TU WIEN Informatics

Reasoning in Financial Knowledge Graphs

Making Industry Sectors Accessible to AI

DIPLOMARBEIT

zur Erlangung des akademischen Grades

Diplom-Ingenieur

im Rahmen des Studiums

Software Engineering & Internet Computing

eingereicht von

Manuel Schüller, BSc
Matrikelnummer 01426298

an der Fakultät für Informatik

der Technischen Universität Wien

Betreuung: Prof. Dr. Emanuel Sallinger
Mitwirkung: DI Markus Nissl
DI Aleksandar Pavlović

Wien, 15. März 2022

Manuel Schüller | Emanuel Sallinger

TU WIEN Informatics

Reasoning in Financial Knowledge Graphs

Making Industry Sectors Accessible to AI

DIPLOMA THESIS

submitted in partial fulfillment of the requirements for the degree of

Diplom-Ingenieur

in

Software Engineering & Internet Computing

by

Manuel Schüller, BSc
Registration Number 01426298

to the Faculty of Informatics

at the TU Wien

Advisor: Prof. Dr. Emanuel Sallinger
Assistance: DI Markus Nissl
DI Aleksandar Pavlović

Vienna, 15th March, 2022

Manuel Schüller | Emanuel Sallinger

Technische Universität Wien
A-1040 Wien ▪ Karlsplatz 13 ▪ Tel. +43-1-58801-0 ▪ www.tuwien.at

Erklärung zur Verfassung der Arbeit

Manuel Schüller, BSc

Hiermit erkläre ich, dass ich diese Arbeit selbständig verfasst habe, dass ich die verwendeten Quellen und Hilfsmittel vollständig angegeben habe und dass ich die Stellen der Arbeit – einschließlich Tabellen, Karten und Abbildungen –, die anderen Werken oder dem Internet im Wortlaut oder dem Sinn nach entnommen sind, auf jeden Fall unter Angabe der Quelle als Entlehnung kenntlich gemacht habe.

Wien, 15. März 2022

Manuel Schüller

Danksagung

An dieser Stelle möchte ich mich bei all jenen bedanken, die mich bei der Umsetzung dieser Diplomarbeit unterstützt und motiviert haben.

Zuerst möchte ich Prof. Dr. Emanuel Sallinger meinen Dank aussprechen, der meine Diplomarbeit betreut und begutachtet hat. Deine umfangreichen Erfahrungswerte, verlässliche Hilfe und konstruktive Kritik waren außerordentlich wertvoll. Vielen Dank zudem an DI Markus Nissl und DI Aleksandar Pavlović. Eure hilfreichen Anregungen und Kommentare haben diese Arbeit in vielerlei Hinsicht verbessert.

Ich danke dem gesamten Team des Knowledge Graph Labs und der Banca d'Italia für die stets produktive Zusammenarbeit und das mir entgegengebrachte Vertrauen. Ohne die überaus wertvollen Datensätze, die sie mir freundlicherweise zur Verfügung gestellt haben, wäre diese Arbeit nicht in dieser Form möglich gewesen.

Ein besonderer Dank gilt allen Teilnehmenden meiner Datenerhebung für ihren wertvollen Beitrag zur Evaluierung meiner vorgestellten Lösung.

Mein tiefster Dank gebührt darüber hinaus meiner Familie, die mich stets in allen Entscheidungen unterstützt hat und auf deren bedingungslose Unterstützung ich immer zählen kann. Ohne euch wäre ich nicht da, wo ich heute bin.

Abschließend möchte ich mich bei meiner Freundin bedanken, die mich vom ersten bis zum letzten Tag meines Studiums begleitet hat und die mir in allen Lebenslagen unerschütterlich zur Seite steht. Dein emotionaler Rückhalt und dein unfehlbares Gespür für Qualität waren von unschätzbarem Wert für diese Arbeit.

Acknowledgements

At this point, I would like to thank all those who supported and motivated me in the realization of this diploma thesis.

First, I would like to express my gratitude to Prof. Dr. Emanuel Sallinger, who supervised and reviewed my work. Your extensive experience, reliable support and constructive criticism were extremely valuable. Many thanks also to DI Markus Nissl and DI Aleksandar Pavlović. Your helpful suggestions and comments have improved this thesis in many ways.

I would like to thank the entire team of the Knowledge Graph Lab and the Banca d'Italia for the consistently productive collaboration and the trust placed in me. Without the highly valuable data they kindly made available to me, this thesis would not have been possible.

The participants of my dataset collection deserve special thanks for their valuable contribution to the evaluation of my proposed solution.

Furthermore, my deepest gratitude goes to my family, who has always supported me in all my decisions and whose unconditional support I can count on at all times. Without you, I would not be where I am today.

Finally, I would like to thank my girlfriend who has been with me from the first to the last day of my studies and who stands by my side in all situations. Your emotional support and infallible sense of quality have been invaluable to this thesis.

Kurzfassung

Wirtschaftszweigsystematiken wie beispielsweise NACE sind eine bewährte Methode zur Klassifizierung wirtschaftlicher Aktivitäten. Sie werden vielfach in der Wirtschaftswissenschaft, im Finanz- und Bankenwesen und in anderen Bereichen zur Gruppierung von Unternehmen eingesetzt, die ähnliche Produkte und Dienstleistungen anbieten und in ähnlichen Märkten operieren. Die meisten Staaten und Wirtschaftszonen nutzen individuell entwickelte Systematiken und forcieren ihre Nutzung, weshalb Unternehmensregister und ähnliche Datensätze oftmals derartige Klassifizierungen enthalten.

Da Wirtschaftszweigsystematiken jedoch üblicherweise als kategorische Codes strukturiert sind, sind sie für numerische Berechnungen und damit viele Anwendungsbereiche künstlicher Intelligenz ungeeignet. Das Potenzial der Branchenklassifizierungen, die viele der qualitativen Eigenschaften eines Unternehmens vereinen, wird dadurch nicht optimal ausgeschöpft. Beispielsweise könnten Behörden durch die Möglichkeit, feindliche Firmenübernahmen vorherzusagen, beim Schutz von Unternehmen mit hoher nationaler Relevanz unterstützt werden. Diese Anwendung umfasst Fragen wie „Welche der Tochtergesellschaften passt am wenigsten in ein gegebenes Unternehmenskonglomerat?“, deren Beantwortung fortschrittliche und wissenschaftlich evaluierte Metriken erfordert.

Derzeit gibt es keine etablierte, nicht-proprietäre Lösung, die diese große Lücke schließt. Die meisten bestehenden Ansätze sind zu vereinfachend und können daher die Nuancen zwischen Branchen nicht adäquat abbilden. Andere sind nicht wissenschaftlich fundiert und basieren auf Daten, die nicht öffentlich zugänglich sind, was ihre Bewertung erschwert.

In dieser Diplomarbeit werden fünf neuartige Methoden zur Quantifizierung der Ähnlichkeit von Wirtschaftszweigen vorgestellt, um bestehende Klassifizierungen für künstliche Intelligenz zugänglich zu machen. Die resultierenden Metriken werden sowohl hinsichtlich ihrer statistischen Eigenschaften als auch im Vergleich zu menschlichen Urteilen bewertet. Um ihre Anwendbarkeit zu verdeutlichen, wird zusätzlich eine Fallstudie durchgeführt.

Unsere Ergebnisse zeigen, dass die Validität und praktische Anwendbarkeit der vorgeschlagenen Metriken stark vom zugrunde liegenden Ansatz sowie der Qualität und Struktur der Eingabedaten abhängen. Insbesondere eine der Metriken erfüllt unsere Erwartungen an eine hochgradig valide und nützliche Ähnlichkeitsmetrik und schließt damit die oben genannte Lücke. Die Ergebnisse der Fallstudie untermauern das hohe Anwendungspotenzial unserer Lösung, da sie in der Lage ist, eine feindliche Firmenübernahme ausschließlich mithilfe der Branchenklassifizierung zu erkennen.

Abstract

Industry taxonomies such as NACE have long been the method of choice for classifying economic activities. They are universally used in economics research, finance, banking, and other areas for grouping similar companies based on the products and services they offer and the markets they operate in. Most countries and economic zones have established their own scheme and enforce its use, which leads business register datasets to often include respective classifications.

However, since industry classification systems are commonly structured as sets of categorical codes, they are mostly unfit to be used for numerical computations as is needed for various artificial intelligence tasks. This is unfortunate especially because the industry classification encapsulates many of the qualitative properties of a company, making it an ideal feature candidate for automated reasoning and machine learning. For example, being able to predict hostile company takeovers supports public authorities in protecting essential enterprises that are of high national relevance. This application involves questions like "Which of its subsidiaries fits the least into a given company conglomerate?", which require advanced metrics based on a scientifically verified approach.

Currently, there is no established, non-proprietary solution that closes this large gap. Most existing approaches are too simplistic and thus fail to convey nuances between industries. Others are non-academic and based on data not available to the public, which makes them difficult to evaluate.

In this thesis, we propose five novel ways of quantifying the similarity between industry sectors so that existing classifications are made accessible to artificial intelligence. The resulting metrics are evaluated both with regards to their statistical properties as well as how they compare to human judgements. Additionally, we conduct a case study in order to exemplify and validate their applicability.

Our results show that the validity and practical applicability of the proposed metrics strongly depend on the underlying approach as well as the quality and structure of the input data. One of the metrics in particular outperforms all others and meets our expectations of a highly valid and usable industry similarity metric, which indeed closes the aforementioned gap. Assessing the case study revealed the high potential of our solution for practical applications, as it is able to detect a hostile company takeover with no information about the involved companies except for their respective industries.

Contents

CHAPTER 1

Introduction

Many economics-related research and business areas require taking a high-level perspective on the economy, for example when monitoring and predicting national growth rates, controlling governmental subsidies and regulations, or assessing supply chain interdependencies. For such activities, it is necessary not to treat companies as individual economic entities but to group them based on mutual properties and behavior. A common way to achieve this is by classifying companies based on the industry sector they operate in. Businesses within the same industry are usually similar in many regards, such as the markets they operate in, the products and services they offer, their employment practices, and how they react to policy changes.

Industry taxonomies have long been the method of choice for implementing such classifications in a standardized manner [BLO03]. They are universally used in economics research, political decision-making, finance, and other domains that generate and consume macroeconomic data [PO16]. Without the existence of well-established industry taxonomies, it would often be impossible to integrate datasets of different origins.

Most countries and economic zones have established their own schemes and enforce their use, which is why business register datasets often include respective classifications by default. For example, all countries within the European Union use the *Statistical classification of economic activities in the European Community* (NACE) standard or some extension of it to classify economic activities [EU08b]. At its core, NACE is a set of codes in a hierarchical structure with each of its four levels being more specific than the one above it. At the highest abstraction level, it distinguishes between 21 main industries, such as “Agriculture, forestry and fishing” and “Financial and insurance activities”. At the most granular level, it distinguishes between a total of 615 industries, for example “Growing of rice” and “Growing of sugar cane”. Every industry is given a unique alphanumerical NACE code.

What makes industry taxonomies such as NACE so remarkable is that they facilitate the reduction of all the common characteristics that companies within the same industry share, which are manifold and potentially hard to measure, to a single dimension. For many practical applications, being provided only with companies' industry codes is sufficient to draw conclusions and make informed decisions. This makes industry taxonomies an indispensable tool for gaining an overall understanding of the economy.

1.1 Problem Statement

Since industry classification systems are typically structured as sets of categorical codes, they are mostly unfit to be used for numerical computations as is needed for various artificial intelligence tasks. For example, consider the following questions:

> *Which of its subsidiaries fits the least into a given company conglomerate?*
>
> *How common is it for a company X to own shares of companies operating in industry Y?*
>
> *If an industry X suffers significant losses due to a global pandemic, which are the five industries that will be affected the most as a consequence?*

It is evident that industry classifications alone are not sufficient to answer questions like these in an automated manner. Different metrics are needed to quantify the relationships between industries.

In particular, there is a substantial research gap regarding the concept of industry similarity and how to quantify it. This is unfortunate especially since being able to express how similar any two industries are as a numeric value could open up a new range of opportunities to automatically process and analyze all the datasets that already contain industry classifications. For example, the research department of the Italian central bank (*Banca d'Italia*), our long-term collaborative partner, expects to be able to derive new knowledge from their company ownership graph by augmenting it with industry similarity metrics. The graph contains data of more than three million Italian companies as well as the ownership relationships among them. It covers the largest part of the nation's economy and is used for various kinds of macroeconomic analysis and decision-making. We describe it in further detail in Section 2.2.1.

A specific use case the Banca d'Italia is interested in is the prediction of hostile foreign company takeovers [BBC+20]. This denotes an ownership change of a company that has previously been controlled by entities of its own country and is now controlled by some foreign entity. An entity, in this context, may be a private or institutional investor or simply another company. Foreign company takeovers are of special interest to government agencies since one of their goals is to prevent enterprises that are considered strategic

to national interests falling into outside control. A way to prevent such takeovers is to prematurely detect them and to enforce intervention policies. Applying industry similarity metrics on a large-scale company knowledge graph facilitates this detection by finding companies that are relatively dissimilar to their parent companies. These sub-companies are considered to be more likely to be sold off than those that lie within the core business of their respective company group [HOR87].

However, there are various challenges that complicate the creation of comprehensive and accurate industry similarity metrics. Most importantly, there is no standard definition of similarity, let alone industry similarity, that is universally agreed upon. This fact, which we expand on in Chapter 3, leads to industry similarity metrics being hard to conceptualize and even harder to evaluate properly. A comprehensive solution needs to take many different aspects into account such as the inherent qualitative information of industry taxonomies, the semantics of textual industry descriptions, the relationship between companies of different industries, and more abstract factors like supply chain interdependencies. The evaluation of such metrics needs to cover both their favorable and unfavorable statistical properties as well as how they compare to the human intuition of industry similarity. The latter is particularly challenging as there is, to the best of our knowledge, no "gold standard" dataset available that could be used for this purpose. Therefore, a comprehensive evaluation inevitably requires the additional effort of gathering a collection of human judgements and assembling a test dataset beforehand.

In conclusion, there are many challenges when it comes to conceptualizing, implementing, and evaluating industry similarity metrics. However, there is significant real-life demand for such metrics that so far has not been met.

1.2 Related Approaches

In this section, we briefly present existing approaches that aim to quantify industry similarity as well as their shortcomings with regards to the requirements described above. We discuss them in further detail in Section 3.

Truncated Industry Codes are commonly used when handling data that already includes industry classifications [BM14, Wun92, MS03]. The idea is to simply ignore the more granular levels of the respective taxonomy and instead group industries very coarsely. Then, if two industries are part of the same general industry, they are assigned a similarity score of 1. If not, they are assigned a similarity score of 0. For example, the industries "Marine fishing" (NACE code `A.03.11`) and "Freshwater aquaculture" (NACE code `A.03.22`) are both included in the more general industry "Fishing and aquaculture" (NACE code `A.03`) and are therefore assigned a similarity score of 1. In contrast, the score between "Marine fishing" and "Central banking" (NACE code `K.64.11`) is 0. The simplicity of this approach makes it easily applicable but unable to reflect any nuances between industries. It only allows for similarity scores of either 0 or 1, which is not how similarity is commonly perceived.

Industry2Vec[1] is an open-source implementation for creating vector representations of industry classifications. These vectors are obtained by training a neural network based on a combination of the truncated industry codes approach described above, word embeddings of textual industry descriptions, and private company ownership data. The similarity score between any two industries is equivalent to the cosine similarity of both their vector representations. Overall, Industry2Vec anticipates some of the concepts of the approaches proposed by this thesis. However, the way it generates industry vectors leaves room for improvement. Apart from the disadvantages of using truncated industry codes that have been discussed before, Industry2Vec models company ownership relations as tree structures, even though graph structures would be closer to reality. More abstract factors, such as supply chain interdependencies or mutual economic contribution, are not taken into account at all. Also, Industry2Vec has not been covered in academic publications and lacks publicly available evaluation data, which makes its validity difficult to assess.

