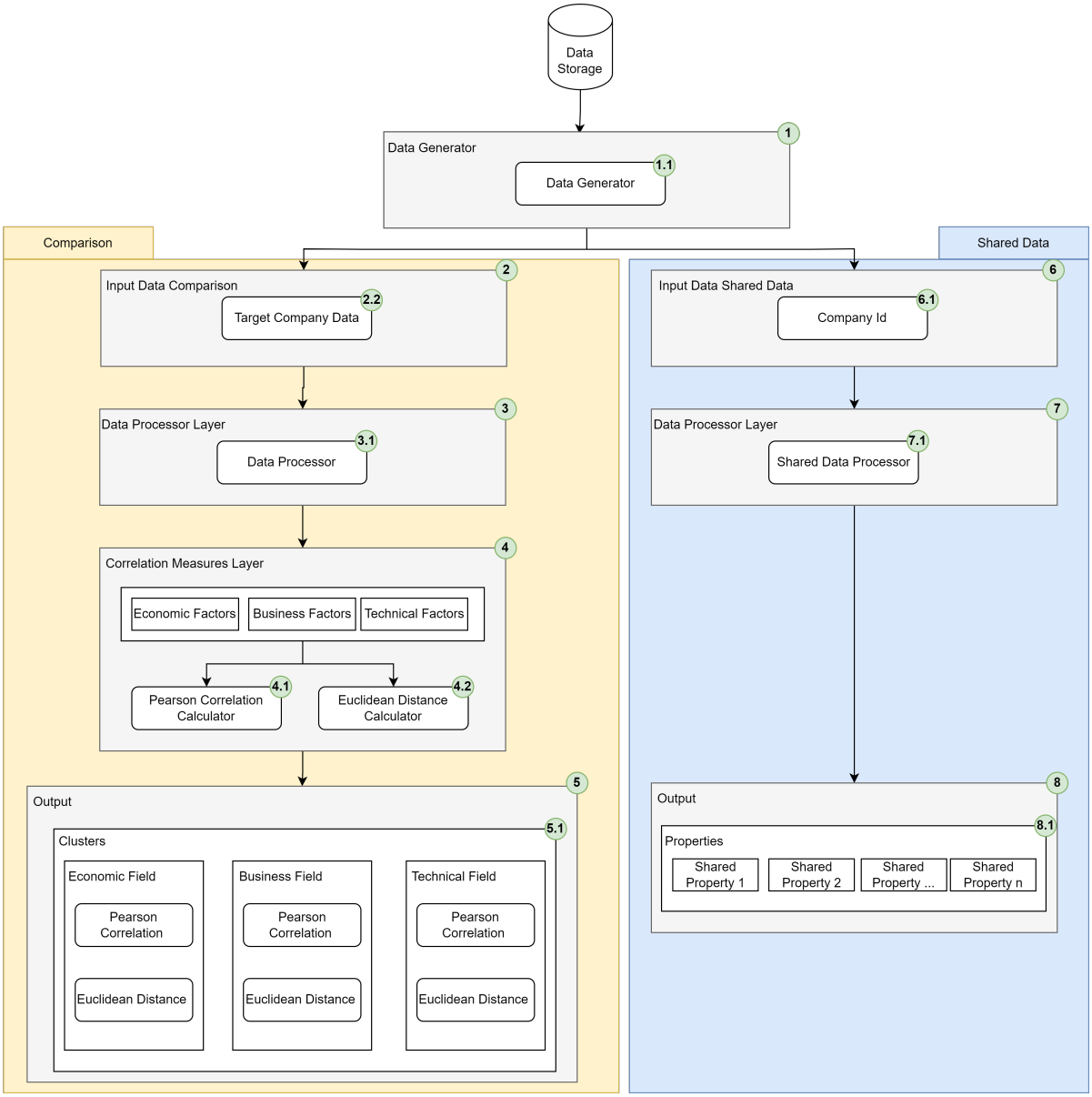


Figure 4.1: Conceptual Architecture of the Approach

Point 2 represents the input information required to find similar companies. It can be seen in *Point 2.1* that in order to compare various companies, the information of the company for which similar companies are to be found is needed. Upon collection, the input data is converted into an appropriate format for continued processing, as shown in Point 3. Following this, *Point 4* focuses on computing the similarity between various companies in comparison to the targeted company, for which similar matches are being identified. This computation involves two distinct correlation measures: the Pearson Correlation and the Euclidean Distance, as depicted in *Points 4.1 and 4.2* correspondingly. These two algorithms are applied in three separate dimensions. The first one is the economic dimension, incorporating economic factors (*e.g.*, cybersecurity investment, and cybersecurity budget) that facilitate an economic analysis of companies. The second one is the business dimension, characterized by its unique business factors, such as amount of revenue or organization size. Finally, the technical dimension (*e.g.*, cloud solution and

network infrastructure) is also taken into account. As a result, for each of the stated dimensions, the similarity for each company is ascertained based on both the Pearson Correlation and the Euclidean Distance.

As a result, a dataset is generated where the Pearson Correlation and the Euclidean Distance for every company are computed for each of the three dimensions. The resultant cluster, represented in *Point 5.1*, provides the opportunity to locate similar businesses within these dimensions. Additionally, it enables the discovery of like companies by utilizing either the Pearson Correlation or the Euclidean Distance.

Once a company has been identified from which the main company wants to receive shared information, the second process, represented on the right side of Figure 4.1 and labeled as *Shared Data*, begins. To access the shared information of a company, the ID of that company needs to be transmitted, which is depicted in *Point 6.1*. This represents the input information. The role of the *Data Processor Layer*, indicated as *Point 7*, is to transmit only those company-specific details that the company has authorized for release. The output, denoted by *Point 8*, is the resultant outcome from the Data Processor Layer, which now comprises a plethora of business information that has been designated for release. The now released information can be consumed.

4.1 Comparison Factors

In the domain of corporate assessment, the evaluation of companies and their comparison with peers requires the identification of specific criteria. Such criteria, referred to as *factors* in this Master Thesis, form the crux of this analysis, allowing for the identification of commonalities and differences among a range of companies. To define these factors, an extensive research effort was carried out within the scope of this study, focusing on the distinct features that set each company apart from others.

Upon discovery, these characteristics underwent a rigorous examination. If they aligned well with the purpose of this work, they were inducted into the portfolio of factors. The creation of this portfolio facilitated a structured approach to comparing various businesses and identifying similarities.

A key facet of this methodology involves categorizing these factors into three distinct dimensions. The first, labeled *Business*, encompasses factors that shed light on a company's size. Factors within this dimension allows to capture an overview of a company's business performance. The second dimension, the *Economic* one, focuses on factors linked predominantly to cybersecurity. It uncovers insights into how a company approaches cybersecurity investments, in terms of both direction and magnitude. Finally, the *Technical* dimension contains factors that speak to a company's IT infrastructure.

By mapping these factors onto these distinct dimensions, a multifaceted lens is created through which companies can be compared. This methodology not only facilitates the identification of similar companies across varied areas but also provides insights into how

these companies, similar in one aspect, perform in other areas. This comprehensive approach lays the groundwork for a nuanced understanding of corporate dynamics and inter-company comparisons. The following will delve into a detailed explanation of the various factors for each dimension.

4.1.1 Business Dimension

The factors in the business dimension describe the company in terms of its location, size, and economic success. Examples of factors include country, revenue, organization size and market share. A complete list of all factors in this dimension is provided in Table 4.1. It should be noted that whenever x appears in the *Accepted Values* column, it refers to the value of the corresponding factor.

Table 4.1: Business Factors

Name	**Description**	**Accepted Values**
Country	Country in which the company is based.	CAN | US | FRA | UK | SCA | GER | ITA | TUR | ESP
Revenue	Yearly revenue of the company.	x > 0
Organization Size	Organization size of the company.	Micro | Small | Medium | Large
Marked Share	The marked share of the company.	$x \geq 0\%$ and $x \leq 100\%$
Growth Rate	The growth rate of the company.	$x \geq$ -100% and $x \leq$ 100%
Remote Employees	The average percent of remote working employees.	$x \geq 0\%$ and $x \leq 100\%$

4.1.2 Economic Dimension

The factors associated with the economic dimension center predominantly on aspects of cybersecurity. They provide insights about businesses, such as the extent of their financial commitments to cybersecurity and the budget allocated for it. Furthermore, these factors helps to understand the number of staff members dedicated to the cybersecurity sector, if any cybersecurity training is implemented, and the level of investment in training. Data related to the expenses of cybersecurity insurance and the most severe cybersecurity threats the company has encountered are also encompassed within these factors. Table 4.2 provides a complete rundown of all the factors pertinent to this dimension, inclusive of their relevant descriptions and acceptable values.

Table 4.2: Economic Factors

Name	Description	Accepted Values
Cybersecurity Investment	The amount the company invest in cybersecurity.	$x \geq 0$
Cybersecurity Budget	The available budget for invest in cybersecurity.	$x \geq 0$
Cybersecurity Staffing	The count of employees who are engaged in the cybersecurity area.	$x \geq 0$
Cybersecurity Training Investment	The amount of money spent for cybersecurity training.	$x \geq 0$
Cybersecurity Insurance Investment	The expenditure on cybersecurity insurance.	$x \geq 0$
Cybersecurity Attack Threat	The most significant cybersecurity threat faced by the company.	Malware | DoS | Man-In-The-Middle | Phishing | SQL Injection

4.1.3 Technical Dimension

To describe the technical dimension of a company, factors are defined such as the type of cloud solution the company uses, and whether the company employs multifactor authentication. The technical factors also encompass the network infrastructure and whether remote access is established. Table 4.3 provides a more detailed explanation of these mentioned factors.

Table 4.3: Technical Factors

Name	Description	Accepted Values
Cloud Solution	The cloud solution of the company.	None | Private | Public | Hybrid
Multi-factor Authentication	Does the company use multi-factor authentication?	None | Multi-factor
Network Infrastructure	The network infrastructure of the company.	LAN | WAN
Remote Access	The company's technology for remote access.	None | VPN

4.2 Separation of Concerns

The goal of this Master Thesis is twofold: firstly, to identify similar companies, and secondly, to consume shared information from a similar company. These are two distinct processes which, although interdependent, are to be handled separately in terms

of procedure and the necessary and provided information. Therefore, a clear process for determining similar companies and their required data is defined, as well as for accessing and processing shared information and their data. What follows is a detailed explanation of the data handling process, as well as ensuring that only information that the company has approved for sharing is shared.

4.2.1 Separation of Company Information

Figure 4.2 illustrates how varied company data is integrated into the two processes. The process on the left employs Correlation Measures to ascertain the similarities between companies. Conversely, the process on the right pertains to the release of a company's information. The diagram distinguishes between four different categories of company data, which are detailed in its legend.

The category of *Full Company Information* encompasses all information about a company that the process can access. This includes data regarding all factors, both shared and non-shared. *Shared Company Information* comprises data about the company that has been deliberately shared and can be accessed by other companies. Additionally, *Comparison Data* includes the information necessary for comparing companies, which are the values of the defined factors. Lastly, *Comparison Result Data*, contains the calculated correlation measure distance for each dimension and correlation measure, along with the ID of the company.

The component labeled with *Point 1* serves as the repository of all company data, holding all accessible information about every company. Given that this component also encompasses information that a company may not wish to share, it is essential that such information is thoroughly safeguarded within this component. This component provides the data for the two primary processes, denoted by *Points 2 and 3*.

The process of identifying similarities between companies is denoted by *Point 2*. At *Point 2.1*, two pre-established correlation measures are utilized. The information needed for this component pertains to the pre-defined factors and their corresponding values. Hence, this component is highlighted in red, indicating that it can only access information available through these pre-established factors. The subsequent step, marked as *Point 2.2*, involves processing the data in such a way that all factor-related information is eliminated. Consequently, the resulting data, denoted by *Point 2.3*, only contains information about the respective distances of the different dimensions and correlation measures. Additionally, the ID of the company is affixed to this distance measurement. This ID is subsequently utilized to access the shared information of the respective company.

The process of accessing a specific company's shared information is depicted by *Point 3*. At *Point 3.1*, the designated company is identified within the dataset of all companies, with the gathered information subsequently shared in the next component, represented by *Point 3.2*. Within this component, all the company information not intended for display is filtered out. Up to this stage, the component retains all available information about the company, hence it is also indicated in purple. This changes in the following step, which signifies the output and is represented by *Point 3.3*. At this point, the component only

holds information that the company has knowingly released. This measure ensures that only shared information is received by another company.

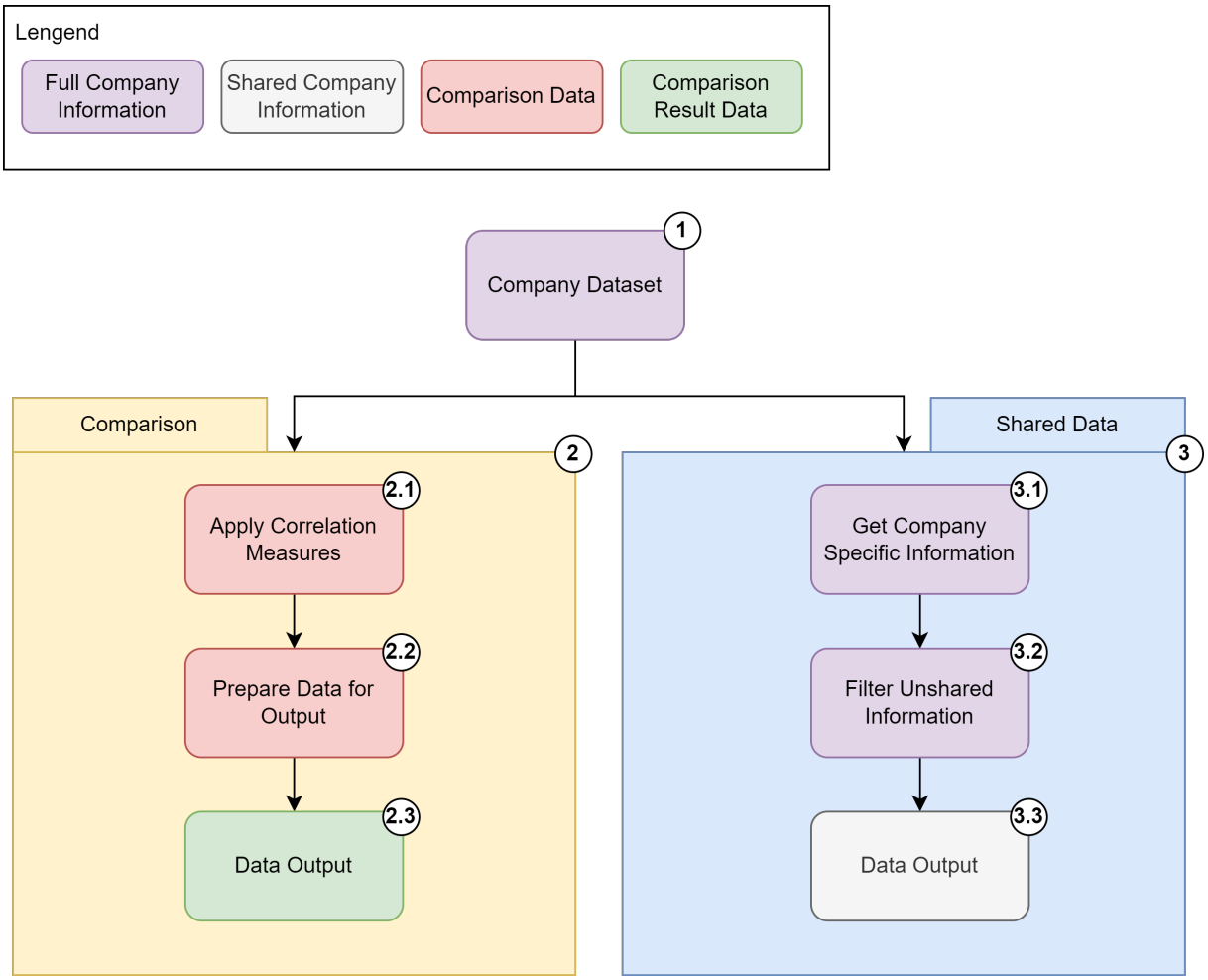


Figure 4.2: Separation of Data

4.3 Data Privacy

When discussing the subject of data exchange and the sharing of information, the issue of data privacy invariably emerges as a recurring theme. The importance of data privacy in this particular work cannot be overstated, particularly in ensuring that company information, not designated for such purpose, isn't inadvertently shared. To guarantee this, meticulous measures are implemented such that every process receives only the precise amount of information about the company that it requires to carry out its designated tasks. The crux of the matter lies in fine-tuning the information flow to ensure the secure performance of tasks without compromising data privacy, thereby achieving a balance between effective data utilization and robust data security.