Industry Similarity via Jaccard Index[2] is based on a simple coefficient used to calculate the similarity of two finite sets. The sets, which in the context of this approach represent industries, are comprised of keywords associated to the respective industry. Computing the similarity between two industries is consequently as simple as dividing the number of keywords they have in common by the total number of keywords that have been assigned to any of them. Although this approach is plausible, there is currently no way of reproducing its results, as it is heavily dependent on the underlying data, which is not available to the public. Similar to Industry2Vec, it also has not been covered in scientific publications and no evaluation data has been released.

In conclusion, there is no established, non-proprietary solution to quantifying industry similarity that closes the aforementioned gap. Most existing approaches are too simplistic and thus fail to convey nuances between industries. Others are non-academic and heavily depend on data not available to the public, which makes them difficult to evaluate.

1.3 Research Questions

As we show in Chapter 3, the currently existing industry similarity metrics are unsatisfactory. This thesis aims to propose a superior solution by exploring and implementing more advanced methods of quantifying industry similarity and taking different data sources into account. In order to give a guideline to the remaining thesis, we define three key research questions:

[1] https://www.sun-analytics.nl/posts/2018-09-06-industry2vec-an-implementation-for-industry-code-vector-representation/ (last accessed 15.03.2022)

[2] https://axialcorps.wordpress.com/2015/05/01/industry-similarity-via-jaccard-index/ (last accessed 15.03.2022)

RQ 1: How can the similarity of industries be quantified?

The metrics presented in Section 3.2 approach the underlying task of quantifying industry similarity in different ways. While all of them are interesting approaches in and of themselves, they are not without shortcomings and leave room for improvement. Novel and more intricate metrics that utilize a broader spectrum of data sources and processing methods could facilitate artificial intelligence tasks.

RQ 2: How does the proposed solution compare to the human notion of industry similarity?

As can be deduced from Section 3.1, there is no "gold standard" definition of industry similarity that is universally agreed upon. This complicates the assessment of newly conceptualized metrics. Gathering a collection of human judgements regarding the similarity between industries will potentially facilitate the evaluation and a deeper understanding of the proposed metrics and their validity.

RQ 3: Which applications can industry similarity metrics be used for in practice?

We present a variety of applications for quantified industry similarities in Chapter 3. To reinforce the significance of research in this field, examining further real-life use cases is beneficial. Additionally, this allows for the evaluation of the proposed metrics and their usability from the perspective of a potential user.

1.4 Methodology and Main Contribution

Our main contribution is conceptualizing, implementing, evaluating, and applying different ways of quantifying the similarity between industry sectors. In order to accomplish this and provide solutions to the research questions stated above, we perform the following steps:

- **Literature research:** An extensive research of academic literature regarding the current state of the art, theoretical concepts, and any other related work is the foundation of this thesis. In particular, we examine industry classification standards such as NACE, definitions and applications of financial knowledge graphs, the notion of similarity from the perspective of psychology and artificial intelligence, and currently existing industry similarity metrics.
- **Prototype implementation**: In order to provide a solution to RQ 1, we conceptualize and implement five similarity metrics based on different approaches and data sources. Our goal is to pre-compute numerical similarity values for all combinations of industry sectors. Although we focus our implementation exclusively on the NACE standard, the underlying approaches can be readily applied to any hierarchically structured industry classification scheme.

In the following list, we briefly present the proposed approaches. They are discussed in detail in Chapter 4.

- **Tree Distance (M1)** views the industry taxonomy as a tree data structure for which paths between industry codes can be calculated. The shorter the path between two industries, the more similar they are considered to be.
- **Description Similarity (M2)** compares the textual descriptions given by the industry taxonomy. The more similar the descriptions of two industries are, the more similar these industries are considered to be.
- **Integrated Ownership (M3)** uses an existing company ownership graph and aggregates its inherent knowledge. The more similar the ownership structures of companies belonging to two industries are, the more similar these industries are considered to be.
- **Supply Chain Interdependency (M4)** quantifies the interdependencies between industries due to them operating within the same supply chain. The more supply chains two industries participate in together, the more similar these industries are considered to be.
- **Economic Contribution (M5)** uses existing monetary-economic data and derives knowledge from the amount of financial transactions between industries. The more two industries contribute to each other economically, the more similar these industries are considered to be.

- **Statistical analysis**: We compare the proposed similarity metrics to each other using methods of descriptive statistical analysis. These include the assessment of their continuity as well as floor and ceiling effects.

  The goal for the analysis is to discuss the plausibility of each metric and assess whether some of them are more preferable than others regarding their statistical properties.

- **Data acquisition**: In order to be able to assess RQ 2, we have to acquire a dataset of human judgements regarding industry similarity with the help of academics and economics experts. This dataset serves as a "gold standard" of industry similarity that the metrics are compared to in order to evaluate their validity.

  To cover a broad spectrum of industries and opinions, we gather a dataset with more than 1,500 data points. The participants are primarily recruited from the members of the Joint Knowledge Graph Labs, which is a joint research project of TU Vienna, Oxford University, and Banca d'Italia. A large portion of the group's activities revolves around company knowledge graphs and big data in the economic governance domain.

- **Case study**: To address RQ 3, we conduct a case study of a real-life hostile foreign company takeover. This includes examining the general concept of hostile company takeovers, the details of the case in question, as well as methods that utilize our

industry similarity metrics and automated reasoning to prematurely detect such takeovers.

In particular, we propose, implement, and apply the following two takeover criteria:

- The **Parent Similarity (TC1)** criterion focuses on the similarity between a conglomerate's parent company and each of its subsidiaries. The subsidiary least similar to the parent company is considered most likely to be sold.
- The **Group Similarity (TC2)** criterion focuses on the similarities among a conglomerate's subsidiaries. The one that is on average least similar to all other subsidiaries is considered most likely to be sold.

In conclusion, we propose ways of quantifying the similarity between industry sectors so that existing classifications can be used for artificial intelligence tasks. We evaluate the resulting metrics both with regards to their statistical properties as well as how they compare to human judgements. Additionally, we conduct a case study in order to exemplify and validate their practical applicability.

1.5 Structure of the Thesis

In Chapter 2, we provide background information about the subjects relevant to this thesis, such as industry classification standards and knowledge graphs. In Chapter 3, we present and discuss state-of-the-art approaches of quantifying industry similarity. In Chapter 4, we thoroughly explain the concepts of our proposed industry similarity metrics and takeover prediction criteria. Details on the implemented prototypes are given in Chapter 5. In Chapter 6, we describe the evaluation of the proposed solutions as well as the utilized methodologies in detail. We also discuss and interpret the results and limitations. Chapter 7 concludes the thesis and provides an outlook for possible future work.

CHAPTER 2

Background

In this chapter, we provide background information about the subjects relevant to this thesis, which includes industry classification standards and knowledge graphs.

2.1 Industry Classification

Industry classification describes the effort of grouping companies, organizations, and other entities based on their economic activities and characteristics. These characteristics may include the markets they operate in, which products and services they offer, how they interact with the financial markets, and their employment practices [PLD05, UN08].

The criteria an industry classification scheme (*industry taxonomy*) is built upon highly depend on its designated use case as well as the available data. They are mainly developed and used by governmental organizations, business information providers, and academics for a variety of purposes, such as economics research, business analytics, risk management, and reporting [PO16]. A topical example of how regulators utilize industry classifications is the restriction of selected economic activities throughout the course of the Covid-19 pandemic, which permitted only those services considered essential to the public [BBG+20].

The industry classification schemes most relevant to this thesis are presented hereinafter.

2.1.1 NACE

The *Statistical classification of economic activities in the European Community*, abbreviated as *NACE*, is the scheme used by the European Union to classify economic activities [EU08b]. It is a standardized set of codes structured as a taxonomy with each

level being more granular than the one above it. The individual hierarchical levels are listed below[1]:

- Level 1: 21 *sections*, encoded as alphabetical letters (A - U)
- Level 2: 88 *divisions*, encoded as two-digit numbers (01 - 99)
- Level 3: 272 *groups*, encoded as three-digit numbers (01.1 - 99.0)
- Level 4: 615 *classes*, encoded as four-digit numbers (01.11 - 99.00)

Each code is accompanied by a textual description, as can be seen in Table 2.1, which also illustrates the hierarchical relationship between the codes. For example, section "`A` - Agriculture, forestry and fishing" covers a wide variety of agriculture-related business activities such as "`A.01` - Crop and animal production, hunting and related service activities", which in turn can be subdivided into "`A.01.1` - Growing of non-perennial crops", "`A.01.2` - Growing of perennial crops", and so on. Generally speaking, higher-level codes contain the activities of all lower-level classifications whose codes start with the same characters. Classifications on the same level are mutually exclusive. A complete list of all NACE codes can be found at the Eurostat metadata server[2].

Code	Level	Description
A	1	Agriculture, forestry and fishing
A.01	2	Crop and animal production, hunting and related service activities
A.01.1	3	Growing of non-perennial crops
A.01.11	4	Growing of cereals (except rice), leguminous crops and oil seeds
A.01.12	4	Growing of rice
A.01.13	4	Growing of vegetables and melons, roots and tubers
...		
A.01.2	3	Growing of perennial crops
A.01.21	4	Growing of grapes
A.01.22	4	Growing of tropical and subtropical fruits
...		
B	1	Mining and quarrying
B.05	2	Mining of coal and lignite
B.05.1	3	Mining of hard coal
B.05.10	4	Mining of hard coal
...		

Table 2.1: A small selection of NACE codes and their textual descriptions.

[1] `https://ec.europa.eu/eurostat/ramon/nomenclatures/index.cfm?TargetUrl=DSP_GEN_DESC_VIEW_NOHDR&StrNom=NACE_REV2&StrLanguageCode=EN` (last accessed 15.03.2022)

[2] `https://ec.europa.eu/eurostat/ramon/nomenclatures/index.cfm?TargetUrl=LST_NOM_DTL_LINEAR&StrNom=NACE_REV2` (last accessed 15.03.2022)

These codes and the economic activities they refer to are standardized for the whole European Union. Each member state may implement its own national adaption by extending the set of NACE codes with additional, more fine-grained levels. Due to its universality, NACE will be the primary classification standard used in this thesis.

2.1.2 ATECO

The *classificazione delle attività economiche (ATECO)* is the Italian adaptation of the European NACE standard [SSN09]. ATECO is a superset of NACE with equal codes referring to equal economic activities. Additionally, ATECO extends the scheme by two even more fine-grained levels:

- Level 5: 918 *categories*, encoded as five-digit numbers (01.11.1 - 99.00.0)
  e.g.: K.66.22.0 - Activities of insurance agents and brokers
- Level 6: 1227 *subcategories*, encoded as six-digit numbers (01.11.10 - 99.00.00)
  e.g.: K.66.22.01 - Insurance broker

As can be deduced from their codes, both examples given are descendants of the NACE class "`A.01.11` - Growing of cereals (except rice), leguminous crops and oil seeds".

ATECO is especially relevant to this thesis since the proposed solution is mostly based on data provided by the Banca d'Italia, which predominantly uses ATECO codes. However, in order to provide more generalizable findings, the hierarchy's lower levels will be omitted and the equivalent NACE codes will be used instead.

2.2 Knowledge Graphs

While the origins of knowledge graphs can be traced back to at least the 1980s [EW16], the recent interest in them can be attributed to a search engine enhancement technology by the same name introduced by Google in 2012[3]. The term "Knowledge Graph" has since been used in various different contexts and with diverging meanings by both academia and industry. Although uniting the existing definitions is therefore difficult, Figure 2.1 gives an overview of the key components knowledge graphs usually have in common [EW16].

In the context of this thesis, a knowledge graph can be considered a knowledge-based system (i.e., a knowledge base and a reasoning engine) with means to integrate various data sources. The individual components and their purpose will be described in detail below.

[3] `https://blog.google/products/search/introducing-knowledge-graph-things-not/` (last accessed 15.03.2022)

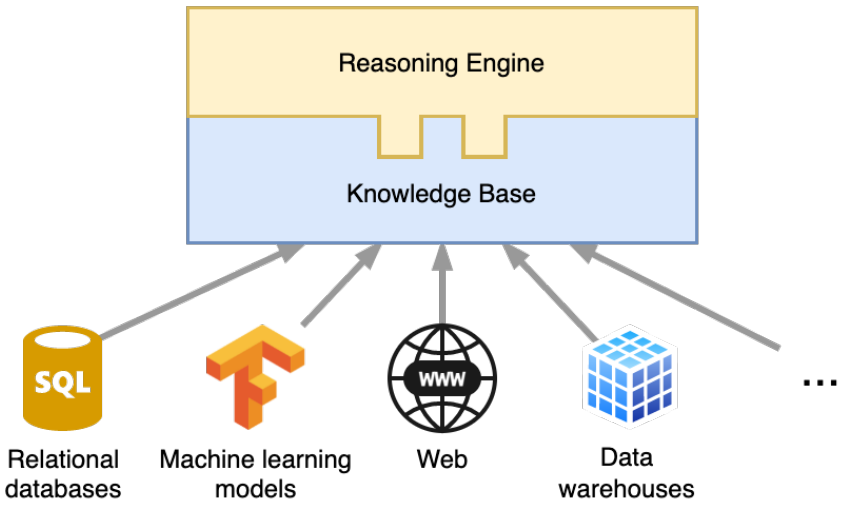


Figure 2.1: Key components of a knowledge graph

- **Knowledge Base:**

  The knowledge base of a knowledge graph can be considered a directed labeled graph that represents an ontology. The nodes and edges of the graph are formed by explicitly declared facts as well as implicit knowledge that is inferred from a set of rules.

  A *fact* is a declarative statement that is considered to be true in the problem domain. Conceptually, it can be compared to a database tuple. The following example shows five facts that together state that A, B, and C are companies, that A controls B, and that B controls C:

  $K = \{$
  $\quad Company(\text{“}A\text{”}),$
  $\quad Company(\text{“}B\text{”}),$
  $\quad Company(\text{“}C\text{”}),$
  $\quad Controls(\text{“}A\text{”}, \text{“}B\text{”}),$
  $\quad Controls(\text{“}B\text{”}, \text{“}C\text{”})$
  $\}$

  Figure 2.2 visualizes this knowledge base as an ontology graph, where each *Company* is depicted as a node and the *Controls* facts are depicted as edges between the respective nodes.

  Note that based only on these three facts, the statements “A controls B” and “B controls C” are true but the intuitively obvious statement “A controls C” is not.

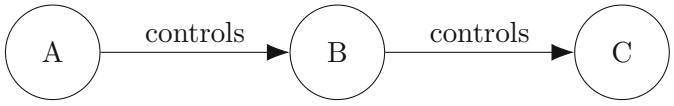


Figure 2.2: The graph visualization of a simple knowledge base

To reflect the transitivity in a control relationship, one could add the explicit fact *Controls*(“*A*”, “*C*”). However, a more suitable way of declaring such knowledge is by using rules.

A *rule* is a formal representation of general or domain-specific knowledge that allows the inference of implicit facts. In order to make the statement “*A* controls *C*” true as well, the following rule is added to the knowledge base:

$$Company(X) \land Company(Y) \land Company(Z) \land \\ Controls(X,Y) \land Controls(Y,Z) \rightarrow Controls(X,Z)$$

This rule states that if X, Y, and Z are companies, X controls Y, and Y controls Z, then X also controls Z.

Unlike conventional knowledge-based systems that have a single homogeneous knowledge base containing all facts and rules, a knowledge graph typically integrates multiple data sources and different means of knowledge representation.

- **Reasoning Engine:**

  The reasoning engine interprets the facts and rules integrated by the knowledge base and can derive new knowledge from them.

  Based on the examples given above, the following data would be inferred by the reasoning engine:

$$\begin{aligned} D = \{ & \\ & Company(\text{“}A\text{”}), \\ & Company(\text{“}B\text{”}), \\ & Company(\text{“}C\text{”}), \\ & Controls(\text{“}A\text{”}, \text{“}B\text{”}), \\ & Controls(\text{“}B\text{”}, \text{“}C\text{”}), \\ & Controls(\text{“}A\text{”}, \text{“}C\text{”}) \\ \} & \end{aligned}$$

  Figure 2.3 visualizes this knowledge as an ontology graph, where each *Company* is depicted as a node and *Controls* facts are depicted as edges between the respective nodes.

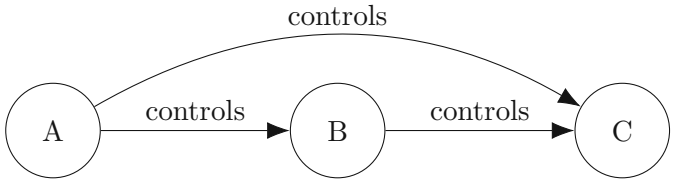


Figure 2.3: The graph visualization of both explicit and implicit knowledge

Apart from logic-based reasoning as shown above, knowledge graphs are often capable of performing embedding-based reasoning [BSV20b], probabilistic reasoning, and others, which, however, will not be covered by this thesis.

- **Data Sources:**

  Depending on the architecture and designated use case of an individual knowledge graph implementation, a knowledge graph integrates one or more data sources. These might be highly heterogeneous and include relational and graph database management systems, web scrapers, machine learning APIs, RDF stores, data warehouse platforms, and more. This allows users to link large amounts of data (*knowledge fragments*) from widely different sources and infer knowledge that would otherwise not be accessible [BSV20a, BBG+20].

  Considering the examples given above, instead of declaring facts explicitly, a knowledge graph would for instance be able to dynamically retrieve person records from a MySQL database and their relationships from a Neo4j graph database.

2.2.1 Applications

Nowadays, knowledge graphs are utilized by many major companies, such as Google, Yahoo, Microsoft, and Facebook [MTB+14]. Applicable use cases include item recommendation in online shopping [BSV20b], link prediction in social networks [WXWZ15], and conversational AI systems, such as speech assistants and chat bots [BSV20b], among many others.

Banca d'Italia Company Ownership Graph

A knowledge graph particularly important to this thesis is the company ownership graph built and maintained by the Italian central bank (*Banca d'Italia*), which acts as the primary data source for some of the proposed industry similarity metrics. It is based on a comprehensive business register containing data of more than three million Italian companies as well as the ownership relationships among them. It covers the largest part of the nation's economy and is therefore well suited as the basis for generalizable analysis.

Each company is described by several properties, such as its legal name, address, and ATECO code. Relationships between companies are either of the type `CONTROL` or `SHARE`. The first one represents any sort of relationship that allows one company to exercise

control over another, which makes them act like a single entity to some extent. The latter one represents one company owning shares of another, whereby the extent of the ownership is stored as a numeric property of the relationship edge. Further details on the structure of the knowledge graph have been covered by Atzeni et al. [ABI+20].

The company ownership graph is stored using Neo4j[4], which is an open-source graph database management system that uses the Cypher query language. Figure 2.4 shows an example query which finds all existing control relationships and returns the names and ATECO codes of the involved companies.

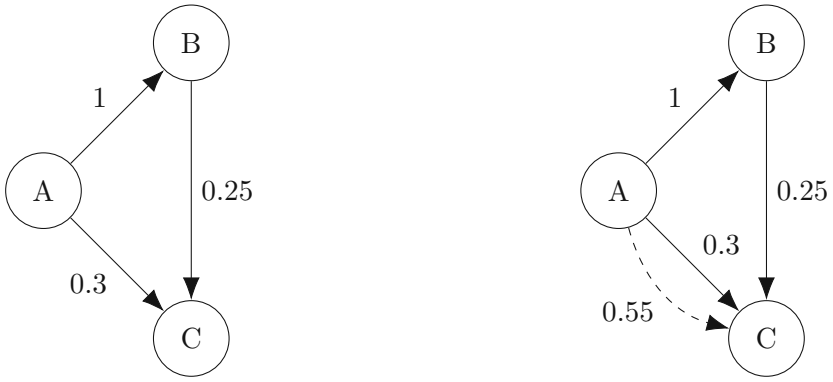
```
MATCH (c1:COMPANY)-[r:CONTROL]->(c2:COMPANY)
WHERE c1.C_ATECO_2007 <> '' AND c2.C_ATECO_2007 <> ''
RETURN c1.DENOMINAZIONE as ParentName, c1.C_ATECO_2007 as ParentCode,
    c2.DENOMINAZIONE as Name, c2.C_ATECO_2007 as Code
```

Figure 2.4: Example Cypher query for the Banca d'Italia company ownership graph

An example how the Banca d'Italia company ownership graph is used is the computation of *integrated ownership*. This figure denotes the accumulated ownership one company has over another through every direct and indirect shareholding. Figure 2.5 illustrates this concept by means of a simple company ownership graph. According to the initial graph which can be seen in Figure 2.5a, no single company seems to control (i.e., own more than 50%) of company C. However, due to the fact that company A controls company B, the ownerships of B need to be added to the ones of A in order to identify the true ownerships of A. Figure 2.5b shows the company ownership graph enriched with the integrated ownership. Here it is made evident that the true ownership of company A over C exceeds 50%.

The theoretical foundations of integrated ownership and some of its applications have been covered in detail by Bellomarini et al. [BBG+20].

[4] https://neo4j.com/ (last accessed 15.03.2022)



(a) A simple company ownership graph

(b) Dashed edges depict integrated ownerships

Figure 2.5: Example of integrated ownership. Nodes depict companies and edges depict ownerships.

2.2.2 Vadalog

Vadalog [BGPS17] is part of the *Value Added Data Systems* (*VADA*) project and denotes both a declarative logic programming language as well as a Knowledge Graph Management System (KGMS) that builds on this very language. VADA [KKA+17] is a research program initiated by the universities of Edinburgh, Manchester, and Oxford with the goal of facilitating the discovery, extraction, integration, access, and interpretation of data.

Since Vadalog has been used in pre-existing research on the Banca d'Italia company knowledge graph as well as other academic and industrial applications, it will be used for the implementation and evaluation of our proposed solution.

Language

The Vadalog language is a logic programming language suitable for knowledge representation used by the Vadalog KGMS to declare facts and rules. It is based on Datalog, more specifically Warded Datalog$^{\pm}$, which extends the base language with existential quantifiers in rule heads in order to enable ontological reasoning while maintaining decidability and tractability [BGPS17]. On top of that, Vadalog enhances its practical applicability by providing additional features, some of which are listed below.

- **Expressions**: Vadalog supports a variety of commonly used algebraic operations (e.g. sum, multiplication, and division of integers and floats), set and list operations (e.g. contains, size, union, intersection), string operations (e.g. substring, contains), boolean operations (e.g. and, or, not), conditions (e.g. equals, less than, greater or equal), and other expressions [BGPS18].

- **Aggregation**: Aggregation (e.g. min, max, sum, count) of numeric values is supported.
- **Data binding**: External data can be integrated via input, bind, and mapping annotations. This allows users to derive facts from various data sources, such as relational databases, graph databases, CSV files, data warehouses, and machine learning models. Similarly, the output of a Vadalog program can be written to different data sinks, most commonly the runtime's standard output or CSV files.
- **Post-processing**: The output of a Vadalog program can be post-processed, for example by ordering the resulting tuples based on one of the values or limiting the number of output tuples.
- **Embedding of external code**: Vadalog supports the execution of external Java and Python code. After declaring the source code and its interface, it can be invoked just like a regular Vadalog expression.

The program shown in Figure 2.6 exemplifies the Vadalog syntax. As can be seen, rules are written as Horn clauses of the form "`<head> :- <body>.`", which is read as "*<head> is considered to be true if <body> is considered true*". Note that rule heads appear on the left hand side of the `:-` symbol, which represents logical implied-by (⟵). The rule body appears on the right hand side and is a conjunction of one or more predicates separated by a comma. Predicate symbols start with lowercase letters and variables start with uppercase letters. All Vadalog statements end with a dot.

```
controls(X, Z) :- controls(X, Y), controls(Y, Z), company(X),
    company(Y), company(Z).
```

Figure 2.6: Vadalog program that defines a rule

Facts are written as clauses without bodies, as shown in Figure 2.7.

```
company("A").
company("B").
company("C").
controls("A", "B").
controls("B", "C").
```

Figure 2.7: Vadalog program that explicitly defines facts

In practice, facts are usually not defined explicitly but integrated from various data sources. For example, the Banca d'Italia company ownership graph is accessed directly through the Neo4j interface, as can be seen in the code snippet below.

```
@input("controls").
@qbind("controls", "neo4j", "", "MATCH
    (c1:COMPANY)-[r:CONTROL]->(c2:COMPANY) RETURN c1.DENOMINAZIONE as
    ParentName, c2.DENOMINAZIONE as Name").
@mapping("controls", 0, "ParentName", "string").
@mapping("controls", 1, "Name", "string").
```

Figure 2.8: Vadalog program that imports facts from a Neo4j database

This code imports all `CONTROL` relationships from the graph and makes them available as `controls(X, Y)` facts, where `X` is the name of the controlling company and `Y` is the name of the controlled company.

KGMS

The Vadalog KGMS is a system for building and maintaining knowledge graphs [BGPS17]. It was specifically designed to allow the integration of Big Data from heterogeneous sources, provides tools for machine learning and analytics tasks, and employs various query optimization techniques.

The architecture and reasoning engine of the Vadalog KGMS were built with the theoretical foundations of Warded Datalog$^{\pm}$ in mind. Compared to systems based on pure Datalog, this restriction allows for significant performance advantages in certain real-world and synthetic scenarios [BGPS18].

CHAPTER 3

Related Work

In this chapter, we provide an overview of the concept of similarity and present and discuss state-of-the-art approaches of quantifying industry similarity.

3.1 Similarity

Although this section does not intend to provide an exhaustive summary of all existing definitions of similarity, it aims at giving an overview that the remainder of this thesis can build upon.

In cognitive psychology, similarity may be defined as the number of environmental properties two persons, objects, concepts, or events have in common [Nob57, Cow17]. These properties include their appearance, typical usage, the context they are usually found in, their location, and the way they came into existence. However, it has explicitly been noted that the interpretation of these properties is highly subjective and different individuals frequently reach differing conclusions regarding the similarity of two objects [Wal58].

In psychology and linguistics, scientists study the process of categorization, which is the human ability to intuitively sort objects and persons into groups [Mat09]. Similarity is a primary factor in major explanatory models of categorization [Gol94, VAS04]. It has been criticised though that the notion of similarity is too unconstrained, context-dependent, based on perception, and thus fails to fully explain empirically observed categorization processes. Additionally, concepts and problems are sometimes categorized more in terms of common goals and solutions rather than their similarity alone [Gol94, Lov02].

In artificial intelligence and machine learning, similarity is used and studied in a wide variety of ways, such as similarity-driven reasoning [Ris06], information retrieval [HVV+06], classification [CGG+09], computer vision [BR94], and natural language processing [MCCD13]. In most of these areas there are different ways of quantifying similarity, with cosine

similarity and euclidean distance being notable examples. Apart from methodological differences, the results of applying AI and machine learning techniques heavily depend on the given input data. Neither domain expert knowledge nor training datasets are completely objective, free of bias, and indefinitely generalizable, which suggests that similarities resulting from them will at best be approximations of human judgements [BGK+18].

To recapitulate, there is no precise definition of similarity that is universally agreed upon and covers all its potential aspects. In the context of this thesis, the literature research lead to the following conclusions:

- An industry similarity metric will be more likely to resemble the human notion of similarity, the more properties it covers and the more diverse these properties are. Therefore, the proposed approach should take a multitude of properties into account.
- Taking more abstract factors than pure similarity into account might lead to more valid results. Whether two industries aim for a common goal, for example, could be assessed by analyzing if they operate within the same supply chain.
- Similarity is typically described as being a highly intuitive activity. Therefore, relying only on theoretical models to evaluate newly proposed industry similarity metrics will most likely be insufficient. Instead, human judgements should be taken into account.

3.2 Existing Industry Similarity Metrics

The following section gives an overview of currently existing industry similarity metrics as well as their respective strengths and shortcomings.

3.2.1 Truncated Industry Codes

One of the most common approaches of determining whether two industry sectors are similar is to compare their classification codes. Given that most industry classification schemes are hierarchical, it is possible to consider them only up to a specific level of detail, depending on the use case. All industries that lie within the same group are consequently deemed to be similar whereas others are not. This binary notion of similarity has been utilized in various contexts such as mergers and acquisitions (M&As) [BRG05, BM14], measuring regional economic diversity [Wun92, Wag00] and corporate diversification [MS03], as well as business angel investment allocation [BBS11].

Table 3.1 exemplifies the concept for four industries represented by their NACE codes, given a truncation at the second level.

	A.01.14	A.01.30	A.03.11	J.62.01
A.01.14	similar	similar	not similar	not similar
A.01.30		similar	not similar	not similar
A.03.11			similar	not similar
J.62.01				similar

Table 3.1: A selection of NACE codes and whether the industries they represent are considered similar based on their codes truncated at the second level. Symmetric values have been omitted for clarity.

For tasks that require numerical values, the similarity can be quantified as follows:

$$Similarity(x, y) = \begin{cases} 1 & \text{if } truncate(code(x), level) = truncate(code(y), level) \\ 0 & \text{otherwise} \end{cases}$$

where x and y are industry sectors, $code(\_)$ corresponds to their code representation in a specified industry classification scheme such as NACE, *level* is the chosen level of detail, and $truncate(\_, \_)$ cuts off the given code at the given level of detail.

The simplicity of this approach makes it easily applicable but discards information that might be helpful for analytics tasks. For example, "Freshwater fishing" (`A.03.12`), "Sale of other motor vehicles" (`G.45.19`), and "Manufacture of motor vehicles" (`C.29.10`) would all be considered equally dissimilar industries based on their truncated NACE codes. Intuitively though, the latter two are arguably more related to each other than to the first one. This shows the shortcomings of the described approach, especially when it comes to inter-sectional industry comparisons.

3.2.2 Industry2Vec

Industry2Vec is an open-source implementation for creating vector representations of industry codes[1]. It was developed by ING Wholesale Banking Advanced Analytics[2] in order to facilitate machine learning processes in the context of banking and finance. The industry classification standard it is based on is the *North American Industry Classification System (NAICS)* [US17], which is similar but not entirely compatible to the aforementioned NACE scheme.

Industry2Vec is highly relevant to this thesis as it offers a ready-to-use implementation as well as an output dataset, which are both publicly available. Its approach is based on a neural network that is trained to classify whether two given industries are similar or not. Each industry input vector consists of the index of its NAICS code concatenated with a vector representation of its textual description, where the latter was obtained

[1] `https://www.sun-analytics.nl/posts/2018-09-06-industry2vec-an-implementation-for-industry-code-vector-representation/` (last accessed 15.03.2022)

[2] `https://www.ing.com` (last accessed 15.03.2022)

by mapping the words to pre-trained GloVe [PSM14] embeddings. The two vectors of the industries in question are then fed into a classification layer, which computes their cosine similarity and labels them as similar or dissimilar using a sigmoid activation function. The target value for each industry pair is generated using a combination of private company ownership data and public morphological similarity:

$$Similarity(x, y) = 0.8 * S_c(x, y) + 0.2 * S_m(x, y)$$

$$S_c(x, y) = \begin{cases} 1 & \text{if any company in industry } x \text{ has a subsidiary in industry } y \\ 0 & \text{otherwise} \end{cases}$$

$$S_m(x, y) = \begin{cases} 1 & \text{if } truncate(code(x), 2) = truncate(code(y), 2) \\ 0 & \text{otherwise} \end{cases}$$

where x and y are industry sectors, $code(\_)$ corresponds to their NAICS code, and $truncate(\_, 2)$ cuts off the given code at the second level.