The possibility of a cyberattacker gaining access to the comprehensive data set comprising sensitive corporate information, although undesirable, cannot be entirely eliminated. In

an effort to curtail the potential damage in such unfortunate incidents, the strategy of data anonymization is adopted. This approach replaces actual company names with randomly generated identification codes within the data set, ensuring an additional layer of protection. The intent of this practice is to shield the company's identity, even in the event of a data breach. Incorporating this technique effectively ensures that, in the event of an attack, the attacker will not be able to directly identify the company to which the compromised data sets belong, thus maintaining confidentiality to a considerable extent.

In a world as diverse as ours, housing as many companies as there are, there exists a plethora of perspectives on the type of information a company should disseminate to other companies. Catering to the demands of company-specific data sharing, an approach is adopted that allows each entity the autonomy to define the nature of information they wish to disclose and those they prefer to retain. It's pertinent to acknowledge that this applies not only to the already mentioned factors but also encompasses additional company-specific information consolidated in the data set. This dynamic configuration of shared information, thus, empowers each company with the ability to dictate the terms of their data sharing, a feature that provides an enhanced level of flexibility. This level of customization in the data sharing process serves as a robust mechanism to ensure the efficient management of proprietary information according to each company's unique needs and discretion.

Chapter 5

Prototype & Implementation

This chapter provides details of the technologies used, the architecture, and describes each component in detail of the architecture necessary to implement the approach described in Chapter 4.

5.1 Technology Stack

In implementing the approach, the existing architecture of SECAdvisor was utilized, as the approach is intended to be integrated into the already existing SECAdvisor tool. Figure 5.1 displays the various layers of the application and the technologies employed for them.

The architecture follows a three-tier design, as outlined in Section 5.2. The foremost layer, known as the User Layer, houses the user interface, which enables the capturing of user interactions and the visualization of data. This layer is crafted using the Angular[1] framework, version 15. Angular, developed by Google, is a TypeScript-based front-end web application framework. TypeScript, in turn, is a programming language derived from JavaScript. Moreover, the design is enhanced with a modern aesthetic using Bootstrap[2] version 5.2.

The second tier of the architecture describes the Business Logic Layer. Its primary responsibility encompasses data processing and preparation. Additionally, this layer handles all intricate calculations. Communication between the user interface and the business layer is facilitated through the Hypertext Transfer Protocol (HTTP). To ensure data is transmitted in a format agnostic to any specific programming language, the JavaScript Object Notation (JSON) is employed. This data format is versatile, enabling seamless data exchange between various subsystems. The implementation of the Business Logic Layer leverages the capabilities of the NestJS[3] framework. Intriguingly, NestJS, which builds upon Node.js[4], shares a similar application structure as Angular.

[1] `https://angular.io/`
[2] `https://getbootstrap.com/`
[3] `https://nestjs.com/`
[4] `https://nodejs.org/de`

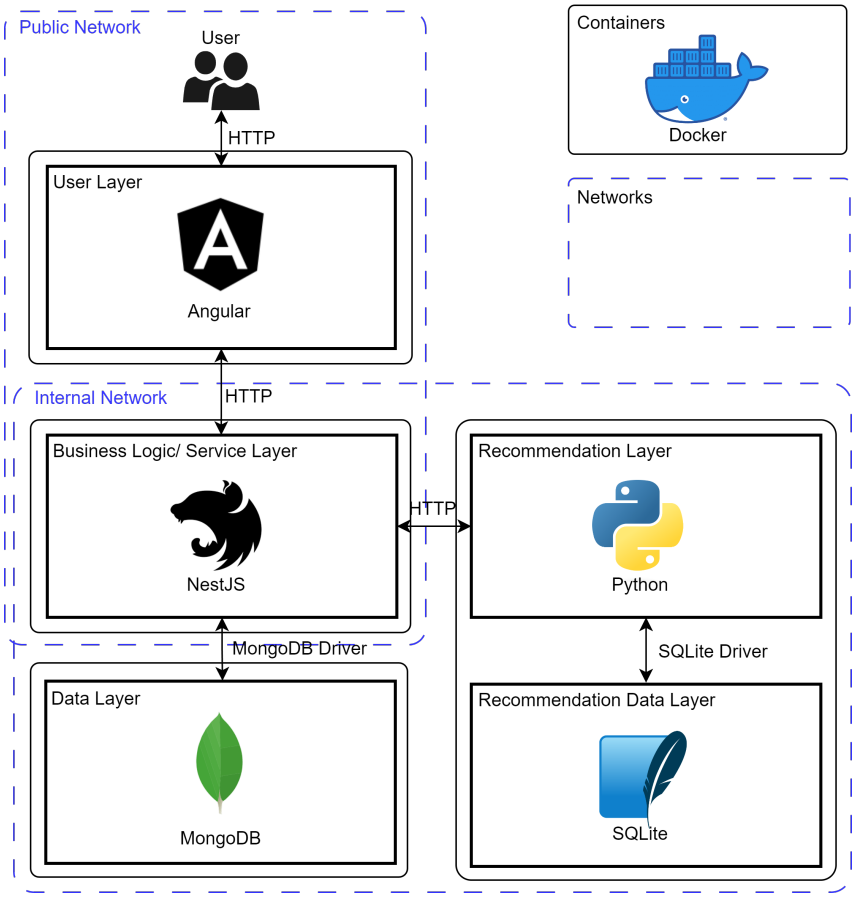


Figure 5.1: Technology Stack based on [44, 35]

The Data Layer tier encompasses the databases essential for storing pertinent application information. Serving as the linchpin for information persistence, MongoDB[5] was the chosen technology for this layer. MongoDB stands out as a document-oriented NoSQL database management system, accommodating data storage in JSON format. This data is subsequently archived within the database as documents. To bridge the application with the database, Mongoose[6] is used. Mongoose, an object data modeling (ODM) tool tailored for Node.js, allows for the creation of schemas, representing the databases' data structures. The conduit linking the business layer and the database primarily operates on HTTP.

The application also includes the Recommendation and Recommendation Data Layer. However, these are not relevant within the scope defined for this work and will not be elaborated upon. To coordinate and run these layers across various networks, Docker[7] and Docker Compose[8] are utilized.

[5] `https://www.mongodb.com/`
[6] `https://mongoosejs.com/`
[7] `https://www.docker.com/`
[8] `https://docs.docker.com/compose/`

5.2 Architecture Overview

Figure 5.2 provides a visual overview of the three separate application layers and their associated roles. For a clearer representation, previously existing components are depicted in a simplified manner. Newly incorporated components stand out with a green highlight, while any alterations to existing components are indicated in blue. The diagram confirms updates made to all three layers, noting the addition of two new components within the Business Logic Layer.

Section 5.3 offers an in-depth exploration of the application's user interface. The subsequent section, 5.4, further investigates the server's capabilities, emphasizing its tasks in data preparation, computational processes, and liaisons with third-party applications. Furthermore, Section 5.5 focuses on the design and structure of the database.

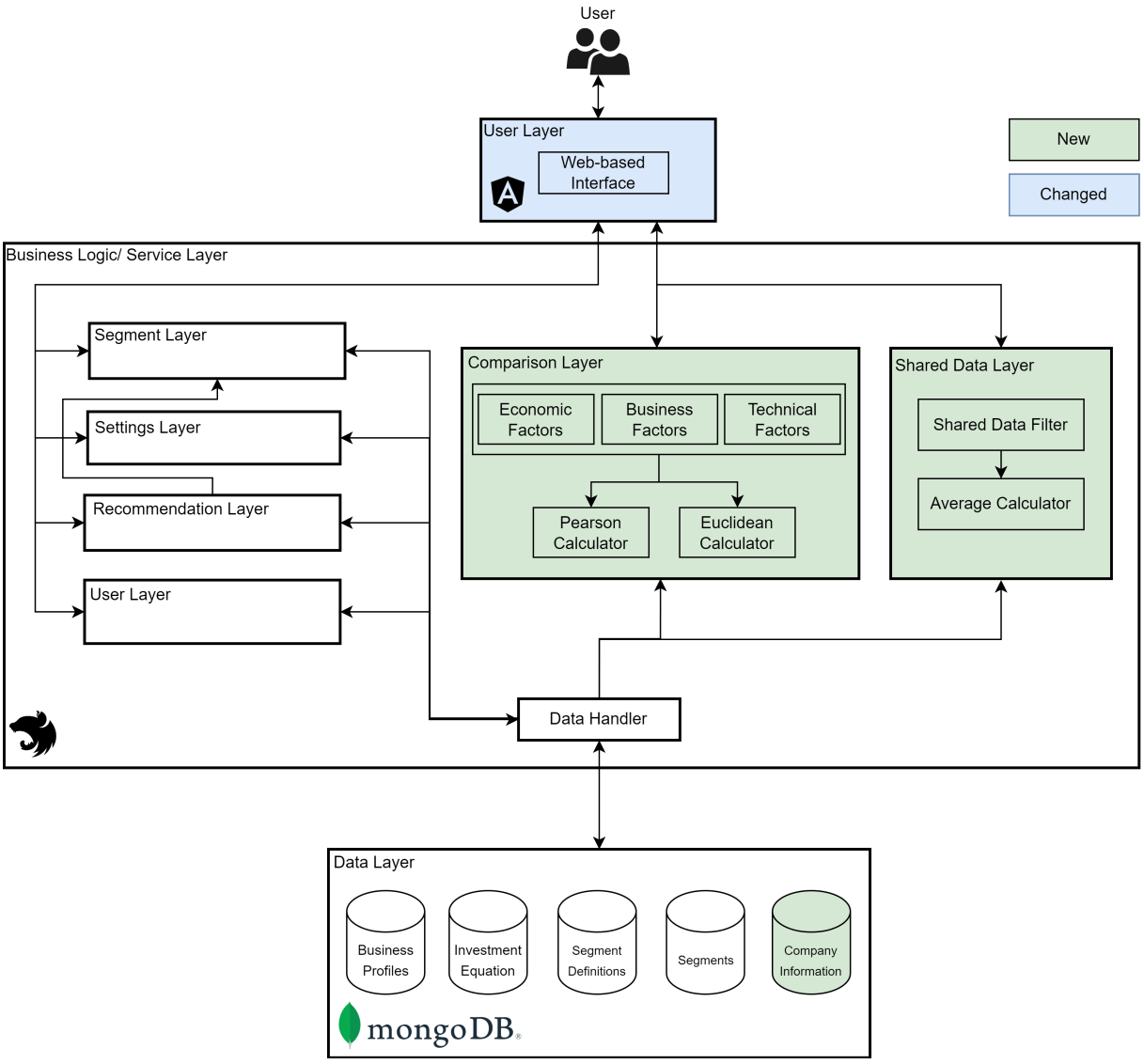


Figure 5.2: Architecture Overview

5.3 User Interface

The user interface set to be implemented drew inspiration from the pre-existing SECAdvisor interface. Emphasis was placed on preserving and building upon the established style and its commendable usability. By ensuring the reusability of specific components, the entire application maintains a consistent and unified appearance and user experience.

Figure 5.3 presents a comprehensive overview of the different pages incorporated within the SECAdvisor tool. These pages are represented by navigation tabs, namely, *Home*, *Business Profile*, *Segments*, *Recommendation*, and *Settings*. Notably, there's an added navigation tab labeled *Analysis Companies* situated at the bottom of the tool's interface. When users engage with this newly-introduced tab by clicking on it, they are seamlessly directed to a fresh page that has been meticulously developed and integrated as a pivotal part of this project.

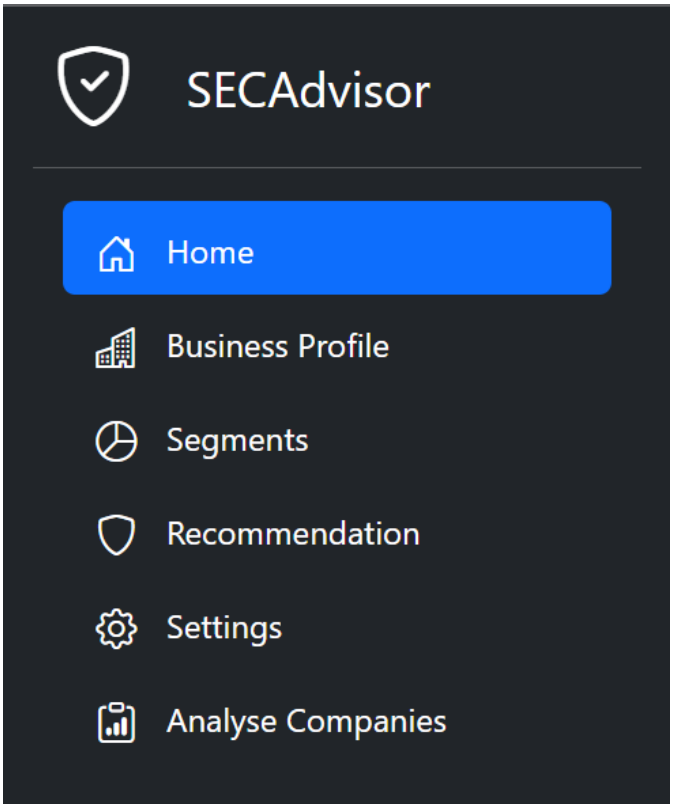


Figure 5.3: Navigation Tabs

Figure 5.4 offers a detailed depiction of the tool as it appears after the user engages with the previously mentioned navigation tab. Upon close observation, one can identify a secondary navigation tab positioned next to the primary navigation tab. This additional secondary tab serves a significant purpose. It allows users to effortlessly toggle between diverse views on a single page. Bearing the title *Company Comparison*, this secondary navigation tab is designed to house two distinct perspectives: the *Chart View* and the *Table View*. Both these views will be subjected to a deeper examination and discussion later within this section. Additionally, in the center of this Figure, there's a blue button

accompanied by a headline. The user is prompted to input details about their company into the tool before proceeding further.

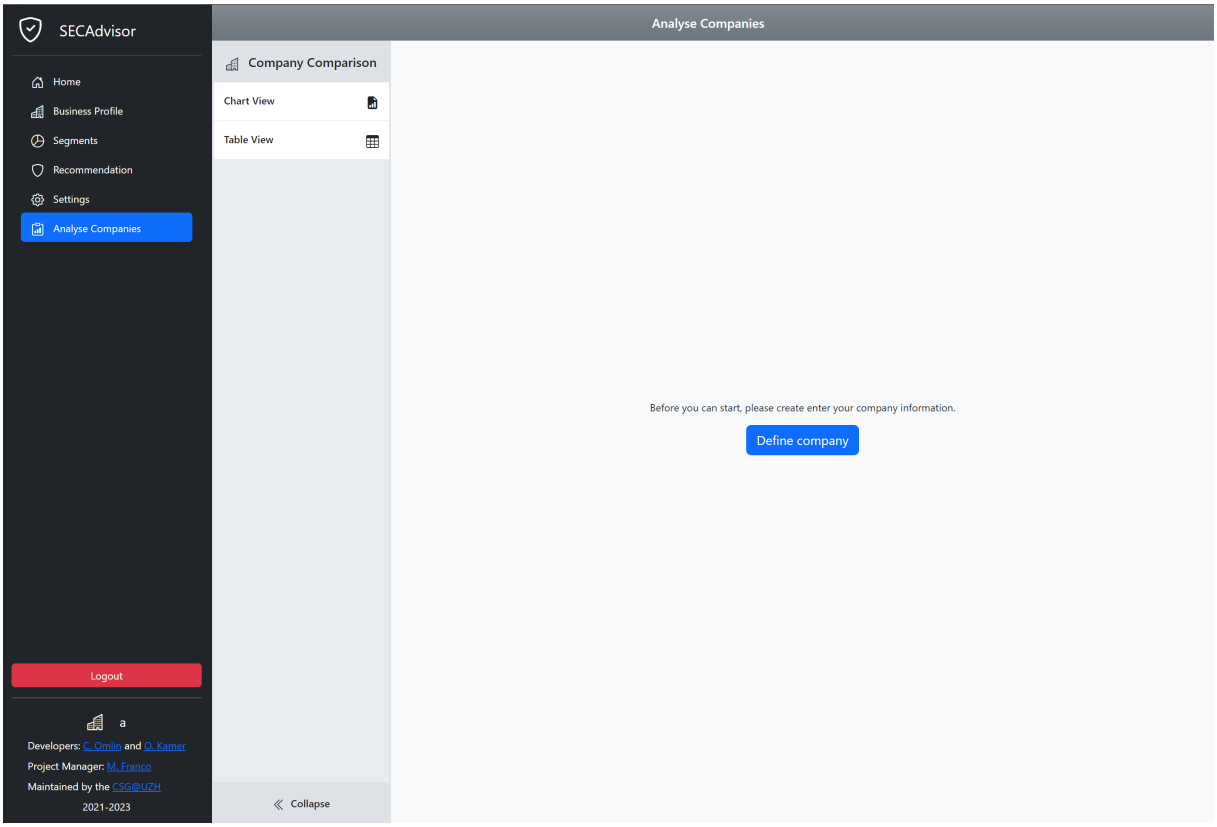
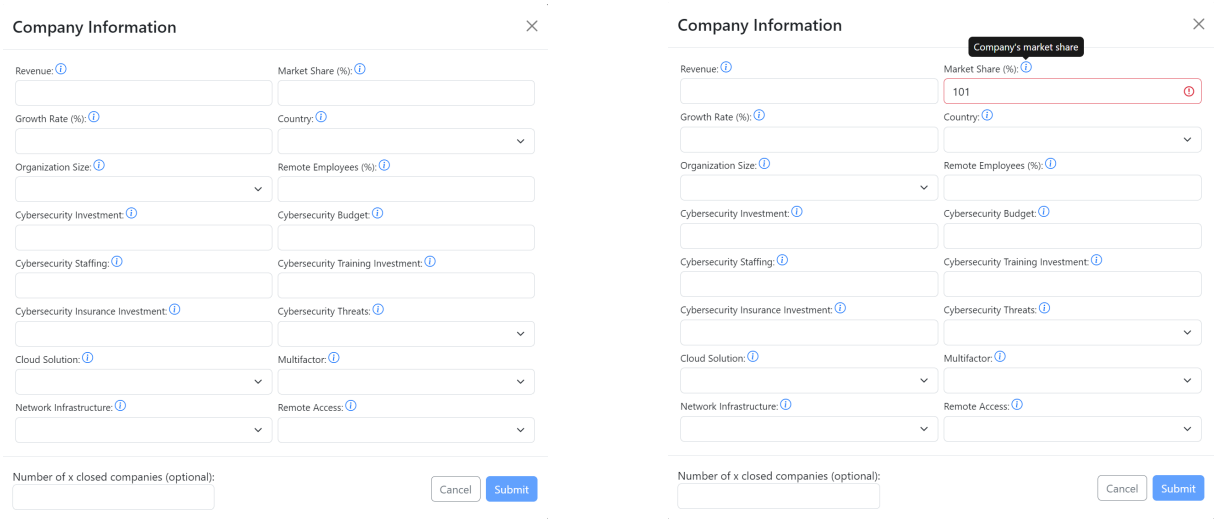


Figure 5.4: Analyse Companies Page

Upon clicking the *Define company* button, the image in Figure 5.5a is revealed. The user is prompted to enter specific company details as outlined in the dialog. Filling out all these details is essential. To emphasize this necessity, the *Submit* button, located at the bottom right of the screen, stays deactivated until all the information is provided correctly. To the left of the *Submit* button, a *Cancel* button is available, enabling the user to close the dialog and return to the previous screen. Additionally, there's an input field at the bottom left where users can indicate how many similar companies they'd like to search for. This input is not mandatory, and its elective status is explicitly marked. If a user opts to skip this field, the search results will display all the relevant companies.

To prioritize user-friendliness and enhance the user experience, a significant focus was placed on incorporating user feedback. An instance of this is captured in Figure 5.5b. Within this illustration, every input field is paired with an info icon. Hovering over this icon provides the user with supplemental guidance about the expected input for that particular field. Should a user input a value outside the approved range, the field becomes emphasized in red, clearly indicating an invalid entry. Additionally, the system is equipped with a keyboard tab index feature, enabling users to effortlessly navigate between various input fields.



(a) Register screen (b) Login screen

Figure 5.5: Company Information Dialog

5.3.1 Chart View

As previously stated, users have the capability to toggle between two distinct views. After entering all the necessary details, they are directed to the *Chart View*. Figure 5.6 offers a visualization of this view, where the choice has been made to display the top 200 prevalent companies for each cluster. The graphic depicts six unique clusters.

In the top-left quadrant, companies are compared based on the business factor, utilizing the Euclidean Distance as a measure. Conversely, the top right segment highlights the business dimension comparison but employs the Pearson Correlation instead. The central row presents companies compared through economic factors, differentiating again between the Euclidean Distance and Pearson Correlation algorithms. Lastly, the bottom row showcases the comparison rooted in technical factors, integrating both correlation techniques.

Additionally, a blue button labeled *Edit Company Information* can be spotted in the top right corner of the image. Clicking on this button reopens the dialog, permitting the user to modify their previously entered company details. Once the dialog is approved with the updated data, the *Chart View* refreshes, presenting the outcomes of the recalculated results.

A crucial aspect to understand is that companies with the closed Euclidean Distance approach a value of zero, whereas those aligning closely in correlation with the provided company data approach a value of 1 when assessed with Pearson Correlation. This layout offers a comprehensive insight into the varied dimensions and correlation measures, assisting users in pinpointing similar companies.

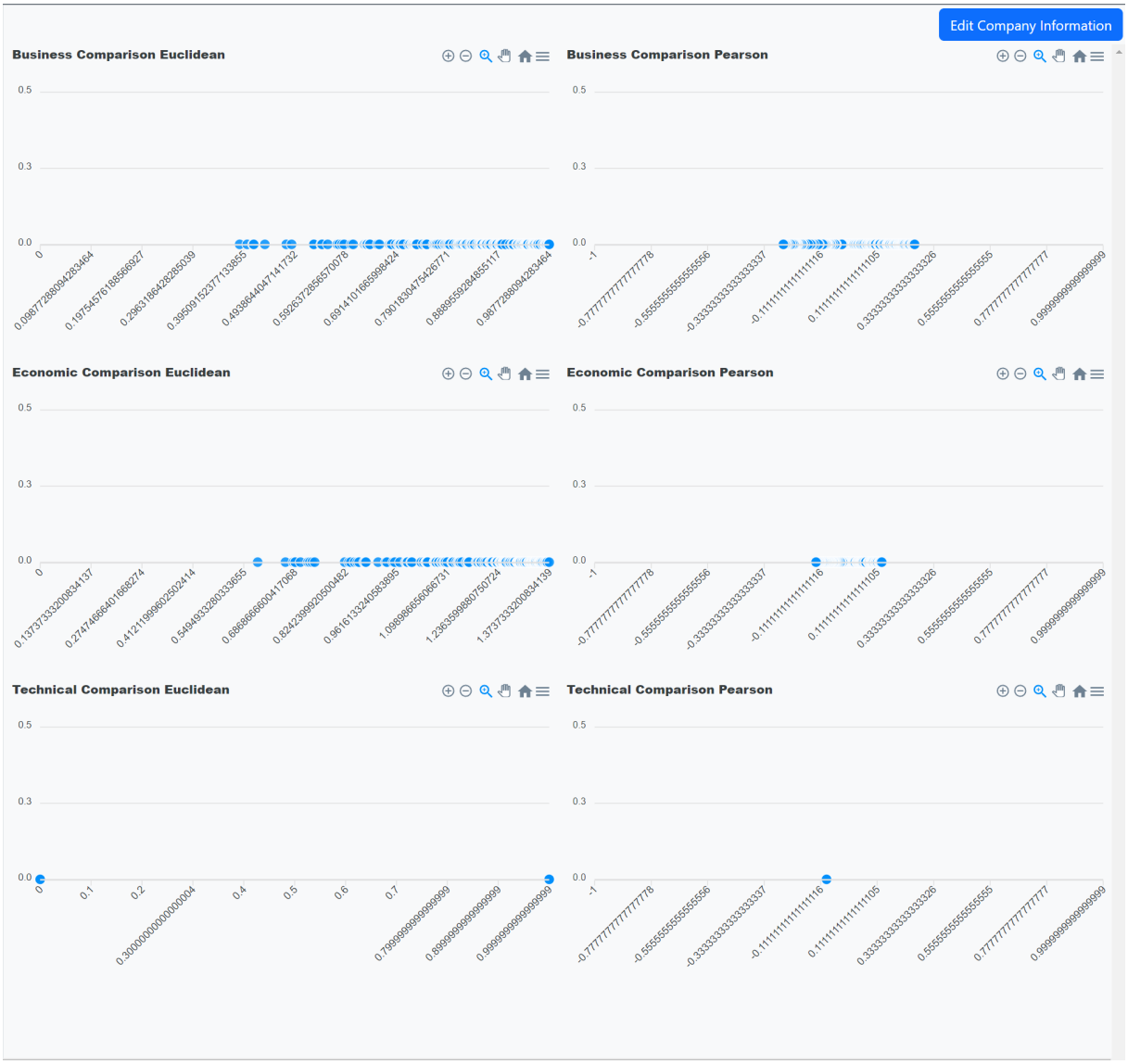


Figure 5.6: Chart View

When a user is curious about a specific company, they can hover the mouse over a point in the chart. This action triggers a small pop-up, revealing the exact correlation measure result and the company's ID. This feature is depicted in the top-left corner of Figure 5.7. In this illustration, there's no specified limit on the number of similar companies to display, so all available companies are shown. This results in overlapping points, making individual entries less distinct. This overlap underscores the importance of allowing users to determine the number of similar companies to be displayed, ensuring a clearer visual representation.

Figure 5.7 highlights another functionality where users can click on a specific point. Selecting such a point initiates a search for that company across all clusters, and if found, the company is emphasized by displaying it in a green color, slightly elevated above the others. In the given illustration, the most similar company in the business dimension, determined using the Euclidean distance, was chosen. Observing the other charts, it's evident that the chosen company appears, marked distinctly with a green dot.

Additionally, a button labeled *Show Shared Information* appears in the top right corner of the image, situated to the left of the *Edit Company Information* button. This button becomes visible once a company is selected. By clicking on it, users can access the publicly shared information of the chosen company.

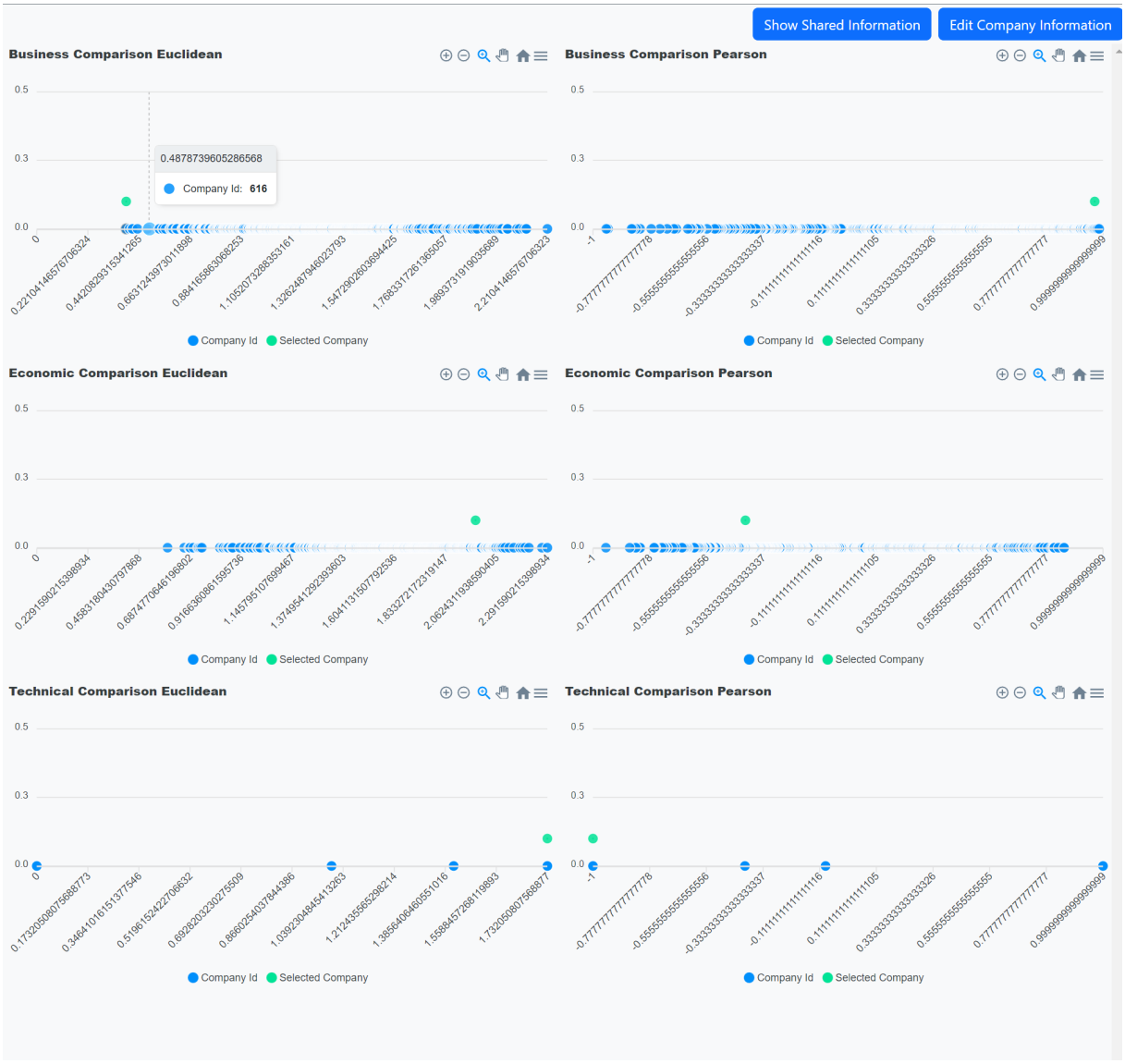


Figure 5.7: Chart View - Company Selection

Upon clicking the *Show Shared Information* button, the view depicted in Figure 5.8 emerges. This view features a table with four distinct columns. The first column lists all attributes that the chosen company has designated for sharing. The second column presents the specific values associated with these attributes for the chosen company. The third column displays values pertaining to the attributes based on the previously provided company information. The fourth column offers insights into the average values from all available companies. To exit the dialog, one can click the button located in the bottom right corner.

The presence of a percentage in an average value suggests that the factor value is not purely numerical but comprises various options. This percentage reflects the extent to which the most popular choice was selected. Moreover, the table outlines attributes that surpass predefined factors. For instance, the table's final row sheds light on the employed breach probability function (BPF), suggesting that the selected company has defined and made available a BPF. Overall, this table grants users a comprehensive perspective on a company's disclosed data, while simultaneously enabling a comparative analysis between the chosen company, the user's company data, and average company values.