Training the network results in an eight-dimensional vector representation for each NAICS code that can be utilized for machine learning and other artificial intelligence purposes. Thus, determining the similarity between any two industries becomes as trivial as computing the cosine similarity of both vectors.

Industry2Vec can be seen as a direct competitor to the metrics proposed in this thesis. It is probable that the system was used productively for at least some time, which would be a indication of the concept's validity. However, Industry2Vec is not without notable shortcomings, some of which are listed below.

- Truncating the NAICS codes strictly at the second level eliminates much of the nuance that the classification standard offers on different levels of the taxonomy.
- Company ownership relationships are modeled as tree structures despite them being more resemblant to graph structures in real life. Apart from that, the extent of the ownership between companies and their subsidiaries is not considered.
- Other, more abstract factors such as supply chain interdependencies or mutual economic contribution are not taken into account at all.
- As mentioned before, Industry2Vec has not been covered in academic publications and the lack of publicly available evaluation data makes it difficult to assess without taking additional efforts.
- There is no one-to-one correspondence between NAICS and NACE codes. Moreover, pre-existing mappings between the two standards do not have a 100% coverage and lead to incomplete sets of NACE codes. The applicability of Industry2Vec for the European economic zone is therefore limited.

- The open-source repository[3] for Industry2Vec has been archived and is no longer being maintained. It is thus unclear to what extent its stakeholders are still confident in the quality and practicality of its core concept.

In conclusion, Industry2Vec is a credible method for quantifying the similarity of industries which employs strong approaches that other, more comprehensive similarity metrics can further build upon.

3.2.3 Industry Similarity via Jaccard Index

Another noteworthy industry similarity metric has been employed by the company Axial[4] in order to enhance the service's search functionality. The metric is based on the Jaccard Index, which is a simple coefficient used to calculate the similarity of two finite sets. It is defined as the ratio between the cardinality of the intersection of two sets and the cardinality of their union:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

The sets, which in the context of this approach represent industries, are comprised of keywords associated with the respective industry. These keywords are continuously created by users during their regular use of the company's services. Computing the similarity between two industries is consequently as simple as dividing the number of keywords both industries have in common by the total number of keywords that have been assigned to any of them. Trivially, this results in a similarity value between 0 and 1.

$$Similarity(x, y) = \frac{|Keywords_x \cap Keywords_y|}{|Keywords_x \cup Keywords_y|}$$

where x and y are industry sectors and *Keywords* corresponds to the set of keywords associated with the respective industry.

While the described approach is plausible, there is currently no way of reproducing its results, as it is heavily dependent on the underlying data, which is not available to the public. Similar to Industry2Vec, it has not been covered in scientific publications and no evaluation data has been released. Even though both its motivation and outcome resemble those of this thesis, the approach is thus hardly suitable for further investigation.

[3] `https://github.com/ing-bank/industry2vec` (last accessed 15.03.2022)

[4] `https://axialcorps.wordpress.com/2015/05/01/industry-similarity-via-jaccard-index/` (last accessed 15.03.2022)

CHAPTER 4

Approach

In this chapter, we present the conceptual solution approach to the research questions specified in Section 1.3 in detail. This includes the detailed description of the proposed industry similarity metrics as well as criteria for an exemplary use case, namely the prediction of hostile company takeovers.

4.1 Industry Similarity Metrics

In order to offer a selection of expressive industry similarity metrics that cover a broad spectrum of application areas, we take a variety of concepts and underlying data into account. These include approaches based on existing classification taxonomies, natural language processing (NLP), and empirically collected data, such as company ownership and mutual economic contribution.

The following list gives an overview of the proposed metrics, each of which will be covered in detail in the remainder of this section.

- **M1 - Tree Distance**

  This metric views a given industry taxonomy (e.g. NACE) as a tree data structure for which paths between industry codes can be calculated. The shorter the path between two industries, the more similar they are considered to be.

- **M2 - Description Similarity**

  This metric compares the textual descriptions given by an industry taxonomy (e.g. NACE). The more similar the descriptions of two industries are, the more similar these industries are considered to be.

- **M3 - Integrated Ownership**

  This metric uses an existing company ownership graph and aggregates its inherent knowledge. The more similar the ownership structures of companies belonging to two industries are, the more similar these industries are considered to be.

- **M4 - Supply Chain Interdependency**

  This metric quantifies the interdependencies between industries due to them operating within the same supply chain. The more supply chains two industries participate in together, the more similar these industries are considered to be.

- **M5 - Economic Contribution**

  This metric uses existing monetary-economic data and derives knowledge from the amount of financial transactions between industries. The more two industries contribute to each other economically, the more similar these industries are considered to be.

4.1.1 M1 - Tree Distance

The *Tree Distance* similarity metric is based on existing industry classification taxonomies such as NACE. It can be considered as a more refined version of the "Truncated Industry Code" approach presented in Section 3.2. Instead of merely representing the similarity of industries as two extremes and thus losing potentially valuable information, the Tree Distance metric expresses their relationship within the scheme in a more graduated manner.

For this, the industry taxonomy is viewed as a tree data structure, i.e., a connected acyclic undirected graph, where nodes constitute industry classes and edges connect them in order to form the classification hierarchy. This modeling allows the application of graph-theoretical concepts such as path finding [DD09], which can be utilized to calculate the distance between two industries within the hierarchy. The basic idea of this metric can thus be summarized as follows:

> *The shorter the path between two industries in the industry taxonomy, the more similar they are.*

In other words, if two industries are similar, this will most likely be reflected in their class codes being similar as well.

As an example, consider Figure 4.1, which shows a small part of the NACE industry taxonomy, depicted as a tree graph. In accordance to the definition of the Tree Distance similarity metric given above, the industries `A.01.1` and `A.01.2` are considered highly similar, as the path between them is just two edges long. In contrast, `A.01.1` and `A.02.1` are considered significantly less similar, as the path between them is five edges long.

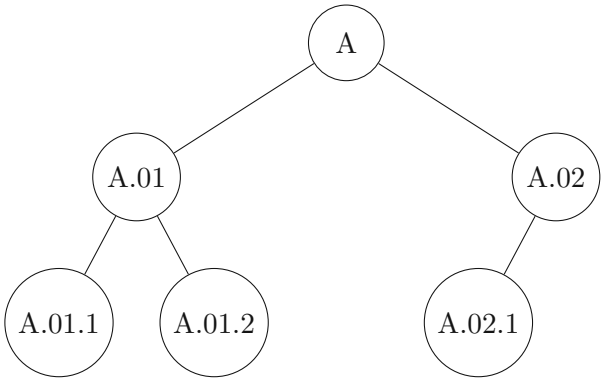


Figure 4.1: A part of the NACE industry taxonomy, depicted as a tree graph.

4.1.2 M2 - Description Similarity

Industry taxonomies such as NACE not only provide a hierarchical classification on the basis of codes, but also a textual description for each class. This is a valuable resource, as they are formulated and refined by regulatory experts who overlook a broad spectrum of markets and economic entities. Quantifying the semantics of these descriptions potentially allows to compute the similarity of any two industries. Thus, the basic idea of the *Description Similarity* metric can be summarized as follows:

> *The more similar the textual descriptions of two industries, the more similar the respective industries are.*

In other words, if two industries are similar, this will most likely be reflected in their descriptions being similar as well.

As an example, consider the short descriptions of the following three industries: “Fund management activities”, “Central banking”, and “Sale of motor vehicles”. In accordance to the definition of the Description Similarity similarity metric given above, the first two industries could intuitively be considered highly similar, as the components of both descriptions indicate a strong relation to the finance sector and usually appear in similar contexts. In comparison, the latter is considered significantly less similar to the other industries, intuitively. It is noteworthy that “Fund management activities” and “Central banking” do not have any particular words in common, from which a high similarity would be immediately derivable. Rather, the semantics of the given descriptions are compared directly.

4.1.3 M3 - Integrated Ownership

The approaches we have presented so far are based solely on existing classification standards. An important piece of information that is not covered by industry taxonomies is the ownership structure of actual companies operating in the industries in question.

Modern economies can be viewed as a complex network of companies owning or partially owning each other [RRT15]. Extracting and aggregating the knowledge implied in these relationships potentially makes the typical ownership structures of industry sectors comparable, not just the ones of individual companies. Thus, the basic idea of the *Integrated Ownership* similarity metric can be summarized as follows:

> *The more similar the ownership structures of companies belonging to two industries, the more similar these industries are.*

In other words, if two industries are similar, this will most likely be reflected in the ownerships of their companies being similar as well.

Consider the company ownership graph depicted in Figure 4.2. Each node represents a company and each edge represents an ownership of one company over another. Nodes are labeled as follows: `<company name>/<industry>`. For simplicity, the ownership percentages are omitted but are assumed to be equal. In accordance to the definition of the Integrated Ownership similarity metric given above, `A` and `B` are considered highly similar, as three out of four ownership relationships are between companies of these industries. In contrast, `A` and `C` are considered significantly less similar and `B` and `C` even less so.

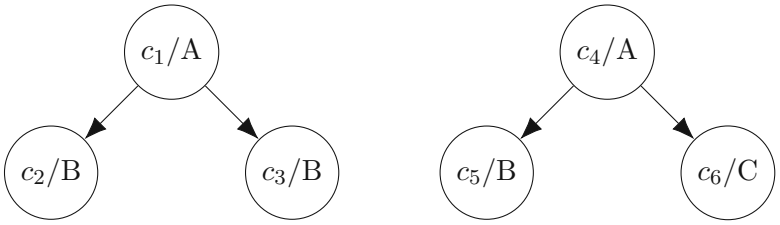


Figure 4.2: A sample company ownership graph. Two companies of industry `A` each own two companies respectively, three of them being of industry `B` and one of them being of industry `C`.

4.1.4 M4 - Supply Chain Interdependency

Companies cannot only be compared to one another other by their area of operation, but also by the supply chains they are involved in. Being part of the same supply chain indicates an interdependency due to the necessity of coordinating logistics, communication, processes, and regulatory compliance all while aligning all participants' economic interests [DHP04]. To some extent, these interdependencies lead towards utilization of and reliance on similar resources such as commodities, means of transport and communication, infrastructure, personnel, and others. Irregularities at one step of the supply chain may have severe consequences for all other companies dependent on it, as could be seen by the repercussions of the Covid-19 pandemic [BFF+20]. Theses concepts can not

only be applied to individual companies, but also to whole industries. Thus, the basic idea of the *Supply Chain Interdependency* similarity metric can be summarized as follows:

> *The more supply chains two industries participate in together, the more similar these industries are.*

In other words, if two industries are similar, this will most likely be reflected in them being part of mutual supply chains as well.

For instance, consider Table 4.1, which contains two exemplary supply chains and a selection of NACE industries that are part of them. In accordance to the definition of the Supply Chain Interdependency similarity metric given above, the industries `C.32.5` and `G.46.46` are considered similar, as they operate within a mutual supply chain. In contrast, `C.32.5` and `C.16.24` are not considered similar, as they do not share a common supply chain.

Supply chain	**Involved industries**
Chemistry	C.32.5 - Manufacture of medical and dental instruments and supplies
	G.46.46 - Wholesale of pharmaceutical goods
	G.47.73 - Dispensing chemist in specialised stores
	Q - Human health and social work activities
	...
Packaging	C.16.24 - Manufacture of wooden containers
	C.22.22 - Manufacture of plastic packing goods
	G.46.76 - Wholesale of other intermediate products
	N.82.92 - Packaging activities
	...

Table 4.1: Two exemplary supply chains and a selection of industries involved in each of them.

4.1.5 M5 - Economic Contribution

Economic interdependency of companies can not only be derived from whether they operate within the same supply chain, as proposed by M4, but also from mutual economic contribution. This contribution can be measured by the cash flows between individual companies as well as whole industries. Unlike supply chain based similarity metrics, which are heavily dependent on the methodology of specifying supply chains, economic contribution can thus be determined by analyzing unbiased financial data.

An extensive economic contribution between industries indicates a high reliance on similar resources and changes in market conditions likely have similar repercussions on both of them. Thus, the basic idea of the *Economic Contribution* similarity metric can be summarized as follows:

The more two industries contribute to each other economically, the more similar these industries are.

In other words, if two industries are similar, this will most likely be reflected in them transferring large amounts of funds among each other as well.

Table 4.2 exemplifies this concept. It shows the economic contribution among the three fictional industries `A`, `B`, and `C`. In accordance to the definition of the Economic Contribution similarity metric given above, `A` and `C` are considered to be highly similar industries, as their mutual contributions exceed those of all other industry pairs. In contrast, `A` and `B` are considered dissimilar due to them contributing comparatively low amounts to each other.

From \ To	A	B	C
A	100	20	1500
B	50	400	2500
C	4000	50	750

Table 4.2: Economic contribution between three fictional industries. All amounts are denoted in millions of euros.

4.2 Hostile Takeover Prediction

This section proposes an example for applying industry similarity metrics in practice, namely the prediction of hostile company takeovers.

During crises such as the COVID-19 pandemic, countries have an increased interest in protecting essential enterprises that are of national relevance, such as in healthcare. Such enterprises play an essential role in the state as a whole and countries need to prevent unlawful takeovers in order to maintain economic integrity and self-sufficiency. However, times of increased market pressure allow foreign entities to easily acquire large amounts of company shares due to low prices, potentially leading to fundamental ownership changes, violating legal frameworks. Thus, many countries have regulations and legal measures in place that enable their governments to prevent them from losing control over strategic enterprises [BBG+20].

However, enforcing these measures presumes the ability to sufficiently detect and predict transactions that might lead to such takeovers. This is by no means trivial, given that national economies form a vast and complex network with millions of legal entities, individuals, and ownership relationships. Additionally, reasons for companies being sold off are diverse, making predictions even more challenging [HOR87].

One promising approach is the analysis of *company conglomerates* regarding the industries they operate in. A conglomerate generally refers to a group of companies that tightly

co-operate. It usually consists of a parent company and multiple subsidiaries. Reasons for forming conglomerates are manifold, including reduction of market risks, increase of efficiency, and product and service improvements[1].

The rationale behind analyzing conglomerates in order to predict hostile takeovers is that in times of economic turmoil they tend to sell off sub-companies that do not contribute to their core business [HOR87]. These events of consolidation expose the sub-companies to the risk of being acquired by foreign investors, which might interfere with national interests, as outlined above. Enriching existing company knowledge graphs with industry similarity metrics allows to automatically reason about conglomerates and efficiently detect any subsidiaries that operate outside of their conglomerate's core businesses. Ideally, this might support authorities and other decision makers in identifying vulnerable strategic companies and taking preemptive measures to prevent any hostile takeover attempts.

Two methods for detecting such vulnerable companies will be proposed hereinafter.

4.2.1 TC1 - Parent Similarity

The *Parent Similarity* takeover criterion focuses on the direct relationship between a single parent company and its subsidiaries. The idea behind this criterion is that the subsidiary least similar to its parent company is most likely to be sold.

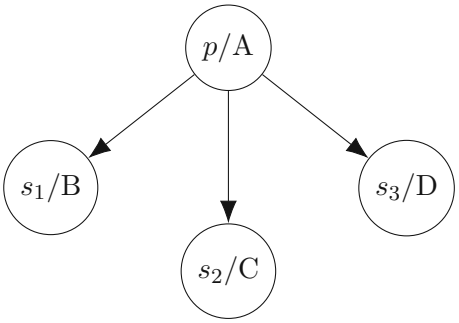
Figure 4.3 exemplifies the application of the Parent Similarity takeover criterion. The company ownership graph given in Figure 4.3a depicts a parent company p and multiple subsidiaries $S = \{s_1, s_2, s_3\}$. Nodes are labeled as follows: `<company name>/<industry>`. Table 4.3b shows values of a generic industry similarity metric as a matrix, whereby 0 equates minimal similarity and 1 equates maximal similarity. The result of augmenting the company ownership graph with these similarities is illustrated in Figure 4.3c. As can be seen, additional similarity edges have been inserted between p and each of its subsidiaries. The resulting knowledge graph can then be used to find the most likely takeover candidate by creating a simple ranking of all eligible companies. Table 4.4d shows that s_2 is considered to be the company most likely to be sold according to TC1.

4.2.2 TC2 - Group Similarity

The *Group Similarity* takeover criterion considers the company group as a whole. Unlike TC1, it largely ignores the parent company and instead focuses on the similarities among its subsidiaries. The idea behind this criterion is that the subsidiary that is on average least similar to all other subsidiaries is most likely to be sold.

Figure 4.4 exemplifies the application of the Group Similarity takeover criterion. The company ownership graph given in Figure 4.4a depicts a parent company p and multiple subsidiaries $S = \{s_1, s_2, s_3\}$. Node naming conventions are identical to those of Figure 4.3.

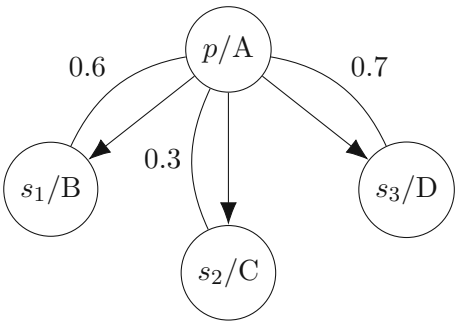
[1] `https://www.britannica.com/topic/conglomerate-business` (last accessed 15.03.2022)



(a) A sample company ownership graph containing four companies p, s_1, s_2, and s_3 that operate within the industries A, B, C, and D, respectively.

	A	B	C	D
A		0.6	0.3	0.7
B				
C				
D				

(b) A sample matrix of pre-calculated industry similarities. Symmetric values and values irrelevant to this example are omitted for clarity.



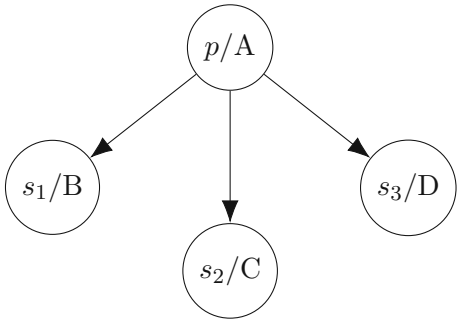
(c) The augmented company ownership graph with the similarities between the companies' industries visualized as edges.

Company	Similarity to p	Rank
s_2	0.3	1
s_1	0.6	2
s_3	0.7	3

(d) Ranking of all subsidiaries of p. The higher the subsidiary, the likelier it is to be sold.

Figure 4.3: Exemplary application of the TC1 takeover criterion.

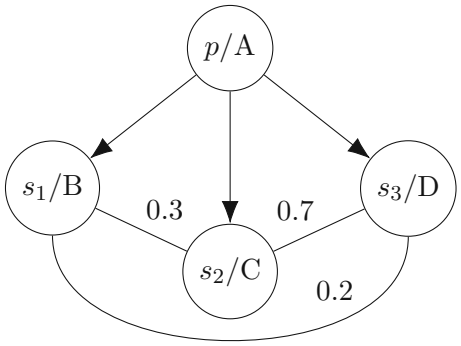
Table 4.4b shows values of a generic industry similarity metric as a matrix. The result of augmenting the company ownership graph with these similarities is illustrated in Figure 4.4c. As can be seen, additional similarity edges have been inserted between all subsidiaries. The resulting knowledge graph can then be used to find the most likely takeover candidate by creating a ranking of all eligible companies. Since each subsidiary has two adjacent similarity edges, its group similarity is obtained by calculating the average of both values. Table 4.4d shows that s_1 is considered to be the company most likely to be sold according to TC2.



(a) A sample company ownership graph containing four companies p, s_1, s_2, and s_3 that operate within the industries A, B, C, and D, respectively.

	A	B	C	D
A				
B			0.3	0.2
C				0.7
D				

(b) A sample matrix of pre-calculated industry similarities. Symmetric values and values irrelevant to this example are omitted for clarity.



(c) The augmented company ownership graph with the similarities between the companies' industries visualized as edges.

Company	Similarity to s_x	Rank
s_1	$\frac{0.3+0.2}{2} = 0.25$	1
s_3	$\frac{0.7+0.2}{2} = 0.45$	2
s_2	$\frac{0.3+0.7}{2} = 0.5$	3

(d) Ranking of all subsidiaries of p. The higher the subsidiary, the likelier it is to be sold.

Figure 4.4: Exemplary application of the TC2 takeover criterion.

CHAPTER 5

Implementation

In this chapter, we describe the implementation of the concepts proposed in Chapter 4. This includes the prototypes of the five industry similarity metrics and the two takeover prediction criteria.

5.1 Industry Similarity Metrics

This section describes the implementation of the concepts outlined in Section 4.1. Since this thesis focuses primarily on the European economic area, the implemented metrics will be based on the NACE classification scheme.

The goal of this step is to pre-compute numerical similarity values for all combinations of industry sectors. Essentially, the output of each metric will be a similarity matrix such as the one depicted in Table 5.1. Since the NACE standard specifies 996 different industry codes, the matrix will contain a total number of 992,016 values. Each one will be a decimal in the range of $[0, 1]$. The higher the value, the higher the implied similarity between the respective industries. Furthermore, we assume $Similarity(x, y) = Similarity(y, x)$ to hold, which means the matrix will be symmetric.

	A	**A.01**	**A.01.1**	...	**U.99.0**	**U.99.00**
A	0.84	0.69	0.59		0.30	0.30
A.01	0.69	0.82	0.70		0.27	0.27
A.01.1	0.59	0.70	0.83		0.23	0.23
...				...		
U.99.0	0.30	0.27	0.23		0.40	0.40
U.99.00	0.30	0.27	0.23		0.40	0.40

Table 5.1: The matrix visualization of a sample industry similarity metric. The axis labels are NACE industry codes.

5.1.1 Preparation of NACE Data

The European Statistical Office (Eurostat) offers an official listing of NACE codes[1], which is openly accessible. The following relevant fields are provided:

- **Code** ... The NACE code of the industry, e.g. `A.01.22`
- **Level** ... The level of the industry within the classification hierarchy
- **Description** ... The title of the industry, mostly consisting of only a few words, e.g. "Growing of tropical and subtropical fruits"; we rename this field to "Title"
- **This item includes** ... Details of which activities are included in the industry class, given either as regular text or as a list; we rename this field to "Details"

The NACE code notation used by Eurostat omits section letters for industry classes at level 2 or higher. However, since this impedes the immediate understanding of a code's affiliation, we normalize all NACE codes to the format `<Section>.<Division>.<Group><Class>`.

Table 5.2 shows the finished preprocessed NACE dataset that the remainder of this section builds upon.

[1] `https://ec.europa.eu/eurostat/ramon/nomenclatures/index.cfm?TargetUrl=LST_CLS_DLD&StrNom=NACE_REV2` (last accessed 15.03.2022)

	Code	Level	Title	Details
1	A	1	Agriculture, forestry and fishing	This section includes the exploitation of vegetal and animal natural resources, comprising the activities ...
2	A.01	2	Crop and animal production, hunting and related service activities	This division includes two basic activities, namely the production of crop products and production of ...
3	A.01.1	3	Growing of non-perennial crops	This group includes the growing of non-perennial crops, i.e. plants that do not last for more than two ...
4	A.01.12	4	Growing of rice	This class includes: • growing of rice (including organic farming and the growing of ...
			...	
995	U.99.0	3	Activities of extraterritorial organisations and bodies	
996	U.99.00	4	Activities of extraterritorial organisations and bodies	This class includes: • activities of international organisations such as the United ...

Table 5.2: Preprocessed NACE dataset

5.1.2 M1 - Tree Distance

We recall the intuition of this metric:

> *The shorter the path between two industries in the industry taxonomy, the more similar they are.*

The implementation of the Tree Distance similarity metric is directly based on the NACE classification hierarchy. At its core, it is a *shortest path metric* [DD09], which means it computes the length of the shortest path (i.e., the *geodesic distance* [HKP12]) between two nodes in a graph and derives a metric from it. In our case, the graph in question is the NACE taxonomy, the nodes are industries, and the derived metric is in the range of $[0, 1]$, where 0 corresponds to minimum and 1 corresponds to maximum similarity.

By default, the NACE standard separates industries into 21 top-level sections, which means that its taxonomy consists of 21 separate trees without any connection among each other. In order to be able to compare industries of different sections, we introduce an artificial root node with all section nodes as its immediate children. This level-0 node

is given the name `__ROOT__` in order to distinguish it from regular industry nodes. It enables paths to be formed between any two nodes of the whole taxonomy.

Since the taxonomy's data structure is a tree, the length of the shortest path between two nodes can be calculated by summing up the depth differences between each node and their lowest common ancestor. Algorithm 5.1, which shows the final implementation of M1, carries out this computation in the lines 1 to 3. To make the resulting value usable as a similarity metric, we normalize it to $[0, 1]$. This is done by dividing the computed path length by the maximum possible path length within the tree, i.e., twice the tree's depth (lines 5 to 6).

However, this naive approach is not robust when comparing codes at different levels of the hierarchy. For example, the computed similarity between `A.01.1` and `A.02` is lower than the one between `A.01` and `A.02`, which is unreasonable considering that `A.01.1` is included within `A.01`. Therefore, the implemented path length computation compensates for any depth differences of the given codes (line 4).

Algorithm 5.1: M1 - Tree Distance similarity

Input: *industries*: Tree<Industry>, $industry_1$: Industry, $industry_2$: Industry
Output: $similarity_{M1}$: Number in $[0, 1]$

1 $lca \leftarrow \text{lowestCommonAncestor}(industry_1, industry_2)$
2 $distance_1 \leftarrow industry_1.level - lca.level$
3 $distance_2 \leftarrow industry_2.level - lca.level$
4 $distance \leftarrow 2 * \min(distance_1, distance_2)$
5 $depth \leftarrow \max(industries.levels)$
6 $similarity_{M1} \leftarrow 1 - distance/(depth * 2)$

Figure 5.1 shows an example of the tree distance calculation between the industries `A.01.1` (level 3) and `A.02` (level 2). As can be deduced from the NACE codes, section `A` is their lowest common ancestor. The red colored nodes indicate the two codes in question. Since the distance measurement compensates for the level differences, `A.01` is chosen as the beginning of the path instead of `A.01.1`. To indicate the actual starting point of the path, the node of `A.01` is colored blue. The orange vertices indicate the shortest path between the codes.

The similarity calculation is broken down as follows:

$lca = \text{lowestCommonAncestor}(\texttt{A.01.1}, \texttt{A.02}) = \texttt{A}$
$distance_1 = \text{level}(\texttt{A.01.1}) - \text{level}(\texttt{A}) = 3 - 1 = 2$
$distance_2 = \text{level}(\texttt{A.02}) - \text{level}(\texttt{A}) = 2 - 1 = 1$
$distance = 2 * \min(2, 1) = 2$
$depth = 4$
$similarity_{M1} = 1 - distance/(depth * 2) = 1 - 2/8 = 0.75$

$\rightarrow \mathbf{similarity_{M1} = 0.75}$

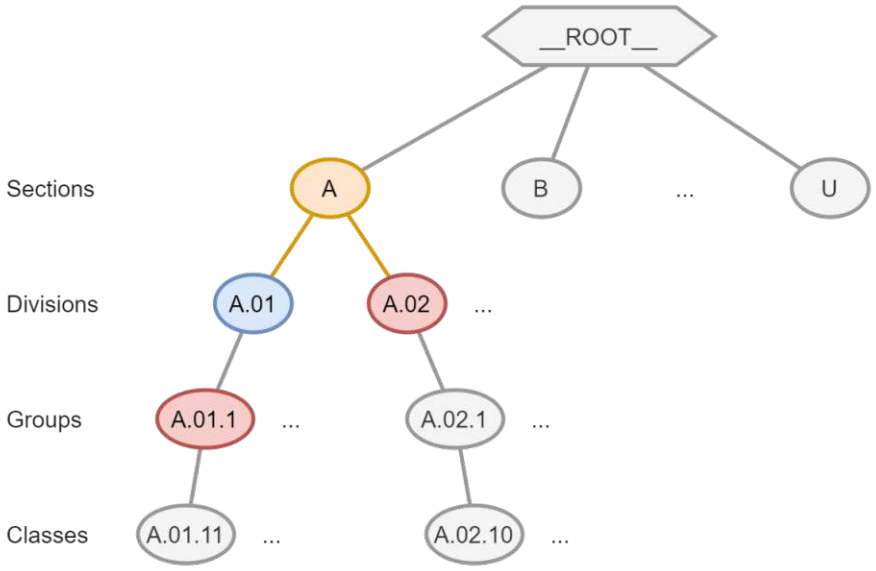


Figure 5.1: Calculation of the tree distance between the industries `A.01.1` and `A.02`.

5.1.3 M2 - Description Similarity

We recall the intuition of this metric:

> *The more similar the textual descriptions of two industries, the more similar the respective industries are.*

In order to quantify the semantics of textual industry descriptions, we employ methods of Natural Language Processing (NLP). More specifically, word embeddings allow individual words to be mapped to high-dimensional vectors that represent their meaning numerically. The semantic similarity of two words can then be computed by calculating the angle between their respective vectors.

Word2Vec [MCCD13] is a widely used algorithm to generate word embeddings. It is based on a shallow neural network that is trained to predict words using their neighboring words within a large text corpus. The hidden layer's weights of a sufficiently trained Word2Vec network can be interpreted as a high-dimensional vector. These vectors have multiple properties which make them usable as word embeddings. For this thesis, the most important characteristic is that Word2Vec vectors of semantically similar words are located closer to one another within the vector space compared to those of dissimilar words. Thus, the similarity of two words can be quantified by calculating the cosine similarity of their respective word embeddings.

Based on these foundations, the process of calculating the M2 industry similarity values is implemented as shown in Figure 5.2. Below, we describe the steps in detail.

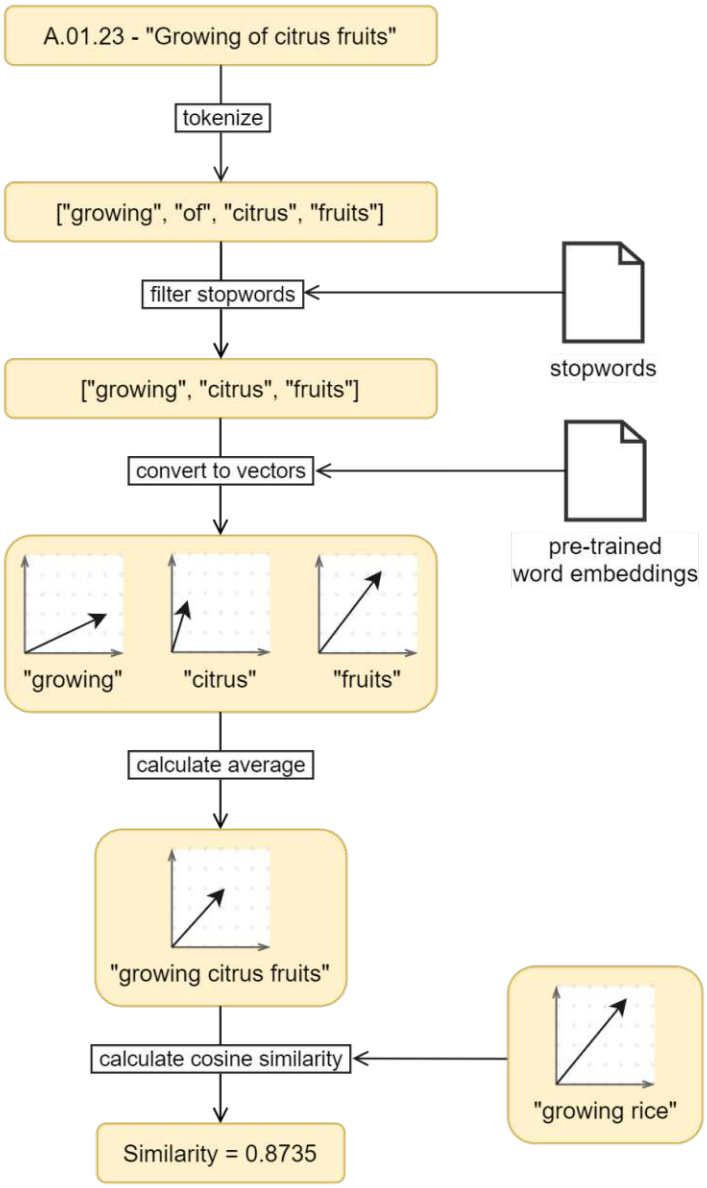


Figure 5.2: Calculation of the Description Similarity between the industries “Growing of citrus fruits” and “Growing of rice”

1. For each industry specified by the NACE standard, both its title and details field are split into separate words, which are then added to a vocabulary.