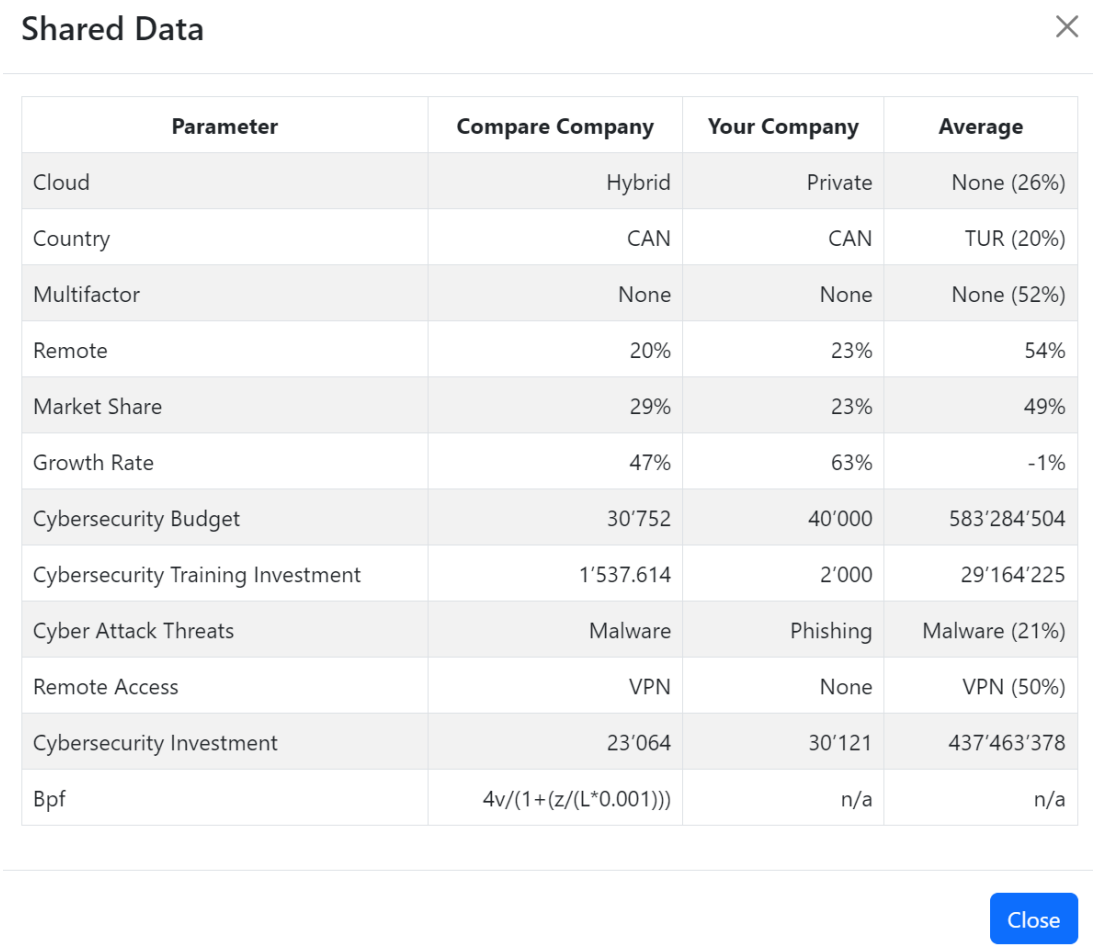


Figure 5.8: Shared Data Dialog

5.3.2 Table View

The *Table View* provides users with an alternative perspective on the correlation measure results. Figure 5.9 presents the *Table View*, showcasing the same data as the *Chart View* but in a distinct format. The input field at the top indicates that *Euclidean Business* is currently chosen, meaning the table displays companies deemed similar based on the Euclidean Distance in relation to business factors. The table's initial row features the company most analogous according to this metric. Consequently, the next row lists the

company that ranks second in similarity, and so forth. Each column signifies the determined values for the correlation measures across various dimensions and algorithms. The final column facilitates access to a company's shared data. Clicking on the "eye" icon triggers the dialog already detailed in Figure 5.8.

Edit Company Information

Comparison:

Euclidean Business

Rank #	Euclidean Distance Business	Euclidean Distance Economic	Euclidean Distance Technical	Pearson Correlation Business	Pearson Correlation Economic	Pearson Correlation Technical	Shared Information
1	0.387	1.969	1.732	0.967	-0.402	-1	
2	0.402	1.773	1.732	0.928	-0.427	-1	
3	0.413	1.693	1.732	0.96	0.352	-1	
4	0.414	1.455	1	0.963	0.086	-0.088	
5	0.436	1.88	1	0.954	-0.38	-0.088	
6	0.478	1.333	1	0.928	0.352	-0.088	
7	0.488	1.217	1.732	0.896	0.159	-1	
8	0.53	1.456	1.414	0.959	-0.133	-0.404	
9	0.531	1.387	1	0.951	0.355	-0.088	
10	0.545	2.038	1	0.903	-0.217	-0.088	
11	0.547	1.615	1	0.901	-0.315	-0.088	
12	0.558	1.826	1.414	0.844	0.253	-0.404	
13	0.577	1.421	1	0.825	-0.229	-0.088	
14	0.58	1.415	1.732	0.886	-0.211	-1	
15	0.586	1.443	1.414	0.91	0.002	-0.404	
16	0.59	2.102	1	0.881	-0.382	-0.088	
17	0.604	1.747	1.414	0.852	-0.128	-0.404	
18	0.607	1.555	1.414	0.937	-0.213	-0.404	
19	0.63	1.337	0	0.807	0.326	1	
20	0.633	1.264	1	0.797	-0.092	-0.088	
21	0.633	1.393	1	0.985	0.277	-0.088	
22	0.635	1.806	1.414	0.872	-0.336	-0.404	
23	0.639	1.457	1.414	0.797	-0.035	-0.404	
24	0.64	1.538	1.414	0.96	0.283	-0.404	

Figure 5.9: Table View

Upon clicking the input field, the view depicted in Figure 5.10 emerges. The dropdown list showcases all available combinations of dimensions and correlation measure algorithms. Once an option is chosen, the table automatically refreshes to display the newly identified similar companies.

Figure 5.10: Table View - Cluster Selection

5.4 Server

The Nest.js framework is utilized for the server, which is referred to as the PublicAPI. This API offers a multitude of endpoints catering to different functionalities. The design of the PublicAPI predominantly adheres to the RESTful API principles, employing the CRUD methodology [51].

Subsequent sections shed light on either new or modified endpoints. While code snippets will illustrate key components, they may not encompass all details and might not function independently. For the sake of clarity, certain imports and context-dependent elements might not be displayed.

5.4.1 Endpoints

To utilize the new features of the PublicAPI, a fresh controller has been established. This controller houses two endpoints. They can be accessed through the ”analyze-companies” route.

The initial endpoint, as presented in Listing 5.1, yields the outcome of the correlation measures applied. This endpoint employs a POST request, where the body provides an object with two properties. The *company* property holds the data about the specific company for which similar companies are being sought, and its expected structure is depicted in Listing 5.3. Meanwhile, the *numberOfClosest* property dictates the quantity of similar companies the PublicAPI should return for each dimension and correlation measure. If this property is not specified, the response will encompass all available companies.

```
16  @Post('')
17  getSimilarCompanies(
18    @Body() body: { company: Company; numberOfClosest?: number },
19  ): Observable<CompanyComparisonDto> {
20    return this.analyseCompaniesService.getSimilarity(
21      body.company,
22      body.numberOfClosest,
23    );
24  }
```

Listing 5.1: Similar Companies Endpoint

Listing 5.2 illustrates the second endpoint. This endpoint is designed to retrieve shared information about a business. It utilizes a GET request, requiring the *companyId* parameter, representing the ID of the company whose shared details are sought. The endpoint's response is an object with two properties. The *company* property provides the shared information about the company. The *average* property, which implements the same interface as the prior property, displays the mean values across all companies. The *SharedCompanyData* interface followed by both properties is illustrated in Listing 5.4.

```
27   @Get('company/:companyId')
28   getSharedCompanyData(
29     @Param('companyId') companyId: string,
30   ): Observable<{ company: SharedCompanyData; average: SharedCompanyData
       }> {
31     return this.analyseCompaniesService.getSharedCompanyData(Number(
      companyId));
32   }
```

Listing 5.2: Get Shared Information Endpoint

```
18 export interface Company {
19   id: number;
20   revenue: number;
21   marketShare: number;
22   growthRate: number;
23   country: Country;
24   organizationSize: OrganizationSize;
25   remote: number;
26   cybersecurityInvestment: number;
27   cybersecurityBudget: number;
28   cybersecurityStaffing: number;
29   cybersecurityTrainingInvestment: number;
30   cybersecurityInsuranceInvestment: number;
31   cyberAttackThreats: CyberAttackThreats;
32   cloud: CloudEnum;
33   multifactor: Multifactor;
34   networkInfrastructure: NetworkInfrastructure;
35   remoteAccess: RemoteAccess;
36 }
```

Listing 5.3: Company Interface

Listing 5.4 shows the *CompanyRawData* interface and its associated properties, depicting an individual company as retrieved from the database. Some properties, like *bpf*, are not encompassed within the entire set of factors and represent additional company information. The *bpf* property indicate the specific breach probability function employed by the company. Moreover, line 76 reveals a property termed *sharedData* that defines the properties a specific company has made public. A more in-depth discussion on this function and its underlying implementation will follow in Subsection 5.4.4.

Line 79 shows the *SharedCompanyData* interface, designated as the return type for the second endpoint. This interface elucidates the specifics that a company has marked for public viewing. Here is how the interface definition is construed:

Using the TypeScript Utility Type *Partial*, it's implied that only select properties of an interface might be established. This means that only specific attributes from the *CompanyRawData* interface are assigned, precisely those the company has earmarked as publicly viewable. The *Omit* TypeScript Utility Type, when paired with the *sharedData* parameter, signifies that the defined interface incorporates every property from *Partial<CompanyRawData>* except for the *sharedData* attribute. The result of this type definition paints a clear picture of the interface encapsulating a company's shared information.

```
57 export interface CompanyRawData {
58   id: number;
59   revenue: number;
60   marketShare: number;
61   growthRate: number;
62   cybersecurityBudget: number;
63   cybersecurityStaffing: number;
64   cybersecurityTrainingInvestment: number;
65   cybersecurityInsuranceInvestment: number;
66   cyberAttackThreats: CyberAttackThreats;
67   networkInfrastructure: NetworkInfrastructure;
68   remoteAccess: RemoteAccess;
69   cybersecurityInvestment: number;
70   cloud: CloudEnum;
71   country: string;
72   multifactor: Multifactor;
73   organizationSize: number;
74   remote: number;
75   bpf: string;
76   sharedData: (keyof CompanyRawData)[];
77 }
78
79 export type SharedCompanyData = Omit<Partial<CompanyRawData>, '
     sharedData'>;
```

Listing 5.4: Company Raw Data Interfcae

5.4.2 Normalization

To ensure that every factor is evaluated on an equal footing, they are normalized. Through this procedure, each factor is given a value ranging from 0 to 1 before applying correlation measures. This subsection delves deeper into several normalization methods for factors. However, it's worth noting that not every factor is comprehensively addressed.

Numerical Factors

For every numeric factor, the range, encompassing the minimum and maximum values, is first established across all companies. This procedure is illustrated in Listing 5.5. The function expects an array of companies as its parameter, with iteration over this array occurring on line 147. The subsequent line, 148, involves iteration over *numericFactors*, which is an array consisting of factors with numeric values. Line 150 assesses whether the boundaries for a particular factor have been initialized. If not, the current factor's value is employed as the initial value. Should a factor's boundary already be set, the instructions from lines 156 to 159 are followed, updating the minimum and maximum values.

```
146 const findBoundaries = (companies: CompanyRawData[]) => {
147   companies.forEach((company) => {
148     numericFactors.forEach((factor) => {
149       const value = company[factor] as number;
150       if (!boundaries || !boundaries[factor]) {
151         boundaries = {
152           ...(boundaries ?? {}),
153           [factor]: { min: value, max: value },
154         };
155       } else {
156         boundaries[factor].min =
157           boundaries[factor].min < value ? boundaries[factor].min :
    value;
158         boundaries[factor].max =
159           boundaries[factor].max > value ? boundaries[factor].max :
    value;
160       }
161     });
162   });
163 };
```

Listing 5.5: Find Bounderies Function

After establishing the minimum and maximum values for each numeric factor, these factors' values are then normalized. This normalization is depicted in Listing 5.6. The function *normalizeNumericFactors* takes in a single entity along with an array containing the names of the numeric factors. Line 137 involves iteration over every numeric factor. Subsequently, on line 138, the previously determined boundaries are retrieved, and on line 139, the specific factor's value is accessed. The normalized value of the factor is then computed on line 141 and allocated to a new attribute. This calculation entails dividing the difference between the factor's value and the minimum value by the span between the maximum value and the minimum value.

To better grasp this calculation, consider an example. Let's say the factor *growthRate* is to be normalized. The minimum value for all *growthRate* is 50, while the maximum stands at 100. If the value needing normalization is 75, the calculation would proceed like this:

$$\frac{75 - 50}{100 - 50} = 0.5$$

Consequently, the normalized value of the *growthRate* factor is 0.5. Intuitively, this is logical, as the value falls precisely midway between the minimum and maximum.

This function allows for the easy normalization of numeric factors using minimal code. The resulting normalized values are then utilized in subsequent correlation measure applications.

```
133 const normalizeNumericFactors = (
134   company: CompanyRawData,
135   props: (keyof CompanyRawData)[],
136 ) => {
137   props.forEach((prop) => {
138     const boundary = boundaries[prop];
139     const value = company[prop] as number;
140
141     (company[`${prop}N` as keyof typeof company] as number) =
142       (value - boundary.min) / (boundary.max - boundary.min);
143   });
144 };
```

Listing 5.6: Normalize Numeric Factors Function

Non-Numerical Factors

Normalizing non-numerical factors is not universally standard because each factor demands unique handling. For every non-numerical factor, a specific method to determine its normalized value must be outlined. Numerical values needs to be allocated to all potential values, which can subsequently be employed for normalization.

Listing 5.7 displays the variable that maps potential organization size values to their corresponding numeric values. In this mapping, the organization size *Micro* is assigned a value of 0, *Small* gets 1, *Medium* is given 2, and *Large* receives 3. This arrangement indicates, for instance, that the size *Small* is closer in value to *Micro* than it is to *Large*. Using this mapping, various organization sizes can be allocated numerical values, facilitating their normalization.

```
1 export const organizationSizeMapping = {
2   Micro: 0,
3   Small: 1,
4   Medium: 2,
5   Large: 3,
6 };
```

Listing 5.7: Organisation Size Mapping

Listing 5.8 showcases the function responsible for normalizing the value of the organization size. The function expects a parameter named *organisationSize* of type *OrganisationSize*. This type is delineated as a string that can be either *Micro*, *Small*, *Medium*, or *Large*. On line 19, the numeric counterpart for the given organization size is determined, which falls within a range of 0-3. Line 20 calculates the normalized value by dividing the numeric value by the total count of the *organisationSizeMapping* variable, which is 4. This calculation yields a result between 0 and 1, effectively normalizing the factor's value. This function is invoked for every company, as well as for the particular company for which similar counterparts are being sought.

```
16 export const getNormalizedOrganizationSize = (
17   organizationSize: OrganizationSize,
18 ) => {
19   const value = organizationSizeMapping[organizationSize];
20   return value / Object.keys(organizationSizeMapping).length;
21 };
```

Listing 5.8: Normalized Organisation Size Function

Listing 5.9 provides an other illustration of mapping values from a non-numeric factor to their numeric counterparts. Every potential company location is paired with a designated numeric value. These numeric assignments is established based on the geographical proximity to Germany, which serves as the reference point. From these values, one can infer, for instance, that Turkey is geographically farther from Germany than Italy is. Leveraging this mapped data, location-based information can now be effectively normalized.

```
3 const countryDistanceMapping = {
4   CAN: -5,
5   US: -4,
6   ESP: -3,
7   UK: -2,
8   FRA: -1,
9   GER: 0,
10  ITA: 1,
11  SCA: 2,
12  TUR: 3,
13 };
```

Listing 5.9: Country Mapping

The corresponding normalization process is detailed in Listing 5.10. The function expects a country's numeric value as an input parameter. Lines 40 and 41 compute the normalized value. This computation is quite similar to the one described in Listing 5.6. The minimum is represented by Canada's value of -5, while the maximum is defined by Turkey's value of 3. The end result is a value between 0 and 1, signifying the normalized value.

The same method is applied to other non-numeric values, including categories like *Cloud*, *Cyber Attack Threat*, *Remote*, and the like. Initially, numeric values are allocated to

```
39 const normalize = (country: number) =>
40   (country - countryDistanceMapping.CAN) /
41   (countryDistanceMapping.TUR - countryDistanceMapping.CAN);
```

Listing 5.10: Normalize Country Function

different variations of the factor. Subsequently, these factors are normalized. However, for certain factors like *Multifactor*, a distinction is made based on whether multifactor authentication is in use or not. In such scenarios, the normalization is adjusted so that a value of 0 is returned when multifactor authentication is not used, and a value of 1 when it is employed. By attributing numerical values to various selections of a factor, correlation measures can be applied through an intermediary step.

5.4.3 Correlation Measure Calculation

A primary capability of the server is the application of two correlation measures, the Euclidean Distance and the Pearson Correlation. The previously normalized values are employed to ascertain the similarities between companies.

The server calculates the Euclidean Distance using the euclidean-distance[9] library, as demonstrated in Listing 5.11. This code snippet computes the Euclidean Distance between a target company and another company from the database for the business dimension. Line 235 features the *euclideanDistance* function offered by the library. It takes two arrays as input parameters. The first array consists of the normalized values of the company for which similar counterparts are sought. Conversely, the second array comprises normalized factor values from a database-listed company. The output array then presents a number denoting the Euclidean Distance.

```
235         euclideanDistanceBusiness: euclideanDistance(compareObject, [
236           company['revenueN' as keyof typeof company],
237           company['marketShareN' as keyof typeof company],
238           company['growthRateN' as keyof typeof company],
239           company['countryN' as keyof typeof company],
240           company['organizationSizeN' as keyof typeof company],
241           company['remoteN' as keyof typeof company],
242         ]),
```

Listing 5.11: Euclidean Distance Calculation

Listing 5.12 illustrates the calculation of the Pearson Correlation for the business dimension. This code snippet bears resemblance to the previous one, but it employs the *calculateCorrelation* function in lieu of the *euclideanDistance* function. This function is sourced from the calculate-correlation[10] library. The outcome is a number that represents the Pearson Correlation.

[9] https://www.npmjs.com/package/euclidean-distance
[10] https://www.npmjs.com/package/calculate-correlation

```
257         pearsonCorrelationBusiness: calculateCorrelation(compareObject,
      [
258           company['revenueN' as keyof typeof company],
259           company['marketShareN' as keyof typeof company],
260           company['growthRateN' as keyof typeof company],
261           company['countryN' as keyof typeof company],
262           company['organizationSizeN' as keyof typeof company],
263           company['remoteN' as keyof typeof company],
264         ]),
```

Listing 5.12: Normalize Numeric Factors Function

In order to guarantee that only authorized information is shared, the server only sends the results of the correlation measures to the frontend. Details about the factors and their normalized values are withheld to protect the data that companies have not approved for release.

5.4.4 Access Shared Data

Companies have the option to share specific information they wish to disclose to other companies. They can designate this via the *sharedData* properties, as previously introduced in Listing 5.4. This property allows companies to determine which attributes should be visible to other companies.

Listing 5.13 outlines the procedure ensuring that only the specified shared information of a company is retrieved from the server. Line 125 displays the parameter required for the *getSharedCompanyInformation* method, representing a company from the database. Line 127 shows the iteration over the *sharedData* property, where only the values of these properties are extracted and combined into an object. The outcome is an object encompassing all shared company details, conforming to the *SharedCompanyData* interface, which has been previously elaborated. This approach guarantees that only the designated information about a company is relayed from the server.

```
124   private getSharedCompanyInformation(
125     company: CompanyRawData,
126   ): SharedCompanyData {
127     return company.sharedData.reduce<SharedCompanyData>(
128       (pre, curr) => ({ ...pre, [curr]: company[curr] }),
129       {},
130     );
131   }
132 }
```

Listing 5.13: Shared Company Information Method

5.5 Database

The primary database for this tool is MongoDB. MongoDB offers a distinct advantage by allowing the storage of more complex models directly within the database [6]. This aligns well with the modern approach adopted in this application. MongoDB greatly streamlines operations because the data is stored in a manner consistent with its use in the application.

In order to read the companies from the database, a Mongoose schema is established. Listing 5.14 presents this schema. The displayed schema represents the manner in which data is stored in the database. The outlined schema closely resembles the one described in Listing 5.5, with the only differences being in the type definition conventions. Due to this schema, it is possible to retrieve type-safe data from the database for further processing.

```
83 export const CompanySchema = new mongoose.Schema({
84   id: Number,
85   revenue: Number,
86   marketShare: Number,
87   growthRate: Number,
88   cybersecurityBudget: Number,
89   cybersecurityStaffing: Number,
90   cybersecurityTrainingInvestment: Number,
91   cybersecurityInsuranceInvestment: Number,
92   cyberAttackThreats: CyberAttackThreats,
93   networkInfrastructure: NetworkInfrastructure,
94   remoteAccess: RemoteAccess,
95   cybersecurityInvestment: Number,
96   cloud: CloudEnum,
97   country: String,
98   multifactor: Multifactor,
99   organizationSize: Number,
100  remote: Number,
101  bpf: String,
102  sharedData: [String],
103 });
```

Listing 5.14: Company Schema

5.6 Challenges

One of the challenges in developing the user interface for the prototype was presenting a large number of companies across the different charts. It became apparent that rendering a large number of data points on the chart was more time-consuming than anticipated by users. Initial efforts to optimize the code for reduced processing did not yield the desired results. This suggests that the ng-apexcharts[11] library inherently requires this duration to render the points. To expedite this process, the decision was made to optional limit

[11] https://www.npmjs.com/package/ng-apexcharts

the number of companies displayed. Users have the flexibility to determine the number of companies they wish to view. An input field enables them to specify this number, subsequently enhancing the performance.

Another challenge emerged when trying to normalize non-numeric values. While numeric values could be normalized by just one single function, each non-numeric factor required a distinct normalization function. It was imperative to ensure that the assignment of numerical values to various factor selection options adhered to a logical framework, facilitating a comparison between different selections. This method worked effectively, but it was not feasible when there were only two choices available. This issue was addressed by applying a logic that if the target company's selection matched the selection of the company being compared, a value of 0 was returned. Otherwise, a value of 1 was returned.

Chapter 6

Evaluation

In order to assess the correctness and usability of the tool, a two-stage evaluation is performed. The first stage emphasizes the accurate computation of correlation measures for the specified factors. The second stage of the evaluation encompasses case studies that validate the tool's functionality using real-world scenarios.

6.1 Factors

The purpose of this evaluation is to verify the accuracy of the correlation measures calculated for the defined factors. Typically, when using the tool, all factors are input values for computing the correlation measures, yielding a single combined result. Due to this amalgamation, it's challenging to ascertain the correctness of a individual factor's calculation. Therefore, in this evaluation, factors are examined individually to understand and validate the accuracy of their calculations.

Furthermore, validating the accuracy of calculations becomes challenging when applying the correlation measures across numerous companies from the database. As a solution, the method involves applying individual factors to a select few companies. This facilitates a clearer understanding of the results and enables a more straightforward assessment of their accuracy.

Given that the Pearson Correlation necessitates multiple factors to establish a correlation, the subsequent discussions will focus solely on the Euclidean distance. Nevertheless, it can be posited that if the normalization calculation for a factor is accurate for the Euclidean distance, it will also hold true for the Pearson Correlation. This is because they employ similar algorithmic approaches, only differing in the specific function invoked for each correlation measure. The distinctions between the Pearson Correlation and the Euclidean distance will be addressed in the second phase of the evaluation.

6.1.1 Environment

In order to apply the correlation measures solely to specific companies, rather than to every company in the database, the Public API has been enhanced with an additional API endpoint. This new endpoint is presented in Listing 6.1. The primary distinction between this new endpoint and the existing one is that, as illustrated on *Line 34*, this endpoint accepts a list of companies as parameters, overriding the companies from the database. This adjustment simplifies the traceability and assessment of individual evaluations.

```
29  @Post('custom-companies')
30  getSimilarCustomCompanies(
31    @Body()
32    body: {
33      company: Company;
34      compareCompanies: CompanyRawData[];
35      numberOfClosest?: number;
36    },
37  ): Observable<CompanyComparisonDto> {
38    return this.analyseCompaniesService.getSimilarity(
39      body.company,
40      body.compareCompanies,
41      body.numberOfClosest,
42    );
43  }
```

Listing 6.1: Custom Companies Endpoint

```
52 with open('assets/mock-companies.json') as companies_file:
53   file_contents = json.load(companies_file)
54
55 response = requests.post(os.path.join(url, 'custom-companies'), json={'
       company': target_company, 'compareCompanies': file_contents })
56 response_json = response.json()
57 print(response_json)
58 eBusinessCompanies = response_json["euclideanDistanceBusiness"]
59 eBusinessValues = list(map(lambda x: x["euclideanDistanceBusiness"],
       eBusinessCompanies))
60
61 # Plot data
62 x = eBusinessValues
63 y = list(map(lambda x: 1, eBusinessValues))
64
65 # Create a line plot
66 plt.plot(x, y, marker='o', linestyle='None', color='b', label='Company')
```

Listing 6.2: Python Evaluation Program

To facilitate and expedite the evaluation of the correlation measures' accuracy for the factors, a Python program has been developed. Listing 6.2 displays a portion of this

script, where the result of the Euclidean distance in the business dimension is analyzed. *Line 52 and Line 53* illustrate the process of reading the list of companies, which serve as the foundation for comparison, from a distinct JSON file, and saving the content into a variable. *Line 55* reveals the invocation of the new endpoint discussed previously, passing both the target company and the list of companies. The outcome of the REST calls is subsequently formatted on *Lines 58 and 59*, ensuring to consider only the values of the Euclidean distance in the business dimension. These values are then presented in a graphic on *Line 66*, with *Line 62 and Line 63* setting the x and y axes, respectively.

The subsequent assessments of the correlation measures calculations for individual factors are conducted using this Python program. The program enables swift modifications to the target company's information without needing user interface interactions. Additionally, the list of companies, which form the basis for comparison, can be easily updated in the JSON file.

6.1.2 Business Factors

Subsequently, for each factor related to the business dimension, a set of companies is compiled to serve as a comparative basis, and a specific target company is identified to find its similar counterparts. The factors *Market Share* and *Remote Employees* are addressed collectively since they employ the same algorithm, attributed to their identical range of permissible values. As a result, only the *Market Share* factor is taken into account, without separately considering the *Remote Employees* factor.

Revenue

To verify the accuracy of the calculations for the *Revenue* factor, three distinct companies with varying revenues are specified. These companies are presented in Table 6.2. The second row of this table highlights the chosen target company, for which a company with the closest revenue is being sought. Upon examining the table, it's clear that the company with ID 3 has revenue most closely aligned with that of the target company.

The outcome of the Euclidean distance is depicted in Figure 6.1. Evidently, the company with ID 3 is the closest match to the target company, exhibiting a distance of 0.15. Following that, the company with a distance of 0.35 ranks as the second most similar, while company 1, with a distance of 0.85, is the least similar. These results align with expectations, affirming that the values have been normalized appropriately and the results are accurate.

Company Name	Company ID	Revenue
Target Company	n/a	$27'000'000
Company 1	1	$10'000'000
Company 2	2	$20'000'000
Company 3	3	$30'000'000

Table 6.1: Revenue Factor - Companies

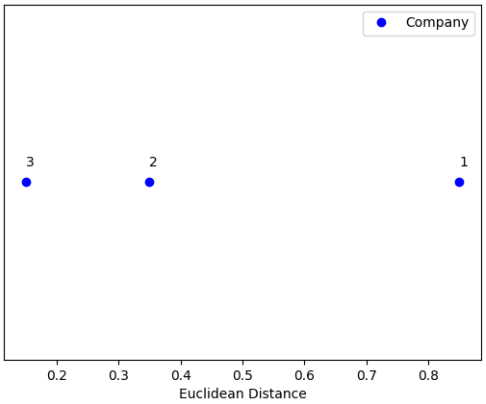


Figure 6.1: Revenue Factor - Euclidean Distance

Country

In order to verify the accuracy of the normalization and the results from the Euclidean distance, three companies were specified along with a corresponding target company, illustrated in Table 6.2. The target company is based in Italy, Company 1 in United Kingdom, Company 2 in Germany, and Company 3 in the USA. It's anticipated that Company 2 will emerge as the most similar to the target, followed by Company 1, and then Company 3.

Company Name	Company ID	Country
Target Company	n/a	ITA
Company 1	1	UK
Company 2	2	GER
Company 3	3	US

Table 6.2: Country Factor - Companies

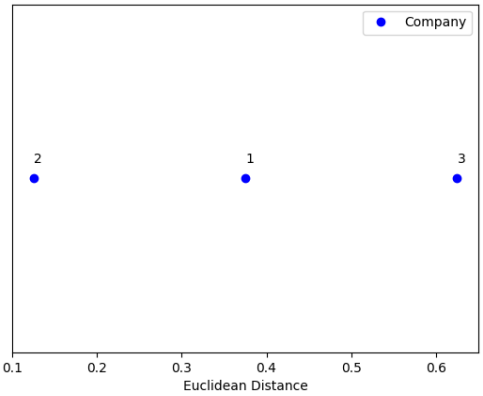


Figure 6.2: Country Factor - Euclidean Distance

Figure 6.2 displays the results, clearly indicating that Company 2 is the nearest to the target company with a distance of 0.125, succeeded by Company 1 and Company 3. This outcome aligns with the anticipated results, reaffirming the accuracy of the calculations.

Organization Size

Table 6.3 presents the three companies chosen for this scenario, along with the target company. The objective is to identify a company with an organizational size closest to that of the target company, which is categorized as *Small*. Company 1 is classified as *Micro*, signifying a the smallest size. Company 2 has a medium-sized organization, while Company 3 is large. It's anticipated that Company 1 and Company 2 will be deemed the closest in similarity, given their identical distance in terms of numerical factor values (as referenced in Listing 5.7). Company 3 should be ranked last, given its significantly different organizational size.

Company Name	Company ID	Org Size
Target Company	n/a	Small
Company 1	1	Micro
Company 2	2	Medium
Company 3	3	Large

Table 6.3: Organization Size Factor - Companies

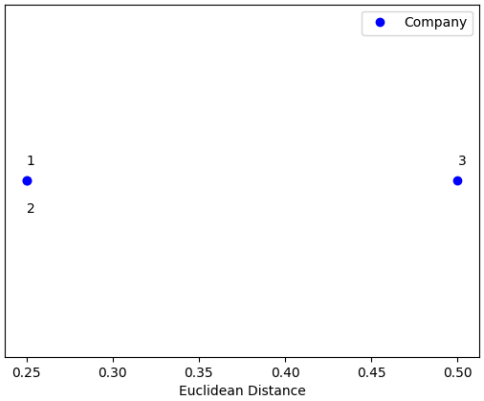


Figure 6.3: Organization Size Factor - Euclidean Distance

Figure 6.3 displays the results of the calculation. It's evident that Company 1 and Company 2 share an identical Euclidean distance, causing their points to overlap. This observation validates the earlier predictions. Both of these companies, being closest to 0, exhibit the greatest similarity to the target company in terms of organizational size, each with a distance of 0.25. With a distance of 0.5, Company 3 is the least similar.

Market Share

For the companies chosen in this scenario, the market shares are as follows: the target company commands a 26% share, Company 1 holds 38%, Company 2 possesses 59%, and Company 3 has 15%. These figures are detailed in Table 6.4. As anticipated, Company 3 should be most similar to the target company, followed by Company 1 and then Company 2.

It's evident from the data in Figure 6.4 that Company 3 exhibits the greatest similarity in terms of market share, boasting a Euclidean distance of 0.11. This is closely followed by Company 1 with a distance of 0.12, while Company 2 concludes with a distance of 0.33. These results are consistent with prior expectations, reinforcing the belief that the factor's normalization is accurate, and the Euclidean distance has been calculated correctly.

Company Name	Company ID	Market Share
Target Company	n/a	26%
Company 1	1	38%
Company 2	2	59%
Company 3	3	15%

Table 6.4: Market Share Factor - Companies

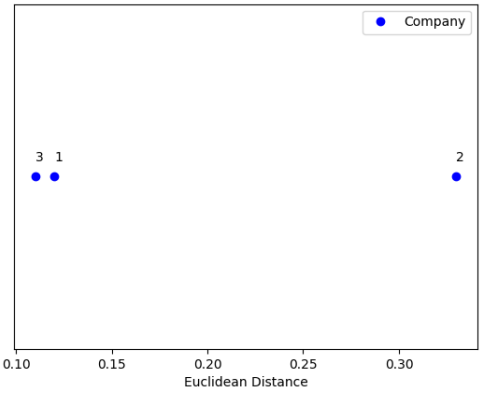


Figure 6.4: Market Share Factor - Euclidean Distance

Growth Rate

Table 6.5 lists the three companies alongside their respective growth rates. Notably, Company 3's growth rate is the nearest to the target value of 2%. Company 2 is the next closest to the target, while Company 1 is expected to rank last.