2. All words in the vocabulary are converted to vectors. For this, we use pre-trained Word2Vec embeddings provided by the Nordic Language Processing Laboratory (NLPL)[2]. The chosen dataset contains a vocabulary of 4,027,169 words and was trained on the English CoNLL17 corpus using the Word2Vec Continuous Skipgram algorithm and a window size of 10.

3. For each NACE industry, its title and details fields are concatenated. We refer to the resulting string as the *description* of the industry. A list containing the embeddings of all words occurring in the respective text is created. Stopwords, i.e., words that carry no inherent meaning, such as "a", "the", and "in" are omitted. The collection of English stopwords used in the implementation was obtained from the website of Ranks NL[3], which provides freely available stopword lists for multiple languages.

4. Based on the *Mean of Words Embeddings (MOWE)* [WTLB15] approach, each list is reduced to its mean in order to obtain a single vector representation of the whole description.

5. To determine the similarity between two NACE industries, the cosine similarity of their respective vectors is calculated. By definition, the output value is in the range $[0, 1]$. The higher the value, the more similar the two industry descriptions are.

It is noteworthy that there are more state-of-the-art approaches to computing document similarity than using MOWE with individual word embeddings. The following list presents the most important ones and the reasons why we have not considered them for our implementation of the M2 metric:

- Doc2Vec [LM14] is an extension of Word2Vec that additionally feeds document indices into the neural network during training. This way, document embeddings are produced instead of word embeddings. Since these document embeddings most likely represent semantics more accurate than the same document's mean word embedding, Doc2Vec could potentially lead to better results than the approach described above.

  Unlike Word2Vec however, there are no pre-trained Doc2Vec embeddings for all-purpose use, as they fit only the particular set of documents they were trained on. This means that in order to use Doc2Vec for the implementation of the Description Similarity metric, it would have to be trained from scratch using the NACE dataset. The problem with this approach is that NACE specifies merely 996 different industries, which is less than a tenth of the smallest training dataset used in the original Doc2Vec paper [LM14]. Therefore, Doc2Vec is not suitable for the purposes of this thesis.

[2] `http://vectors.nlpl.eu/repository/` (last accessed 15.03.2022)
[3] `https://www.ranks.nl/stopwords` (last accessed 15.03.2022)

- Bidirectional Encoder Representations from Transformers (BERT) [DCLT18] is an advanced language model that achieves state-of-the-art results in a variety of NLP tasks. Unlike Word2Vec, BERT does not simply provide a single embedding for each word in its vocabulary. Instead, a word's representation depends on its context within the input text. For example, BERT yields different results for the word "bank" depending on whether it is used to describe a financial institute or a river bank.

  The main reason BERT was not considered suitable for the M2 implementation is that it is not intended to be used with long input texts. Pre-trained BERT models usually have a limit of 512 tokens, which is shorter than many of NACE's detailed industry descriptions. Increasing this limit is usually considered impractical due to the computation complexity that grows quadratically in terms of input length. There are extensions to BERT that improve on this aspect, such as BERT-AL [ZWSC19] or DocBERT [ARTL19]. However, these either lack open source implementations or are intended for other use cases than similarity computation.

5.1.4 M3 - Integrated Ownership

We recall the intuition of this metric:

> *The more similar the ownership structures of companies belonging to two industries, the more similar these industries are.*

The foundation of the Integrated Ownership similarity metric implementation is an extensive knowledge graph that models companies and the ownership relations among them. The Banca d'Italia provided such a knowledge graph including data of more than three million Italian companies. Details about its content and structure have already been covered in Section 2.2.1.

Since many companies are organized as complex shareholding structures, transitive ownership relations are not apparent in the initial graph. Therefore, the *integrated ownership* between all companies is used instead. This key figure denotes the accumulated ownership that one company has over another through every direct and indirect shareholding, as discussed in Section 2.2.1. The integrated ownerships dataset used for the implementation of the M3 metric was also provided by the Banca d'Italia and is based on the *Baldone Ownership* defined by Bellomarini et al. [BBG+20].

Based on these foundations, the process of calculating the M3 industry similarity values is implemented as shown in Figure 5.3. Below, we describe the steps in detail.

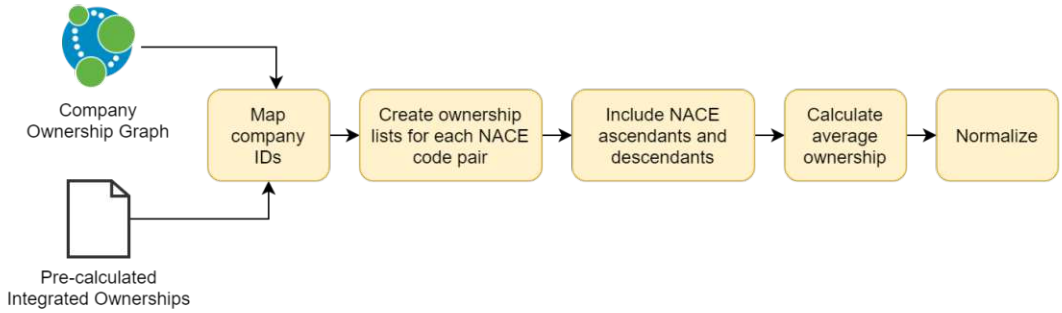


Figure 5.3: Calculation steps of the Integrated Ownership similarity

1. For each integrated ownership relation in the dataset, the NACE codes of both adjacent companies' IDs are looked up in the company knowledge graph.
2. For each found NACE code pair, a list of the percentages of all related integrated ownerships is created. The ownership direction is ignored while doing so, since it is irrelevant to the final similarity metric. This results in a symmetric matrix whose axes are NACE codes and whose values are percentage lists of differing lengths.
3. When looking up the ownerships between industry `X` and `Y`, what is implicitly asked for is not only the ownerships between `X` and `Y` specifically, but also those of their ancestors and descendants in the NACE hierarchy. To substantiate this claim, consider the following examples:
   - $company_1$ with $code_1$=`A.01` owns 30% of $company_2$ with $code_2$=`B.05`. When looking up the ownerships of `A` and `B.05`, this relation should be reflected in the result, since `A` encompasses `A.01` by definition. The same applies for the ascendants of `B.05`.
   - $company_1$ with $code_1$=`A.01` owns 30% of $company_2$ with $code_2$=`B.05`. Since it is not apparent from the data which specific sub-industry $company_1$ belongs to, any descendant of `A.01` (e.g. `A.01.1`, `A.01.11`, `A.01.2`, ...) might implicitly own `B.05`. This uncertainty is resolved by simply adding the ownership relation between $company_1$ and $company_2$ to all descendants of `A.01` and `B.05`.

   Therefore, the ownership list of each code pair is extended by all the ownership percentages of both code's ascendants and descendants.
4. In order to make the ownership lists comparable to each other, all the implicit 0% share relationships between industries need to be included into the ownership lists as well. Otherwise, a single 61% ownership between two industries A and B would result in a higher similarity than a hundred 60% ownerships between C and D, even though ownerships between the latter two are much more common. Therefore, the lists are normalized by padding them with zeros.

5. For each NACE code pair, the arithmetic mean of its normalized ownership list is calculated. The resulting values are then min-max normalized to the range $[0, 1]$.

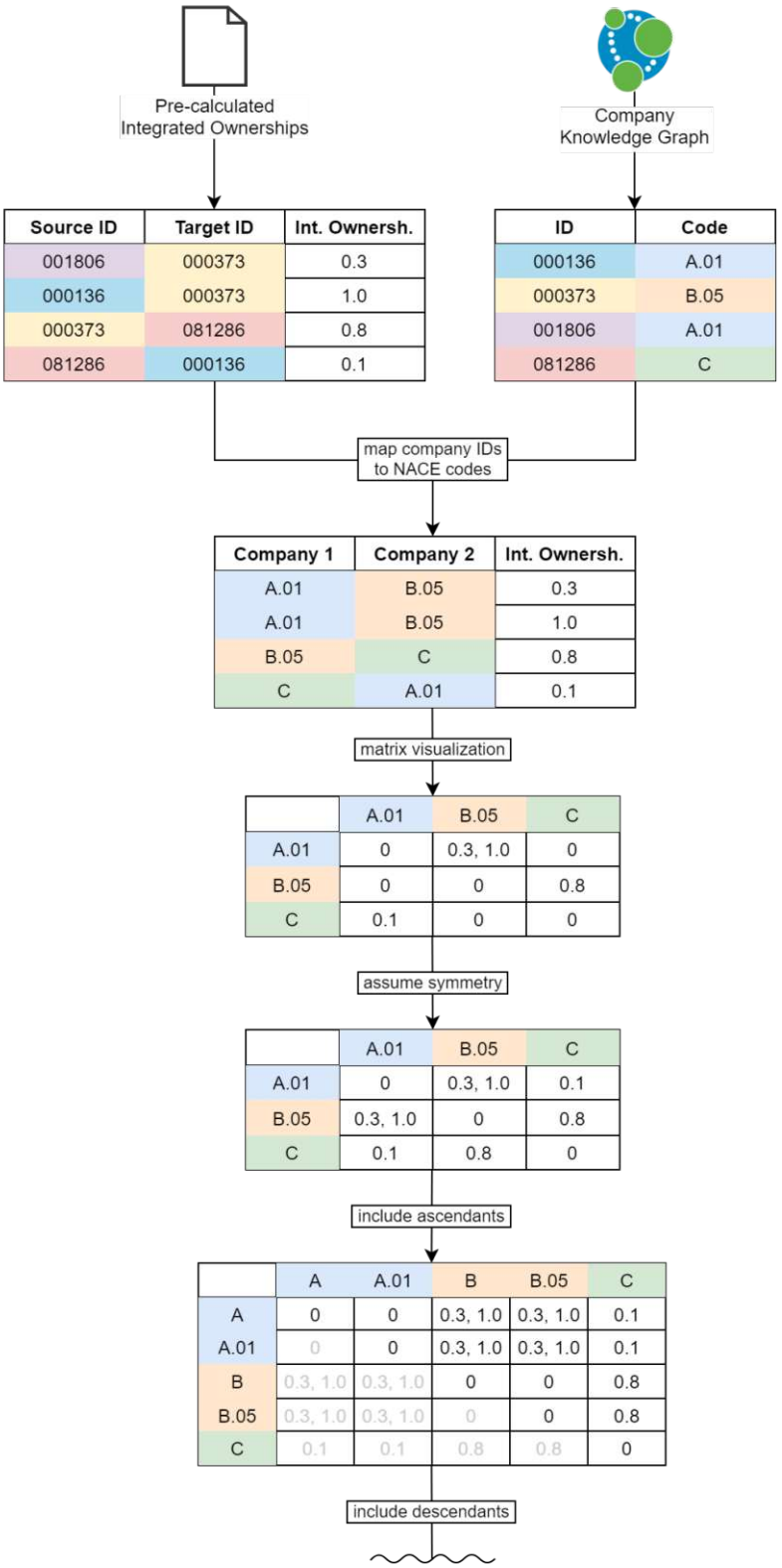
Figure 5.4 shows the example of two small sample datasets of companies and integrated ownerships that are being processed in order to obtain M3 industry similarity values.

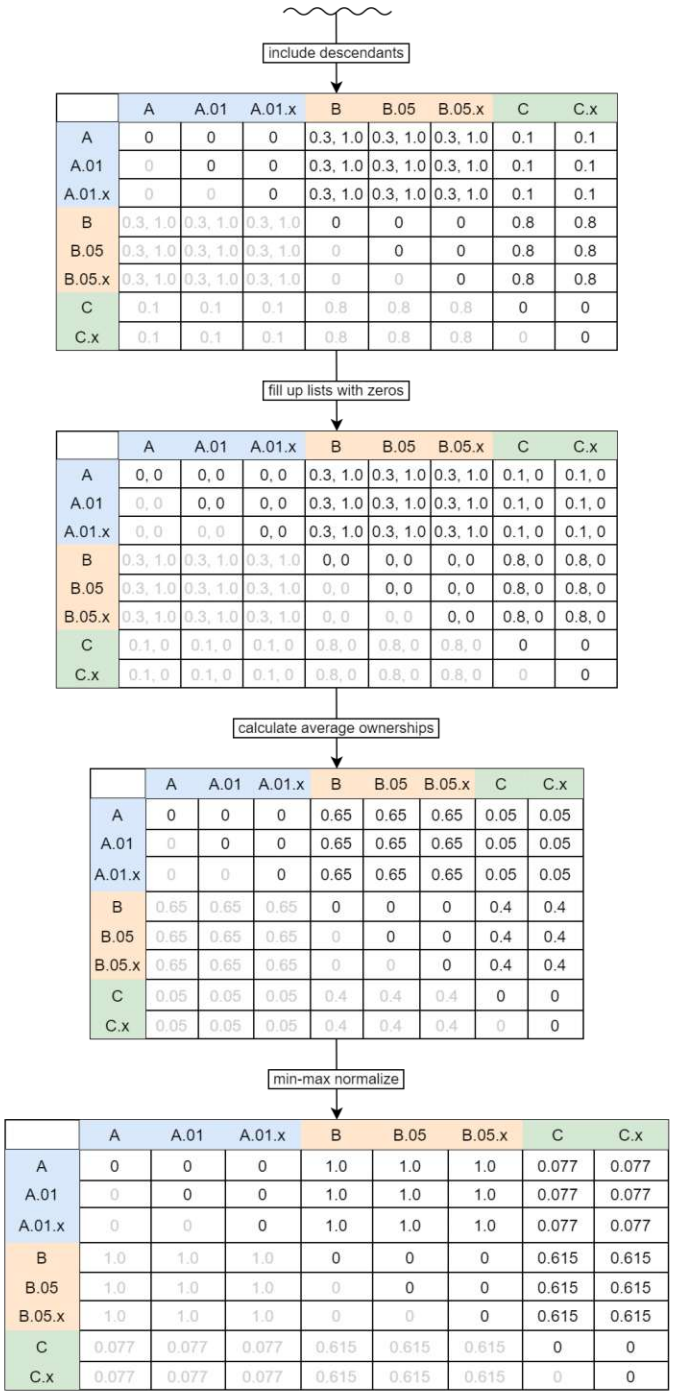
It is important to note that since a pre-calculated integrated ownership dataset was available to be used in our implementation, we did not have to take care of any transitive relations between companies. They had already been made explicit in the input data, which is why it sufficed to consider solely the immediate neighbors of each node in the integrated ownership graph. However, there are applications where only a regular company ownership graph is available or there is no efficient way to compute integrated ownerships. In these cases, we suggest the use of approaches such as node2vec [GL16] as an alternative way to compute the ownership similarities.