Company Name	Company ID	Growth Rate
Target Company	n/a	2%
Company 1	1	18%
Company 2	2	-5%
Company 3	3	8%

Table 6.5: Growth Rate Factor - Companies

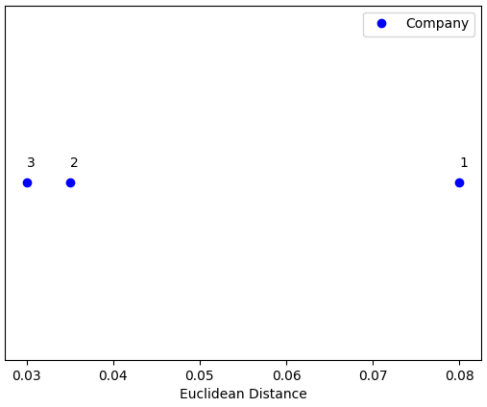


Figure 6.5: Growth Rate Factor - Euclidean Distance

As depicted in Figure 6.5, the anticipated outcomes are realized. Company 3, with a distance of 0.03, is the closest to the target company. It's followed by Company 2 at a distance of 0.035, while Company 1, with a distance of 0.08, is the least similar. The alignment of results with expectations suggests that both the normalization and the Euclidean distance calculations are accurate.

6.1.3 Economic Factors

Next, the normalization and calculation of the Euclidean distance for economic factors are scrutinized. Given that the factors *Cybersecurity Investment*, *Cybersecurity Budget*, *Cybersecurity Staffing*, *Cybersecurity Training Investment*, and *Cybersecurity Insurance Investment* share the same accepted value range (as indicated in Table 4.2), the same algorithm is utilized for their normalization. Consequently, the focus will primarily be on *Cybersecurity Investment*. If its normalization and Euclidean distance calculations prove accurate, the same can be inferred for the other mentioned factors.

Cybersecurity Investment

Table 6.6 displays the selected companies along with their cybersecurity investments. Company 1 is anticipated to be the closest to the target company, given its cybersecurity investment is same to the $150,000 mark. It's expected to be followed by Company 2 and then Company 3.

Company Name	Company ID	Cybersecurity Investment
Target Company	n/a	$150'000
Company 1	1	$150'000
Company 2	2	$220'000
Company 3	3	$310'000

Table 6.6: Cybersecurity Investment Factor - Companies

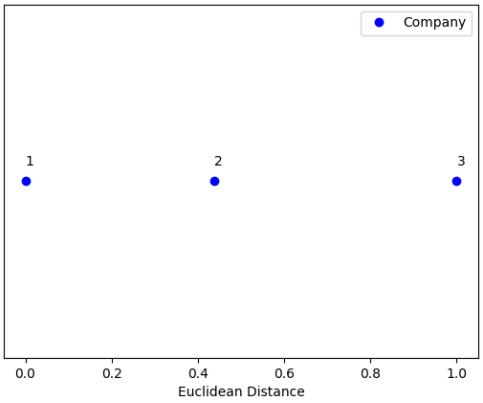


Figure 6.6: Cybersecurity Investment Factor - Euclidean Distance

Figure 6.6 illustrates the Euclidean distances for the chosen companies. Notably, Company 1 has a distance of 0, indicating that its cybersecurity investment matches that of the target company. Company 2 follows, at a distance of 0.44, being the next closest to the target, while Company 3 trails as the least similar. These outcomes align with prior expectations, reaffirming the accuracy of the normalization and the Euclidean distance calculations.

Cybersecurity Attack Threat

For this particular use case, the selected companies are detailed in Table 6.7. The target company has identified malware as its primary cyber attack threat. All other companies,

except for Company 3, have highlighted different cyber threats. This suggests that only Company 3 should register a distance of 0, with the rest displaying a distance of 1.

As illustrated in Figure 6.7, Company 3, with a distance of 0, aligns perfectly with the target company in terms of cyber attack threat. Both Company 1 and Company 2 register a distance of 1, as their identified cyber attack threats differ from the target company's. This outcome validates the initial assumptions.

Company Name	Company ID	Cyber Attack Threat
Target Company	n/a	Malware
Company 1	1	DoS
Company 2	2	Phishing
Company 3	3	Malware

Table 6.7: Cybersecurity Attack Threat Factor - Companies

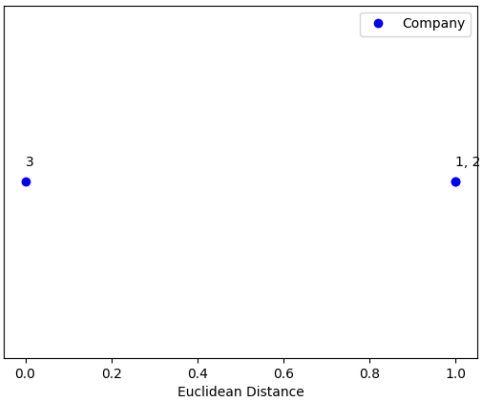


Figure 6.7: Cybersecurity Attack Threat Factor - Euclidean Distance

6.1.4 Technical Factors

This section delves into the technical factors, examining the accuracy of their normalization and Euclidean distance calculations. For each factor, three benchmark companies have been selected for comparison.

Cloud Solution

In Table 6.8, the cloud solutions of three selected companies are presented for this specific use case. While the target company utilizes a private cloud solution, Company 1 lacks a cloud solution, Company 2 employs a public cloud, and Company 3 operates a hybrid cloud solution. It's anticipated that Company 1 and Company 2 will collectively be most aligned with the target company based on the similarities in their cloud solutions, leaving Company 3 as the least similar.

The expected result is confirmed in Figure 6.8. Company 1 and Company 2 are closest to the target company, each with a distance of 0.25. Conversely, Company 3, with a distance of 0.5, is the most distant from the target company. This alignment with initial expectations indicates that both the normalization and the Euclidean distance were calculated correctly.

During the evaluation of this use case, it was observed that the calculations were initially flawed. The target company was consistently compared to itself, leading to a distance of

0 for every company. This use case enabled the identification and rectification of a bug in the program.

Company Name	Company ID	Cloud Solution
Target Company	n/a	Private
Company 1	1	None
Company 2	2	Public
Company 3	3	Hybrid

Table 6.8: Cloud Solution Factor - Companies

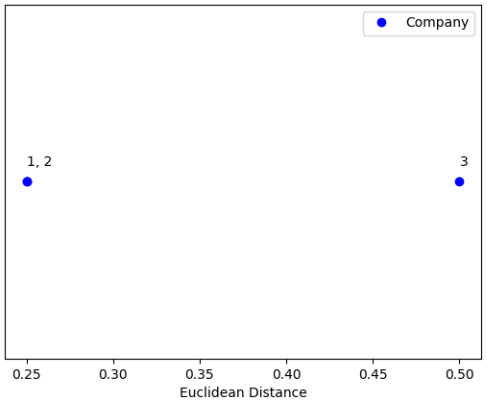


Figure 6.8: Cloud Solution Factor - Euclidean Distance

Multi-factor Authentication

The data in Table 6.9 indicates that neither the target company nor Company 1 implement multi-factor authentication, while Company 2 and Company 3 do. As a result, Company 1 is anticipated to have a distance of 0, aligning with the target company's lack of multi-factor authentication. Conversely, Company 2 and Company 3 are expected to have a distance of 1 due to their utilization of multi-factor authentication.

Company Name	Company ID	Multi-factor Authentication
Target Company	n/a	None
Company 1	1	None
Company 2	2	Multifactor
Company 3	3	Multifactor

Table 6.9: Multi-factor Authentication Factor - Companies

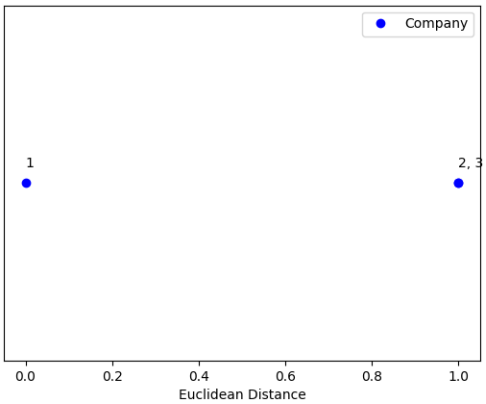


Figure 6.9: Multi-factor Authentication Factor - Euclidean Distance

The results presented in Figure 6.9 offer a clear insight into the evaluation process. As initially projected, Company 1 aligns perfectly with the target company, resulting in a distance value of 0. On the other hand, both Company 2 and Company 3 exhibit a distance of 1. This indicates that they are less similar to the target company. The consistency

between the anticipated outcomes and the actual results lends significant confidence to the accuracy and reliability of the calculations performed.

Network Infrastructure

Table 6.10 reveals the network infrastructures adopted by various companies. Specifically, the target company utilizes a LAN network infrastructure. In contrast, Company 1 employs a WAN network infrastructure. Both Company 2 and Company 3 have chosen a LAN setup, mirroring the choice of the target company. From this data, it is anticipated that Company 2 and Company 3 will register a distance value of 0, due to their infrastructural alignment with the target company. Meanwhile, Company 1, with its unique WAN setup, is expected to have a distance value of 1.

Company Name	Company ID	Network Infrastructure
Target Company	n/a	LAN
Company 1	1	WAN
Company 2	2	LAN
Company 3	3	LAN

Table 6.10: Network Infrastructure Factor - Companies

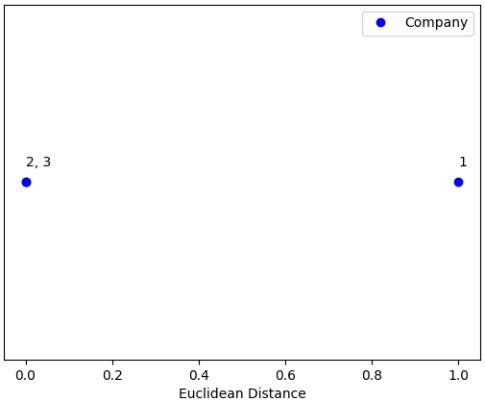


Figure 6.10: Network Infrastructure Factor - Euclidean Distance

The findings displayed in Figure 6.10 align with the expectations. Both Company 2 and Company 3 have an Euclidean distance of 0, indicating their similarity to the target company. In contrast, Company 1 has a greater Euclidean distance, highlighting its divergence from the target's network infrastructure.

Remote Access

Table 6.11 reveals that both the target company and Company 2 lack remote access, while Company 1 and Company 3 utilize VPN for remote access. Given these details, it is anticipated that Company 2 will have a distance of 0, while Companies 1 and 3 should each register a distance of 1.

The results in Figure 6.11 validate the initial assumption. Company 2 aligns perfectly with the target company, evident by its distance of 0, since neither employ remote access. On the other hand, both Company 1 and Company 3 have a distance of 1, given their use of a remote access method that differs from the target company's. This consistency

in outcomes affirms the correctness of both the normalization process and the Euclidean distance calculation for this specific factor.

Company Name	**Company ID**	**Remote Access**
Target Company	n/a	None
Company 1	1	VPN
Company 2	2	None
Company 3	3	VPN

Table 6.11: Remote Access Factor - Companies

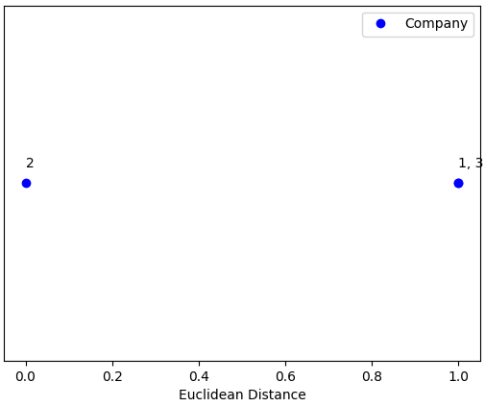


Figure 6.11: Remote Access Factor - Euclidean Distance

6.2 Scenarios

Three distinct scenarios are presented below, where similar companies are sought for a target company across all three dimensions. In each scenario, a database containing 1'000 diverse companies is utilized as the foundation for comparison. Once similarities among the companies are identified, the top two most similar companies are further analyzed. Additionally, the outcomes of two correlation measures are juxtaposed to examine their similarities and differences.

For all scenarios, it is considered an hypothetical company named Alanga AG, a company headquartered in Germany. The company is keen on leveraging the recently implemented tool to pinpoint another enterprise that mirrors its own in terms of revenue, organizational size, market share, and growth trajectory. By meticulously studying the shared data and insights from this analogous company, Alanga AG intends to make well-informed adjustments to its cybersecurity budget, ensuring it aligns with industry best practices and standards. More detailed information regarding Alanga AG is provided below.

Alanga AG
Country: GER
Revenue: $ 7 million
Market Share: 9%
Growth Rate: 4%
Organization Size: Small
Remote: 45%
Cybersecurity Budget: $ 7'000
Cybersecurity Investment: $ 6'000
Cybersecurity Training Investment: $ 0
Cybersecurity Insurance Investment: $ 1'000
Cybersecurity Staffing: 3
Cybersecurity Attack Threat: Man-In-The-Middle
Cloud: Public
Multifactor: None
Network Infrastructure: LAN
Remote Access: None

6.2.1 Business

For this scenario, Alanga AG is in pursuit of companies with similarities in their business factors and search criteria. Consequently, the company will seek counterparts within the business dimension. Figures 6.12 and 6.13 display the search outcomes based on the Euclidean distance and the Pearson Correlation, respectively.

In both figures, the top 10 most analogous companies are displayed. Upon examining the two figures more closely, it is observed that the company with ID 796 ranks first in Euclidean distance (values close to zero are better) but places third in Pearson correlation

(values close to one are better). Conversely, the company with ID 861, which tops the Pearson correlation, is not found within the top 10 for Euclidean distance. Another company, with the ID 472, appears in the top 10 for both metrics. To gain a clearer understanding of these varied outcomes, a deeper exploration into the profiles of the two most similar companies for each correlation measure is necessary.

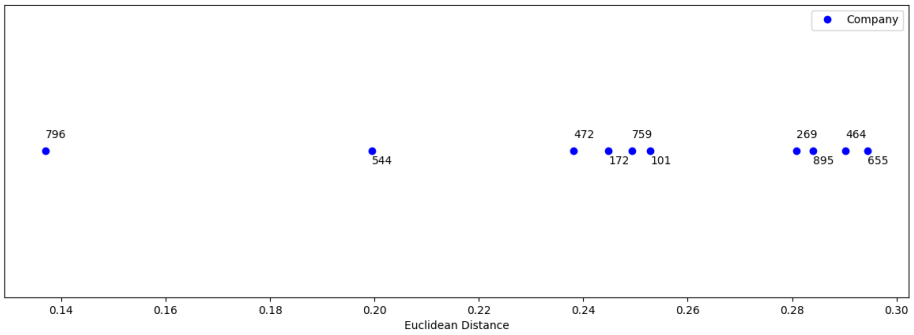


Figure 6.12: Similar Business Companies - Euclidean Distance

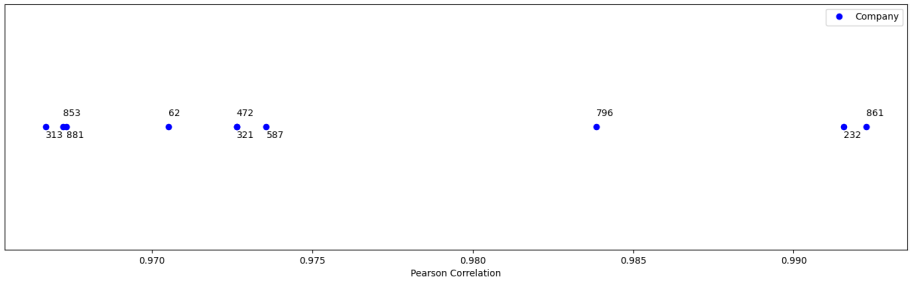


Figure 6.13: Similar Business Companies - Pearson Correlation

Tables 6.12 and 6.13 display the top two companies ranked by Euclidean distance and Pearson correlation, respectively. The initial column provides details about Alanga AG, while the final column presents the average values. According to Table 6.12, Company 796 is the closest match to Alanga AG based on Euclidean distance, with Company 544 ranking as the second closest. On the other hand, as depicted in Table 6.13, Company 861 emerges as the closest match based on Pearson correlation, followed by Company 232.

When examining the similarities between Alanga AG and Company 796, which is identified as the most similar to Alanga AG based on Euclidean distance, the factors *Market Share* and *Organization Size* are identical to Alanga AG's values. Additionally, the attributes *Country*, *Growth Rate*, and *Remote* closely resonate with Alanga AG's data. Nevertheless, there is a variance of $ 1.5 million in the "Revenue" factor relative to Alanga AG.

Company 544, which ranks second in Euclidean distance, is based in Germany, similar to Alanga AG. On the other hand, its discrepancy in revenue is more pronounced than that of Company 796. Moreover, when evaluating the *Growth Rate* and *Remote* factors, Company 544 diverges more from Alanga AG compared to Company 796. In summary, Company 544 aligns closely with Alanga AG in terms of the *Country* attribute, but it

does not fare as well in the *Revenue*, *Market Share*, *Growth Rate*, and *Remote* factors in comparison to Company 796.

Table 6.12: Business Comparison Results - Euclidean Distance

	Alanga AG	**Company 796**	**Company 544**	**Average**
Correlation Measure Rank	-	1	2	-
Country	GER	ITA	GER	TUR (20%)
Revenue	$ 7'00'000	$ 5'547'726	$ 5'239'487	$ 116'656'900'817
Market Share	9%	9%	7%	48%
Growth Rate	4%	9%	-22%	-3%
Organization Size	Small	Small	Small	Large (31%)
Remote	45%	40%	60%	54%

Table 6.13: Business Comparison Results - Pearson Correlation

	Alanga AG	**Company 861**	**Company 232**	**Average**
Correlation Measure Rank	-	1	2	-
Country	GER	TUR	UK	TUR (20%)
Revenue	$ 7'00'000	$ 881'740'146	$ 610'612'322	$ 116'656'900'817
Market Share	9%	10%	2%	48%
Growth Rate	4%	73%	-10%	-3%
Organization Size	Small	Medium	Medium	Large (31%)
Remote	45%	80%	47%	54%

Upon examining Companies 861 and 232, as outlined in Table 6.13 and deemed most similar based on the Pearson correlation, their resemblance to Alanga AG is not immediately apparent. Company 861 is classified as a medium-sized organization, and its revenue is notably higher than Alanga AG's. Additionally, across all factors, the values surpass those of Alanga AG. However, given that the values are consistently higher by a similar degree relative to Alanga AG, the Pearson correlation identifies Company 861 as the most similar to Alanga AG. A similar pattern emerges for Company 232, though its *Growth Rate* is an exception, being lower than that of Alanga AG.

Comparing the average values with Alanga AG reveals that these values deviate the most among all previously analyzed companies, primarily due to significantly higher revenues and the associated organizational size. Additionally, the market share considerably exceeds that of Alanga AG. The only similarity lies in the number of remote employees, which is comparable to Alanga AG's count.

Based on these comparisons, Company 796 emerges as the most aligned with the anticipated results, closely followed by Company 544; both were evaluated using the Euclidean distance metric. In contrast, the two companies assessed through the Pearson correlation significantly deviate from Alanga AG's profile to be deemed similar. A direct comparison

between the average values and Alanga AG underscores the unsuitability of relying solely on average values for such assessments.

In the context of business dimension comparisons, the Euclidean distance metric seems more reliable than the Pearson correlation, given its ability to identify companies closer to the reference entity. Aligning with this observation, Alanga AG also identifies Company 796 as the most analogous entity and subsequently adjusts its cybersecurity budget to mirror that of Company 796.

6.2.2 Economic

In a subsequent phase, Alanga AG aims to tailor the Breach Probability Function (BPF) [24, 22] to better suit its needs. They currently employ a standardized BPF function that does not align seamlessly with their company's specific requirements. To refine this function, they're exploring companies that are similar in the economic dimension, given that these factors correlate with cybersecurity. The objective is to identify a company that closely mirrors Alanga AG in terms of cybersecurity aspects, and then adapt insights from this company to enhance the BPF function. The updated economic details about Alanga AG, which were adjusted based the prior scenario, are provided below.

Alanga AG - Updated
Cybersecurity Budget: $ 40'000
Cybersecurity Investment: $ 30'000
Cybersecurity Training Investment: $ 3'000
Cybersecurity Insurance Investment: $ 6'000
Cybersecurity Staffing: 5
Cybersecurity Attack Threat: Man-In-The-Middle

Figure 6.14 displays the top 10 companies most akin, from the economic perspective, to Alanga AG when assessed using Euclidean distance, while Figure 6.15 showcases the top 10 companies when evaluated through Pearson correlation. Observing Figure 6.14, one can discern two distinct clusters proximate to 0. The primary cluster, positioned close to 0, comprises four companies with IDs 274, 887, 790, and 702. Based on Euclidean distance metrics, Company 274 exhibits the least distance from Alanga AG, followed closely by Company 887.

Turning attention to Figure 6.15, which presents Pearson correlation outcomes, discernible clusters can again be seen. Notably, one cluster is particularly close to a value of 1 and consists of three companies. Among them, Company 234 demonstrates the highest correlation with Alanga AG, with Company 35 ranking second. Interestingly, Company 274 stands out as the only company appearing in the top 10 for both methodologies: it holds the foremost position in Euclidean distance and the fourth in Pearson correlation.

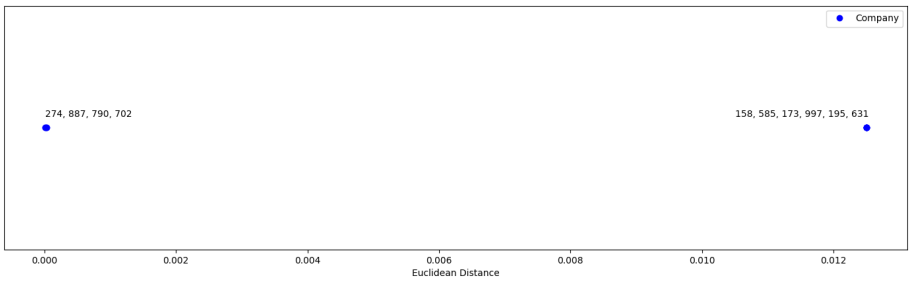


Figure 6.14: Similar Economic Companies - Euclidean Distance

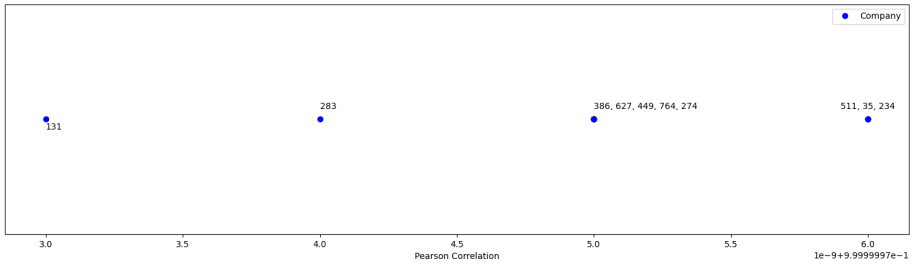


Figure 6.15: Similar Economic Companies - Pearson Correlation

Table 6.14 provides a consolidated view comparing Alanga AG with Company 274, Company 887, and the average values. When analyzing Company 274, which is closest to Alanga AG as per the Euclidean distance, it becomes evident that factors like *Cybersecurity Budget*, *Cybersecurity Investment*, *Cybersecurity Training Investment*, and *Cyber Attack Threat* align closely with or match the data for Alanga AG. The only deviation is seen in *Cybersecurity Insurance Investment*, which stands at roughly half the value reported for Alanga AG.

When analyzing the factors for Company 887, Company 274, and Alanga AG, it's evident that Company 887 aligns more closely with Alanga AG in terms of *Cybersecurity Budget* and *Cybersecurity Investment*. However, it diverges significantly from Alanga AG compared to Company 274 in *Cybersecurity Training Investment* and *Cybersecurity Insurance*. The relative closeness of the former two factors does not compensate for the disparity in the latter, leading to Company 887 securing the second position.

Table 6.15 contrasts the top two companies based on the Pearson correlation against Alanga AG and the average values. In assessing the similarity between Alanga AG and Company 234, which boasts the highest correlation with Alanga AG as per the Pearson correlation, it's evident that aside from the *Cyber Attack Threat* and *Cybersecurity Insurance Investment* factors, all other factors are approximately half the value of those for Alanga AG. The *Cybersecurity Insurance Investment* is merely a quarter of what's reported for Alanga AG, while the *Cyber Attack Threat* factor aligns perfectly with Alanga AG's value. Given that the majority of the factors are approximately half of Alanga AG's values, this company has been identified as having the highest correlation to Alanga AG.

Table 6.14: Economic Comparison Results - Euclidean Distance

	Alanga AG	**Company 274**	**Company 887**	**Average**
Correlation Measure Rank	-	1	2	-
Cybersecurity Budget	$ 40'000	$ 44'365	$ 35'844	$ 583'284'504
Cybersecurity Investment	$ 30'000	$ 33'274	$ 26'883	$ 437'463'378
Cybersecurity Training Investment	$ 3'000	$ 2'218	$ 1'792	$ 29'164'255
Cybersecurity Insurance Investment	$ 6'000	$3'105	$2'509	$40'829'915
Cyber Attack Threat	Man-In-The-Middle	Man-In-The-Middle	Man-In-The-Middle	Malware (21%)

Table 6.15: Economic Comparison Results - Pearson Correlation

	Alanga AG	**Company 234**	**Company 35**	**Average**
Correlation Measure Rank	-	1	2	-
Cybersecurity Budget	$ 40'000	$ 24'043	$ 601'373	$ 583'284'504
Cybersecurity Investment	$ 30'000	$ 18'032	$ 451'030	$ 437'463'378
Cybersecurity Training Investment	$ 3'000	$ 1'202	$ 30'068	$ 29'164'255
Cybersecurity Insurance Investment	$ 6'000	$ 1'683	$ 42'096	$ 40'829'915
Cyber Attack Threat	Man-In-The-Middle	Man-In-The-Middle	Man-In-The-Middle	Malware (21%)

When assessing Company 35, most factor values surpass those of Alanga AG, with the exception of *Cyber Attack Threat*, which matches Alanga AG's value. The majority of factors are ten to fifteen times those of Alanga AG. Due to this significant correlation, Company 35 holds the second rank. When comparing Alanga AG to the average values, it's evident that all factors are significantly higher than those of Alanga AG, and the *Cyber Attack Threat* factor also does not align.

From the residual analysis, Company 274 emerges as the closest to Alanga AG, followed by Company 887, both identified using the Euclidean distance. While the companies ranked first and second via the Pearson correlation are more similar to Alanga AG than

the average values, they are less so than those determined by the Euclidean distance. Thus, when seeking companies similar in the economic dimension, the Euclidean distance appears more reliable than the Pearson correlation. Alanga AG should avoid making decisions based on the average values, as they deviate significantly from its own metrics. Consequently, Alanga AG has recognized Company 274 as most aligned with its profile and will adjust its BPF in accordance with Company 274.

6.2.3 Technical

The latest scenario focuses on the technical dimension, where Alanga AG searches for companies with similar or identical technical specifications based on the technical information provided. Alanga AG uses a public cloud solution and has not implemented multi-factor authentication. Furthermore, their network infrastructure is based on LAN, and remote access is not supported. For this scenario, it is used the updated economic information of Alanga AG is used together with the initial information.

Figure 6.17 displays the top ten companies with the least Euclidean distance, while Figure 6.17 shows the top ten companies with the highest Pearson correlation. Notably, in the Euclidean distance metric, these ten companies have a distance of 0, indicating they are identical to Alanga AG. Similarly, in the Pearson correlation, these companies have a value of 1, signifying a perfect correlation with Alanga AG. Additionally, companies with ID 160, 511, 866, and 818 appear in both sets of results. The slight variations in results are attributed to the fact that more than ten companies have either a Euclidean distance of 0 or a correlation of 1, hence not all of them are displayed.

Table 6.16 presents a comparison among Alanga AG, Company 160, Company 818, and the average values. Given that all ten companies have a distance of 0, two companies were chosen randomly for comparison. Upon examining the respective factor values, it's evident that Alanga AG, Company 160, and Company 818 all share the same values. This implies that these companies are identical in the technical dimension.

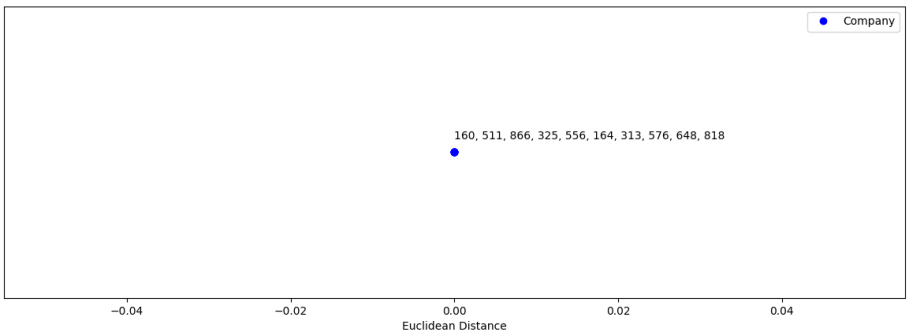


Figure 6.16: Similar Technical Companies - Euclidean Distance

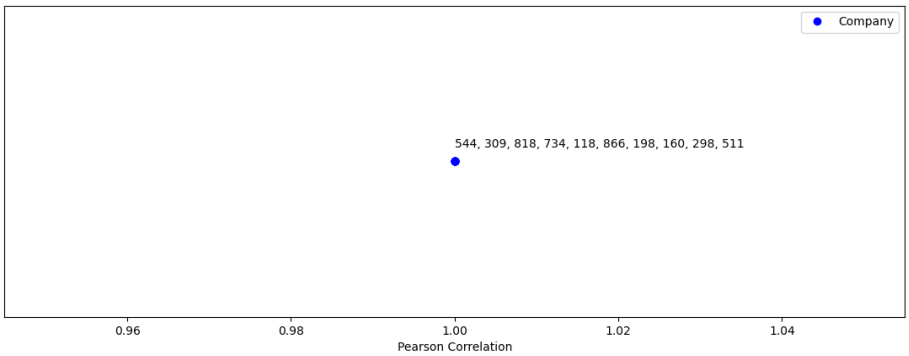


Figure 6.17: Similar Technical Companies - Pearson Correlation

Table 6.17 compares Alanga AG to Company 511, Company 866, and the average values. From the ten companies that received a Pearson correlation score of 1, Company 511 and Company 866 were randomly selected for this comparison. Similar to the observations from the previous table, all factor values match those of Alanga AG. This indicates that both Company 511 and Company 866 are technically identical to Alanga AG.

Table 6.16: Technical Comparison Results - Euclidean Distance

	Alanga AG	**Company 160**	**Company 818**	**Average**
Correlation Measure Rank	-	1	1	n/a
Cloud	Public	Public	Public	None (26%)
Multifactor	None	None	None	None (52%)
Network Infrastructure	LAN	LAN	LAN	WAN (35%)
Remote Access	None	None	None	VPN (51%)

Table 6.17: Technical Comparison Results - Pearson Correlation

	Alanga AG	**Company 511**	**Company 866**	**Average**
Correlation Measure Rank	-	1	1	-
Cloud	Public	Public	Public	None (26%)
Multifactor	None	None	None	None (52%)
Network Infrastructure	LAN	LAN	LAN	WAN (35%)
Remote Access	None	None	None	VPN (51%)

Upon closer examination of the average values in comparison to those of Alanga AG, it becomes clear that the *Mulfication* factor is the only one that corresponds precisely with Alanga AG. All the other factor values present a contrast to those observed in Alanga AG. This observation serves as a potent reminder of the inherent risks and potential inaccuracies when relying solely on average values for decision-making purposes.