Node2vec is an algorithm used to generate node embeddings, which are high-dimensional vector representations that capture the respective node's community structure. It is based on generating random walks within the graph and training a Skipgram model to predict the probability of nodes co-occurring in the same random walk and within a certain window. The resulting node embeddings are located closer to each other in the vector space the more similar their community structures are. Consequently, the similarity of two companies' ownership structures can be quantified by calculating the cosine similarity of their respective vector representations.

The node2vec algorithm is an extension to DeepWalk [PARS14] but offers two hyperparameters for controlling whether the random walks should prefer breadth first search or depth first search. This makes it particularly well suited for sparse graphs such as the Banca d'Italia company ownership graph [DG18]. Also, the algorithm is able to take edge weights into account when generating the random walks, which makes it possible to capture the semantics of different ownership percentages between companies.

However, there are certain drawbacks of using node2vec compared to our pre-calculated integrated ownerships: First, the algorithm is fundamentally non-deterministic due to its dependency on generating walks randomly. This means that the resulting similarity values might differ for each training run, which is detrimental to the their reproducibility. Second, the outcomes are highly dependent on the configuration of the hyperparameters such as the maximum walk length. Due to both of these characteristics it is not guaranteed that relations between all possible nodes in the graph are accurately captured by the node embeddings. For example, if the shortest path between two companies is longer than the maximum random walk length, their relation would be completely ignored by the algorithm and information that could eventually contribute to the ownership similarities is lost. Therefore, we retracted from using node2vec for the purposes of this thesis.





	A	A.01	A.01.x	B	B.05	B.05.x	C	C.x
A	0	0	0	0.3, 1.0	0.3, 1.0	0.3, 1.0	0.1	0.1
A.01	0	0	0	0.3, 1.0	0.3, 1.0	0.3, 1.0	0.1	0.1
A.01.x	0	0	0	0.3, 1.0	0.3, 1.0	0.3, 1.0	0.1	0.1
B	0.3, 1.0	0.3, 1.0	0.3, 1.0	0	0	0	0.8	0.8
B.05	0.3, 1.0	0.3, 1.0	0.3, 1.0	0	0	0	0.8	0.8
B.05.x	0.3, 1.0	0.3, 1.0	0.3, 1.0	0	0	0	0.8	0.8
C	0.1	0.1	0.1	0.8	0.8	0.8	0	0
C.x	0.1	0.1	0.1	0.8	0.8	0.8	0	0

	A	A.01	A.01.x	B	B.05	B.05.x	C	C.x
A	0, 0	0, 0	0, 0	0.3, 1.0	0.3, 1.0	0.3, 1.0	0.1, 0	0.1, 0
A.01	0, 0	0, 0	0, 0	0.3, 1.0	0.3, 1.0	0.3, 1.0	0.1, 0	0.1, 0
A.01.x	0, 0	0, 0	0, 0	0.3, 1.0	0.3, 1.0	0.3, 1.0	0.1, 0	0.1, 0
B	0.3, 1.0	0.3, 1.0	0.3, 1.0	0, 0	0, 0	0, 0	0.8, 0	0.8, 0
B.05	0.3, 1.0	0.3, 1.0	0.3, 1.0	0, 0	0, 0	0, 0	0.8, 0	0.8, 0
B.05.x	0.3, 1.0	0.3, 1.0	0.3, 1.0	0, 0	0, 0	0, 0	0.8, 0	0.8, 0
C	0.1, 0	0.1, 0	0.1, 0	0.8, 0	0.8, 0	0.8, 0	0	0
C.x	0.1, 0	0.1, 0	0.1, 0	0.8, 0	0.8, 0	0.8, 0	0	0

	A	A.01	A.01.x	B	B.05	B.05.x	C	C.x
A	0	0	0	0.65	0.65	0.65	0.05	0.05
A.01	0	0	0	0.65	0.65	0.65	0.05	0.05
A.01.x	0	0	0	0.65	0.65	0.65	0.05	0.05
B	0.65	0.65	0.65	0	0	0	0.4	0.4
B.05	0.65	0.65	0.65	0	0	0	0.4	0.4
B.05.x	0.65	0.65	0.65	0	0	0	0.4	0.4
C	0.05	0.05	0.05	0.4	0.4	0.4	0	0
C.x	0.05	0.05	0.05	0.4	0.4	0.4	0	0

	A	A.01	A.01.x	B	B.05	B.05.x	C	C.x
A	0	0	0	1.0	1.0	1.0	0.077	0.077
A.01	0	0	0	1.0	1.0	1.0	0.077	0.077
A.01.x	0	0	0	1.0	1.0	1.0	0.077	0.077
B	1.0	1.0	1.0	0	0	0	0.615	0.615
B.05	1.0	1.0	1.0	0	0	0	0.615	0.615
B.05.x	1.0	1.0	1.0	0	0	0	0.615	0.615
C	0.077	0.077	0.077	0.615	0.615	0.615	0	0
C.x	0.077	0.077	0.077	0.615	0.615	0.615	0	0

Figure 5.4: Example data processing steps of the Integrated Ownership similarity

5.1.5 M4 - Supply Chain Interdependency

We recall the intuition of this metric:

> *The more supply chains two industries participate in together, the more similar these industries are.*

The foundation of the M4 similarity metric implementation is a 2012 report[4] on production chains and regions created by the Italian ministry of economic development (*Ministero dello sviluppo economico*). The purpose of the analysis was to gain an overview of the most important supply chains in Italy and how different industry sectors work together. It defines a total of 17 supply chains, such as "Agribusiness", "Construction", and "Packaging".

Each supply chain is accompanied by a list of ATECO codes that operate in them. Some of these codes are more high-level than others, so our implemented pre-processing steps explicitly associate their descendants to the respective supply chains.

Since the report was not intended for automated data processing, some items of the ATECO code lists are given in an unstructured form. For example, the "Metallurgy and Steel" supply chain contains all companies with the code `C.25.9` but specifically excludes `C.25.99`. We resolve special cases like these by manually adjusting the data accordingly during pre-processing.

Based on the pre-processed supply chain data, the calculation of the M4 industry similarity value of two given ATECO codes is implemented as follows:

1. The number of supply chains that both ATECO codes occur in together is counted.
2. The count is divided by the maximum number of supply chain co-occurrences of any two ATECO codes. For the given dataset, the maximum number of supply chain co-occurrences is 2, as there is no code pair that occurs in more than 2 supply chains together. The resulting similarity value is in the range $[0, 1]$.

Our approach essentially corresponds to an application of the Jaccard Index, which we have already described in Section 3.2.3. It is defined as the ratio between the size of the intersection of two sets and the size of their union. With regard to the implementation of our M4 metric, those sets are the supply chains that each of the two industries in question is part of.

As already discussed, ATECO codes up to four levels deep are identical to NACE codes. Therefore, the results of calculating the Supply Chain Interdependency can be readily used for comparing NACE industries as well.

[4] `https://www.indire.it/lucabas/lkmw_file/ITS/Brochure%20Filiere%20-def.pdf` (last accessed 15.03.2022)

Products (CPA) \ Industries (NACE)	**A.01**	**A.02**	**A.03**	**B**	...	**S.95**	**S.96**	**T-U**
A.01	7329.7	0.1	0	0.4		0.5	29.3	0
A.02	0	16.4	1.8	0.6		1.5	1	0
A.03	0	0.1	34.8	0		0.2	0.4	0
B	52	0	0	602.1		6.5	57.7	0
...					...			
S.95	0	0	0	5.8		2.1	1.5	0
S.96	0	0.5	0	0.2		0.1	256.6	0
T-U	0	0	0	0		0	0	0

Table 5.3: An excerpt of the input data for the M5 similarity metric. Each value denotes the amount of money an industry spends on certain products. All amounts are given in millions of euros.

5.1.6 M5 - Economic Contribution

We recall the intuition of this metric:

> *The more two industries contribute to each other economically, the more similar these industries are.*

The foundation of the M5 similarity metric implementation is a 2015 report[5] created by the Italian National Institute of Statistics (*Istat*). It summarizes the amount of money Italian companies spend on goods and services to run their businesses.

Table 5.3 shows an excerpt of the data provided by the report. As can be seen, companies are grouped by their level 1 or 2 NACE codes whereas goods and services are grouped by level 1 or 2 CPA codes. *Classification of Products by Activity (CPA)* is a classification taxonomy used within the European Union to categorize products with common characteristics [EU08a]. Its structure and categories are equivalent to those of the NACE taxonomy. In the context of this thesis, CPA and NACE are used interchangeably.

Based on this input data, we implement the process of calculating the M5 industry similarity values as shown in Figure 5.5. Below, we describe the steps in detail.

[5] `http://www.diss.uniroma1.it/moodle2/mod/folder/view.php?id=7284` (last accessed 15.03.2022)

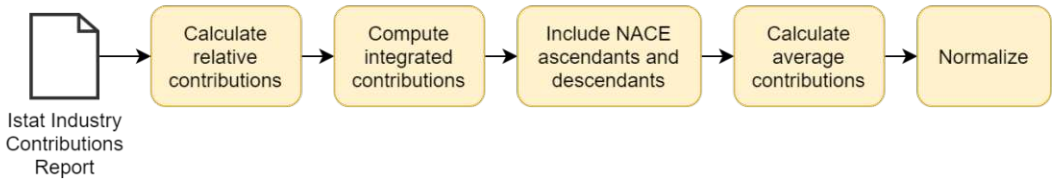


Figure 5.5: Calculation steps of the Economic Contribution similarity

1. Each industry's expenses are divided by the total expense of that industry in order to calculate relative contributions.

2. Similar to the M3 ownership input data, transitive relations are not apparent in the initial data provided by the report. To make the indirect contributions between industries explicit, the *integrated contribution* of all NACE/CPA code pairs is computed as shown in Algorithm 5.2.

   The algorithm takes as input the matrix of economic contributions between all NACE/CPA codes and outputs a matrix of the same size that contains the integrated economic contributions. In lines 3 to 6, the output matrix is initialized by setting all values to empty lists. Lines 8 to 15 define the function `fillIntegratedContributions`(*toOrig*, *to*, *factor*). Its purpose is to recursively gather the contributions of all industries that directly and indirectly contribute to *toOrig* and populate the lists in the integrated contribution matrix respectively. The function iterates over all industries, multiplies the relative contribution of (*from*, *to*) by *factor*, and adds the result to the matrix (lines 9 to 11). If the calculated contribution is below the threshold $\epsilon = 0.00001$, the function terminates (line 12). Else, `fillIntegratedContributions` is invoked again such that it recursively walks the tree of dependencies of *toOrig* using the calculated contribution as the new $factor$ parameter (line 13). Since $factor$ converges towards zero, the function is guaranteed to terminate in a finite number of steps. In lines 17 to 19, `fillIntegratedContributions` is invoked once for each CPA code with $to = toOrig$ and $factor = 1.0$ as initial parameters. Afterwards, the lists in the output matrix are each reduced to their arithmetic mean (lines 21 to 23). The last step is to normalize the integrated contribution values such that the columns sum up to 100% (lines 25 to 27) in order to provide relative contributions percentages.

   At its core, the algorithm is based on the ϵ-*Baldone Ownership* definition by Bellomarini et al. [BBG+20], which we adapted to the domain of economic contributions. We chose $\epsilon = 0.00001$ since tests showed that any further reduction of the threshold did not change the resulting values in a significant way.

3. Since the final metric does not discern between industry and product codes, the direction of the cash flow is irrelevant. Therefore, the integrated contribution

Algorithm 5.2: M5 - Integrated Economic Contribution

Input: *contributions*: Matrix of economic contributions between all NACE/CPA codes

Output: *integratedContributions*: Matrix of integrated economic contributions between all NACE/CPA codes

```
1  allCodes ← contributions.labels
2
3  integratedContributions ← contributions
4  for cell in integratedContributions do
5  |  cell.value ← [ ]
6  end
7
8  Function fillIntegratedContributions(toOrig, to, factor):
9  |  for from in allCodes do
10 |  |  contribution ← contributions[from][to] ∗ factor
11 |  |  append contribution to integratedContributions[from][toOrig]
12 |  |  if contribution > 0.0001 then
13 |  |  |  fillIntegratedContributions(toOrig, from, contribution)
14 |  |  end
15 |  end
16
17 for code in allCodes do
18 |  fillIntegratedContributions(code, code, 1.0)
19 end
20
21 for cell in integratedContributions do
22 |  cell.value ← mean(cell.value)
23 end
24
25 for cell in integratedContributions do
26 |  cell.value ← cell.value/sum(cell.column)
27 end
```

matrix is turned symmetric by replacing each value with the arithmetic mean of $contribution_{from,to}$ and $contribution_{to,from}$.

4. When looking up the contribution between industry `X` and `Y`, what is implicitly asked for is not only the contribution between `X` and `Y` specifically, but also those of their ancestors and descendants in the NACE/CPA hierarchy. To substantiate this claim, consider the following examples:

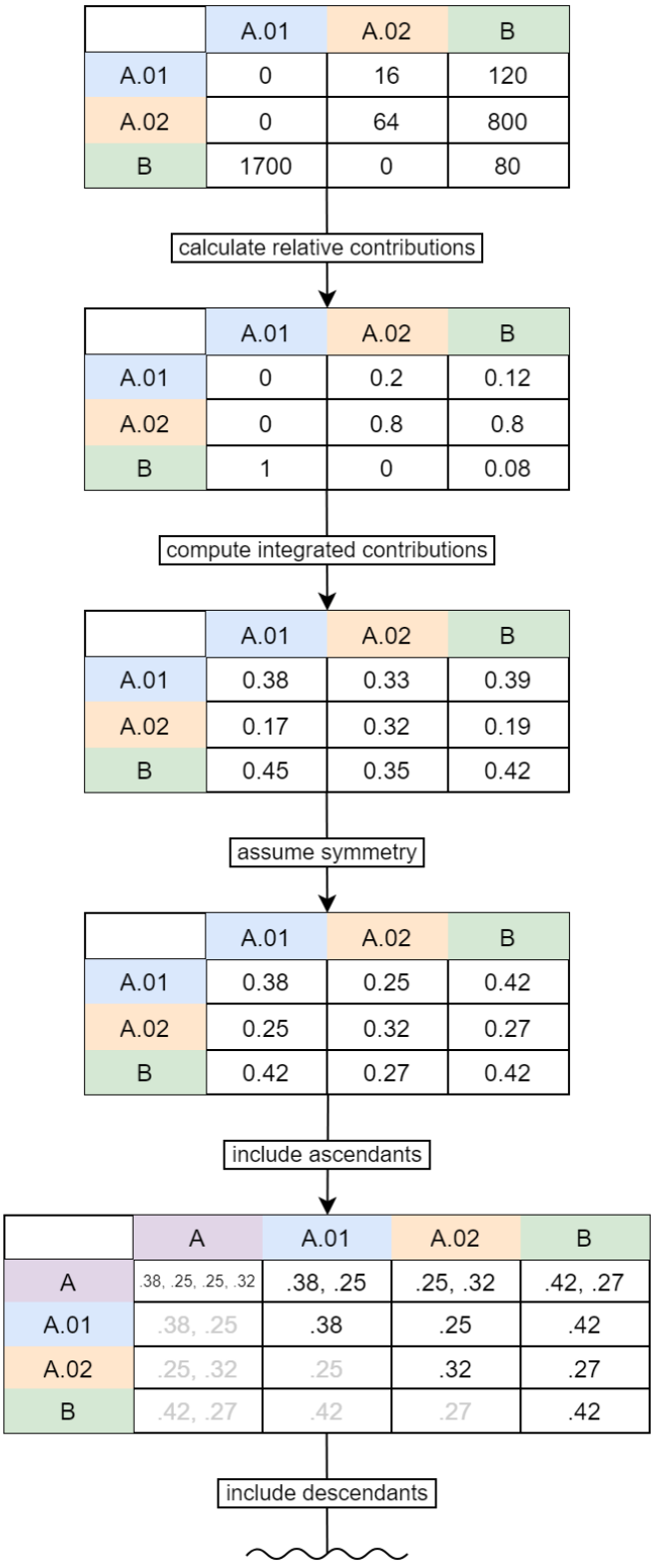
   - $industry_1$ with $code_1$=`A.01` contributes 30% to $industry_2$ with $code_2$=`B.05`. When looking up the contribution of `A` and `B.05`, this relation should be reflected in the result, since `A` encompasses `A.01` by definition. The same applies for the ascendants of `B.05`.
   - $industry_1$ with $code_1$=`A.01` contributes 30% to $industry_2$ with $code_2$=`B.05`. Since it is not apparent from the data which specific sub-industry $industry_1$ belongs to, any descendant of `A.01` (e.g. `A.01.1`, `A.01.11`, `A.01.2`, ...) might implicitly contribute to `B.05`. This uncertainty is resolved by simply adding the contribution relation between $industry_1$ and $industry_2$ to all descendants of `A.01` and `B.05`.