When examining and contrasting the outcomes derived from the two correlation measures, it becomes challenging to definitively state which method produced more accurate results. This is mainly because both measures successfully pinpointed companies that have factors identical to those of Alanga AG. Given these observations, it can be inferred that in the realm of the technical dimension, both correlation measures demonstrate a similar level of efficacy and reliability.

6.2.4 Discussion and Limitations

Upon reviewing the various scenarios, it is evident that in each case, companies either identical or very similar to Alanga AG were identified. However, there was a distinction between the results derived from the Euclidean distance and the Pearson correlation. In the business and economic dimension, the Euclidean distance yielded more precise results, identifying companies more akin to Alanga AG, compared to the Pearson correlation. When it comes to the technical dimension, it is inconclusive which correlation measure is superior, as both managed to identify the same identical companies.

Examining the different factor values across various dimensions reveals that the business and economic factors predominantly consist of numeric values. In contrast, the technical dimension's factor values are strictly integer values, like 0, 1, or 2. This suggests that the Pearson correlation might yield more precise results when working with integer values. Based on this observation, it's recommended that greater emphasis be placed on the residuals of the Euclidean distance in the business and economic dimensions, while in the technical dimension, the residuals from both correlation measures should be equally considered.

As limitation of this work, we can highlight that the *Data Generator* was defined with arbitrary data that generates random values (within a define range) for all relevant business, economic, and technical factors. However, although those factors were carefully selected, the values being generated might be different of real-world scenarios since it is hard to define baselines for data generation. To address this issue, industry reports and approaches like [21] can be used in order to check the performance of the proposed approach in real-world sectors.

Chapter 7

Conclusions and Future Work

In the digital era we live in, the importance of cybersecurity is growing due to the rising number of cyberattacks year after year. It is crucial for companies, regardless of size, to prioritize investments in cybersecurity. However, the challenge lies in determining the best investment method, as there is no one-size-fits-all solution. A common approach to address this challenge is information sharing. Organizations can refine and adapt their cybersecurity strategies by accessing shared cybersecurity data from other companies.

To effectively benefit from shared information among companies, one must first identify which companies are relevant for consultation. It is important to discern which companies align closely with the ones seeking information. To determine this alignment, companies are compared across three dimensions: business, economic, and technical. Segmenting company information in this way allows for a more targeted comparison rather than a broad, holistic view. Specific factors have been established within each dimension, representing key attributes of a company that best define and characterize each dimension.

An application has been developed to calculate and visually present the similarities between various companies. Two correlation measures were employed to gauge these similarities: the Euclidean distance and the Pearson correlation measure. Upon a thorough evaluation comparing the outcomes of these measures, it was observed that the Euclidean distance yielded more accurate results in the business and economic dimensions than the Pearson correlation. By "more accurate," it is implied that companies deemed highly similar by the Euclidean distance shared more characteristics with the reference company than those identified by the Pearson correlation.

However, in the technical dimension, it was inconclusive as to which correlation measure was superior. Both methods identified companies having the same factor values as the reference company. Given that the technical dimension primarily comprises integer values, while the other two dimensions predominantly contain numerical values, it can be deduced that the Euclidean distance is more reliable for comparing numerical values relative to the Pearson correlation. For comparisons involving integer values, the outcomes from both correlation measures should be weighed with equal importance.

Currently, the method to identify similar companies uses a dataset of companies generated by the *Data Generator* and subsequently stored in the database. For a broader and more

real-world-oriented analysis, future work would involve substituting these database-stored companies with actual real-world companies. By adopting this approach, one can delve deeper into the practical implications and ascertain if the observations and findings are consistent and valid when contextualized in real-world scenarios and organizations. Such an endeavor would validate the existing methodology and furnish more comprehensive insights regarding its wider applicability.

The developed application, as it stands, operates independently without any interfaces to other systems. For future work, a logical step would be its integration into another platform, such as SecAdvisor. This would allow a company searching for similar entities to access and utilize shared data, including vital details about the BPF. By leveraging this BPF information, companies can modify and refine their own BPF. Such an integration would augment the application's functionality and significantly elevate the value of the SecAdvisor application for its user base.

Bibliography

[1] Saeed Ahmed, Anila Kousar, Noor Gul, Junsu Kim, and Su Min Kim. Euclidean distance-based machine learning scheme to detect vehicle hacking cyber-attacks. , pages 350–351, 2022.

[2] Mostofa Ahsan, Rahul Gomes, Md. Minhaz Chowdhury, and Kendall E. Nygard. Enhancing machine learning prediction in cybersecurity using dynamic feature selector. *Journal of Cybersecurity and Privacy*, 1(1):199–218, 2021.

[3] Charles Arthur. Businesses unwilling to share data, but keen on government doing it. June 2023, [Online] `https://www.theguardian.com/technology/2010/jun/29/business-data-sharing-unwilling`, last visit June 2023.

[4] Sirajuddin Asjad. The rsa algorithm. 12 2019.

[5] Li Kelly Olusola Bhavsar, Roy. Anomaly-based intrusion detection system for iot application. pages 2730–7239, 5 2023.

[6] Alexandru Boicea, Florin Radulescu, and Laura Ioana Agapin. Mongodb vs oracle – database comparison. In *2012 Third International Conference on Emerging Intelligent Data and Web Technologies*, pages 330–335, 2012.

[7] Jason Brownlee. Cyber information sharing: Building collective security. July 2023, [Online] `https://machinelearningmastery.com/distance-measures-for-machine-learning/`, last visit July 2023.

[8] Jacquelyn Bulao. How many cyber attacks happen per day in 2023? May 2023, [Online] `https://techjury.net/blog/how-many-cyber-attacks-per-day/#gref`, last visit May 2023.

[9] BYJU's. Euclidean distance. July 2023, [Online] `https://byjus.com/maths/euclidean-distance/`, last visit July 2023.

[10] Louis Columbus. 2023 cybersecurity forecasts: Zero trust, cloud security will top spending. May 2023, [Online] `https://venturebeat.com/security/2023-cybersecurity-forecasts-zero-trust-cloud-security-will-top-spending/`, last visit May 2023.

[11] The Cybersecurity and Infrastructure Security Agency (CISA). Automated indicator sharing (ais). June 2023, [Online] `https://www.cisa.gov/topics/cyber-threats-and-advisories/information-sharing/automated-indicator-sharing-ais`, last visit June 2023.

[12] Erica Galvez Edward Juhn. Incentivizing data sharing among health plans, hospitals, and providers to improve quality, Dec 2022.

[13] Leek J. T. Ellis, S. E. How to share data for collaboration. In *The American statistician*, page 53–57, 2018.

[14] Embroker Team. 2023 must-know cyber attack statistics and trends. May 2023, [Online] `https://www.embroker.com/blog/cyber-attack-statistics/`, last visit May 2023.

[15] P. Faber and R.B. Fisher. Pros and cons of euclidean fitting. In Bernd Radig and Stefan Florczyk, editors, *Pattern Recognition*, pages 414–420, Berlin, Heidelberg, 2001. Springer Berlin Heidelberg.

[16] Massimo Russo Tian Feng. The new tech tools in data sharing. July 2023, [Online] `https://www.bcg.com/publications/2021/new-data-sharing-tools-helping-companies-find-value`, last visit July 2023.

[17] Muriel Figueredo Franco. *CyberTEA: a Technical and Economic Approach for Cybersecurity Planning and Investment*. PhD thesis, Communication Systems Group (CSG), University of Zurich, February 2023.

[18] M. Franco, B. Rodrigues, and B. Stiller. MENTOR: The Design and Evaluation of a Protection Services Recommender System. In *15th International Conference on Network and Service Management (CNSM 2019)*, pages 1–7, Halifax, Canada, October 2019. IEEE.

[19] M. F. Franco, L. Z. Granville, and B. Stiller. CyberTEA: a Technical and Economic Approach for Cybersecurity Planning and Investment. In *36th IEEE/IFIP Network Operations and Management Symposium (NOMS 2023)*, pages 1–6, Miami, USA, 2023.

[20] Muriel Franco, Erion Sula, Bruno Rodrigues, Eder Scheid, and Burkhard Stiller. Protectddos: A platform for trustworthy offering and recommendation of protections. In Karim Djemame, Jörn Altmann, José Ángel Bañares, Orna Agmon Ben-Yehuda, Vlado Stankovski, and Bruno Tuffin, editors, *Economics of Grids, Clouds, Systems, and Services*, pages 28–40, Cham, 2020. Springer International Publishing.

[21] Muriel Figueredo Franco, Fabian Künzler, Jan von der Assen, Chao Feng, and Burkhard Stiller. Rcvar: an economic approach to estimate cyberattacks costs using data from industry reports. Preprint, jul 2023.

[22] Muriel Figueredo Franco, Christian Omlin, Oliver Kamer, Eder John Scheid, and Burkhard Stiller. SECAdvisor: a Tool for Cybersecurity Planning using Economic Models, 2023. 2304.07909, cs.CR, `https://arxiv.org/abs/2304.07909`.

[23] Louise Gaille. 12 advantages and disadvantages of correlational research studies. July 2023, [Online] `https://vittana.org/12-advantages-and-disadvantages-of-correlational-research-studies`, last visit July 2023.

[24] Lawrence A. Gordon, Martin P. Loeb, William Lucyshyn, and Lei Zhou. The impact of information sharing on cybersecurity underinvestment: A real options perspective. *Journal of Accounting and Public Policy*, 34(5):509–519, 2015.

[25] Maarten Grootendorst. 9 distance measures in data science. July 2023, [Online] `https://towardsdatascience.com/9-distance-measures-in-data-science-918109d069fa`, last visit July 2023.

[26] Katya Hill. How to share information with team members effectively. July 2023, [Online] `https://www.joinassembly.com/blog/how-to-share-information-with-team-members-effectively`, last visit July 2023.

[27] HIMSS. Information sharing: What is it? how to do it? why does it matter? July 2023, [Online] `https://www.himss.org/resources/information-sharing-what-it-how-do-it-why-does-it-matter`, last visit July 2023.

[28] IBM Corporation. Cost of a data breach 2022. May 2023, [Online] `https://www.ibm.com/security/data-breach`, last visit May 2023.

[29] Imperva. Anonymization. July 2023, [Online] `https://www.imperva.com/learn/data-security/anonymization/`, last visit July 2023.

[30] ISAO. Introduction to information sharing. July 2023, [Online] `https://ciasisao.org/standards-doc/isao-300-1/`, last visit July 2023.

[31] Zahra Jadidi, Shantanu Pal, Mukhtar Hussain, and Kien Nguyen Thanh. Correlation-Based anomaly detection in industrial control systems. *Sensors (Basel)*, 23(3), February 2023.

[32] Roemer J Janse, Tiny Hoekstra, Kitty J Jager, Carmine Zoccali, Giovanni Tripepi, Friedo W Dekker, and Merel van Diepen. Conducting correlation analysis: important limitations and pitfalls. *Clinical Kidney Journal*, 14(11):2332–2337, 05 2021.

[33] Zeng Zhu Jiang Chen Jiang, Shen. Cancelable biometric schemes for euclidean metric and cosine metric. pages 2523–3246, 2 2023.

[34] Zhiwei Jiang. The data-sharing advantage: A strategy for unrestricted innovation. July 2023, [Online] `https://www.forbes.com/sites/forbestechcouncil/2021/11/01/the-data-sharing-advantage-a-strategy-for-unrestricted-innovation/`, last visit July 2023.

[35] Oliver Kamer and Chrisitan Omlin. SECAdvisor 2.0: Visualizations and Extensions for Cybersecurity Economics Analysis, January 2023. Master Project, Communication Systems Group, Department of Informatics, Universität Zürich UZH, Zürich, Switzerland.

[36] Sai Praneeth Karimireddy, Wenshuo Guo, and Michael I. Jordan. Mechanisms that incentivize data sharing in federated learning. 2022.

[37] Peyman Kor. The danger of averages: Why data scientists shouldn't rely on average values for business decision making. June 2023, [Online] https://medium.com/geekculture/the-danger-of-averages-why-data-scientists-shouldnt-rely-on-average-values-for-business-decisions-b9969092b1d8, last visit June 2023.

[38] Roberta Kwok. When should companies share information with competitors? July 2023, [Online] https://insights.som.yale.edu/insights/when-should-companies-share-information-with-competitors, last visit July 2023.

[39] Jong-Ho Lee. Minimum euclidean distance evaluation using deep neural networks. *AEU - International Journal of Electronics and Communications*, 112:152964, 2019.

[40] Nate Lord. Data protection: Data in transit vs. data at rest. July 2023, [Online] https://www.digitalguardian.com/blog/data-protection-data-in-transit-vs-data-at-rest, last visit July 2023.

[41] Chavez Martínez. In defence of the simple: Euclidean distance for comparing complex networks, April 2018.

[42] Hiroki Miyahara. Solving for similarity using company exposures and euclidean distance. June 2023, [Online] https://venturebeat.com/security/2023-cybersecurity-forecasts-zero-trust-cloud-security-will-top-spending/, last visit June 2023.

[43] Renáta Myšková and Michal Kuběnka. Information sharing in the context of business cooperation – as a source of competitive advantage. *Journal of International Studies*, 12:169–182, 09 2019.

[44] Christian Omlin. A Gordon-Loeb-based Visual Tool for Cybersecurity Investments, January 2022. Bachelor Thesis, Communication Systems Group, Department of Informatics, Universität Zürich UZH, Zürich, Switzerland.

[45] Blue Pencil. What is data cleansing? July 2023, [Online] https://www.blue-pencil.ca/data-cleansing-what-is-it-and-why-is-it-important/, last visit July 2023.

[46] Lisa Perez. Why businesses aren't sharing more data. June 2023, [Online] https://theodi.org/article/why-businesses-arent-sharing-more-data, last visit June 2023.

[47] Juhi Ramzai. Clearly explained: Pearson v/s spearman correlation coefficient. July 2023, [Online] https://towardsdatascience.com/clearly-explained-pearson-v-s-spearman-correlation-coefficient-ada2f473b8, last visit July 2023.

[48] Rick G. Randall and Stuart Allen. Cybersecurity professionals information sharing sources and networks in the u.s. electrical power industry. *International Journal of Critical Infrastructure Protection*, 34:100454, 2021.

[49] Heiko Richter and Peter R. Slowinski. The data sharing economy: On the emergence of new intermediaries, Dec 2018.

[50] Margaret Rouse. Information sharing. July 2023, [Online] `https://www.techopedia.com/definition/24839/information-sharing`, last visit July 2023.

[51] Maryam Sulemani. Crud operations explained: Create, read, update, and delete, Apr 2021.

[52] Shaun Turney. Pearson correlation coefficient (r) | guide examples. July 2023, [Online] `https://www.scribbr.com/statistics/pearson-correlation-coefficient/`, last visit July 2023.

[53] World Economic Forum (WEF). Cyber information sharing: Building collective security. June 2023, [Online] `https://www3.weforum.org/docs/WEF_Cyber_Information_Sharing_2020.pdf`, last visit June 2023.

[54] Einat Weiss. How to convince customers to share data after gdpr. July 2023, [Online] `https://hbr.org/2018/05/how-to-convince-customers-to-share-data-after-gdpr`, last visit July 2023.

Abbreviations

AIS	Automated Indicator Sharing
API	Application Programming Interface
BPF	Breach Probability Function
CSR	Corporate Social Responsibility
CSV	Comma Separated Value
DDoS	Distributed Denial-of-Service
DNN	Deep Neural Network
ICS	Industrial Control Systems
IDS	Intrusion Detection System
GCD	Greatest Common Divisor
GDPR	General Data Protection Regulation
HTTP	Hypertext Transfer Protocol
IoT	Internet of Things
JSON	JavaScript Object Notation
ODM	Object Data Modeling
PII	Personally Identifiable Information
PoC	Proof-of-Concept
RCVaR	Real Cyber Value at Risk
REST	Representational State Transfer
RSA	Rivest-Shamir-Adleman
SDG	Sustainable Development Goals
SME	Small and medium-sized enterprises
SQL	Structured Query Language
URL	Uniform Resource Locator

List of Figures

List of Tables

Appendix A

Installation Guidelines

In order to run the application a recent version of Docker as well as Docker compose is required.

- Clone repository
  `git clone https://github.com/sec-advisor/cybersecurity-investment-tool.git`
- Checkout branch
  `git checkout feat/master-thesis`
- Build and start docker compose
  `docker compose up -build`