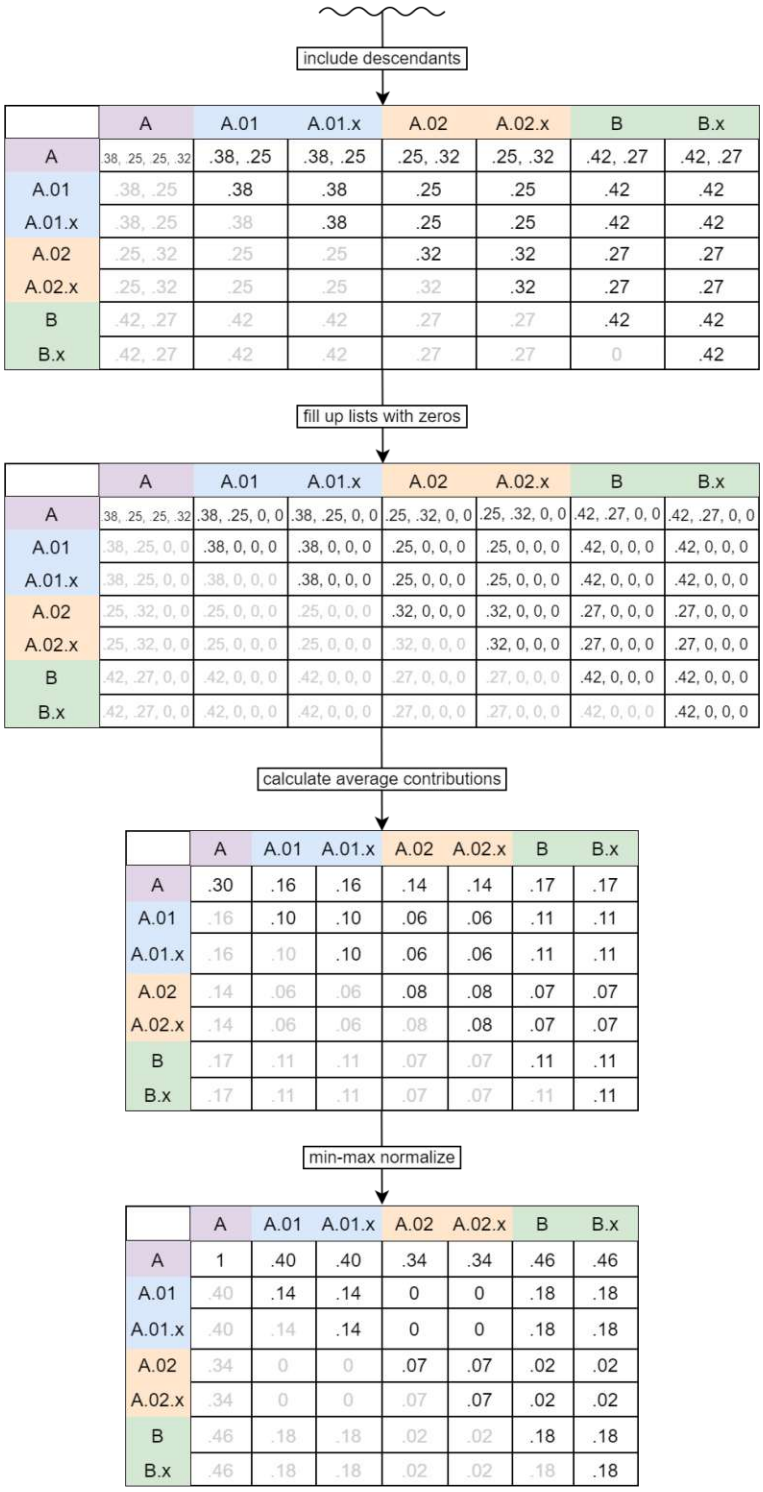
   To implement this, the contribution between each code pair is turned into a list that contains all the contribution percentages of both code's ascendants and descendants.

5. In order to make the contribution lists comparable to each other, they need to be normalized. Otherwise, the resulting similarity values would be skewed based on the length of the lists, which due to the step above is strongly correlated with the level of the respective industry codes. To compensate for this effect, the contribution lists are padded with zeros.

6. For each code pair, the arithmetic mean of its normalized contribution list is calculated. The resulting values are then min-max normalized to the range $[0, 1]$.

Figure 5.6 shows the example of a small sample dataset of contributions that are being processed in order to obtain M5 industry similarity values.





	A	A.01	A.01.x	A.02	A.02.x	B	B.x
A	.38, .25, .25, .32	.38, .25	.38, .25	.25, .32	.25, .32	.42, .27	.42, .27
A.01	.38, .25	.38	.38	.25	.25	.42	.42
A.01.x	.38, .25	.38	.38	.25	.25	.42	.42
A.02	.25, .32	.25	.25	.32	.32	.27	.27
A.02.x	.25, .32	.25	.25	.32	.32	.27	.27
B	.42, .27	.42	.42	.27	.27	.42	.42
B.x	.42, .27	.42	.42	.27	.27	0	.42

	A	A.01	A.01.x	A.02	A.02.x	B	B.x
A	.38, .25, .25, .32	.38, .25, 0, 0	.38, .25, 0, 0	.25, .32, 0, 0	.25, .32, 0, 0	.42, .27, 0, 0	.42, .27, 0, 0
A.01	.38, .25, 0, 0	.38, 0, 0, 0	.38, 0, 0, 0	.25, 0, 0, 0	.25, 0, 0, 0	.42, 0, 0, 0	.42, 0, 0, 0
A.01.x	.38, .25, 0, 0	.38, 0, 0, 0	.38, 0, 0, 0	.25, 0, 0, 0	.25, 0, 0, 0	.42, 0, 0, 0	.42, 0, 0, 0
A.02	.25, .32, 0, 0	.25, 0, 0, 0	.25, 0, 0, 0	.32, 0, 0, 0	.32, 0, 0, 0	.27, 0, 0, 0	.27, 0, 0, 0
A.02.x	.25, .32, 0, 0	.25, 0, 0, 0	.25, 0, 0, 0	.32, 0, 0, 0	.32, 0, 0, 0	.27, 0, 0, 0	.27, 0, 0, 0
B	.42, .27, 0, 0	.42, 0, 0, 0	.42, 0, 0, 0	.27, 0, 0, 0	.27, 0, 0, 0	.42, 0, 0, 0	.42, 0, 0, 0
B.x	.42, .27, 0, 0	.42, 0, 0, 0	.42, 0, 0, 0	.27, 0, 0, 0	.27, 0, 0, 0	.42, 0, 0, 0	.42, 0, 0, 0

	A	A.01	A.01.x	A.02	A.02.x	B	B.x
A	.30	.16	.16	.14	.14	.17	.17
A.01	.16	.10	.10	.06	.06	.11	.11
A.01.x	.16	.10	.10	.06	.06	.11	.11
A.02	.14	.06	.06	.08	.08	.07	.07
A.02.x	.14	.06	.06	.08	.08	.07	.07
B	.17	.11	.11	.07	.07	.11	.11
B.x	.17	.11	.11	.07	.07	.11	.11

	A	A.01	A.01.x	A.02	A.02.x	B	B.x
A	1	.40	.40	.34	.34	.46	.46
A.01	.40	.14	.14	0	0	.18	.18
A.01.x	.40	.14	.14	0	0	.18	.18
A.02	.34	0	0	.07	.07	.02	.02
A.02.x	.34	0	0	.07	.07	.02	.02
B	.46	.18	.18	.02	.02	.18	.18
B.x	.46	.18	.18	.02	.02	.18	.18

Figure 5.6: Example data processing steps of the Economic Contribution similarity

5.2 Takeover Criteria

In this section, we describe the implementations of the takeover prediction criteria proposed in Section 4.2. They are based on the following data:

- A company ownership dataset such as the Banca d'Italia company ownership knowledge graph presented in Section 2.2.1.
- A pre-computed industry similarity metric such as the ones presented in Section 5.1. In our particular case, we use the most preferable of the five metrics according to our evaluation in Section 6.

5.2.1 TC1 - Parent Similarity

Since the *Parent Similarity* criterion takes only the similarity between a parent company and each of its subsidiaries into account, its implementation is simple. It consists of the following steps:

1. For each company s_x controlled or owned by company c, look up the pre-calculated $Similarity(s_x, c)$ between both company's NACE codes. Let $TC1Score(s_x) = Similarity(s_x, c)$.
2. Rank all companies s_x by their *TC1Score*. The companies with the lowest scores are assumed to be the most likely takeover candidates.

5.2.2 TC2 - Group Similarity

The *Group Similarity* implementation computes the average similarity for each subsidiary to all subsidiaries of their mutual parent company.

Unlike TC1, the TC2 criterion needs to take the ownership percentage between c and each s_x into account. The reason for this is illustrated by the following example: Company c owns only a small percentage of s_1 and a large percentage of s_2. Then, the similarity of s_1 to all other s_x should influence their scores less than their similarity to s_2 does. This is done by introducing an *ImplicitShare*, which is calculated as follows:

$$ImplicitShare(s_i) = \begin{cases} ownership\ percentage & \text{if } c \text{ owns a share of } s_i \\ 1.0 & \text{if } c \text{ controls } s_i \end{cases}$$

As already mentioned in Section 2.2.1, the Banca d'Italia company ownership graph discerns between `SHARE` and `CONTROL` relationships. For the purposes of the TC2 implementation, each `CONTROL` relationship is considered equivalent to an implicit 100% `SHARE` relationship.

Based on these foundations, the *Group Similarity* criterion is implemented as follows:

1. For each company s_x controlled or owned by company c, create a list of tuples $[(s_x, s'_1), (s_x, s'_2), ..., (s_x, s'_{|S|})]$, where $s'_{1...|S|}$ are the companies controlled or owned by c (i.e., the "sibling companies" of s_x).

2. For each tuple (s_x, s'_y), get both company's NACE codes and look up the pre-calculated $Similarity(s_x, s'_y)$ between those two codes.

3. For each s'_y, calculate the $ImplicitShare(s'_y)$

4. For each s_x, calculate the weighted arithmetic mean as follows:

$$AverageSiblingSimilarity(s_x) = \frac{\sum_{y=1}^{|S'|} Similarity(s_x, s'_y) * ImplicitShare(s'_y)}{|S|}$$

   The result is referred to as $TC2Score(s_x)$ and reflects how similar company s_x is to all other companies controlled by c.

5. Rank all companies s_x by their $TC2Score$. The companies with the lowest scores are assumed to be the most likely takeover candidates.

CHAPTER 6

Evaluation

In this chapter, we evaluate the industry similarity metrics we have proposed in Chapter 4 and implemented in Chapter 5. The structure of this chapter is based on the three research questions defined in Section 1.3. For each question, we describe the utilized methodology to answer it, conduct the respective steps, and present and interpret the results. We conclude this chapter by discussing the limitations of both our solution and our evaluation methods.

6.1 Statistical Analysis

RQ 1: How can the similarity of industries be quantified?

In this section, we describe the methodology used to answer RQ 1, which consists of a descriptive statistical analysis. Afterwards, we present our results and discuss them.

6.1.1 Methodology

In order to evaluate the plausibility of the implemented industry similarity metrics, we conduct a descriptive data analysis of their output values. For each metric, we perform the following actions:

- **Visualization**: First, we visualize the metric's output values using a histogram. This allows us to gain an overview of their distribution, central tendency, and continuity and also facilitates the interpretation of the key figures described below.
- **Continuity**: We compute the number of distinct output values. A very low number indicates that either the metric's fundamental approach or our concrete implementation cannot accurately model nuances between industries. This degrades

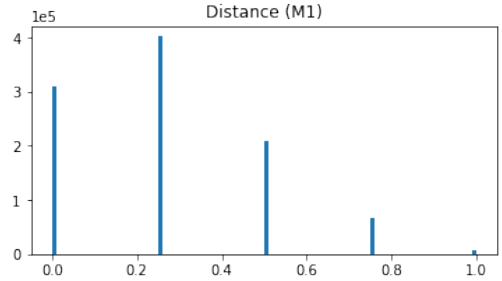
the practical usability of the metric, since applications that are based on comparisons or rankings are not feasible if a large portion of industry pairs yields the exact same similarity score. Therefore, metrics that generate a continuous output value range (i.e., a high number of distinct values) are generally preferable.

- **Floor and ceiling effects**: We compute the five most common output values and how often they appear. Along with the visualizations, this step will reveal if there are any statistical floor or ceiling effects. A significant floor or ceiling effect indicates that the metric's value range is too constrained to accurately model the range of possible industry similarities. As a consequence, a considerable portion of similarities is mapped to the minimum or maximum similarity score. This leads to a potentially inaccurate reflection of the actual variety of the data at the top or bottom of the output scale, which makes it impossible to discern any differences there. Similarly to a low continuity, this degrades the practical usability of the respective metric. Therefore, metrics whose output values do not exhibit floor or ceiling effects are generally preferable.

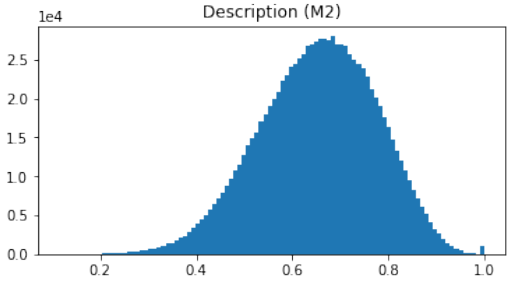
6.1.2 Results

Visualization

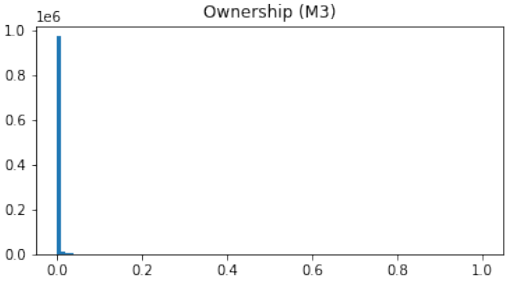
Figure 6.1 shows the histograms of the output values of each industry similarity metric. As can be seen, the less frequent bins of M3, M4, and M5 are barely visible. In order to make them apparent to the viewer, we also provide the histograms with their y-axes set to a logarithmic scale in Figure 6.2.



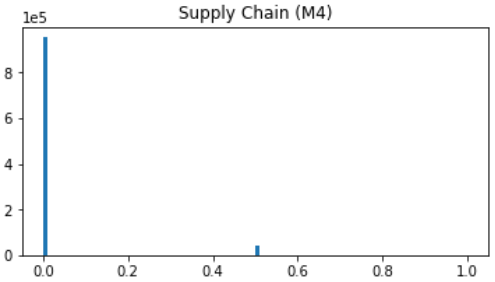
(a) Tree Distance (M1)



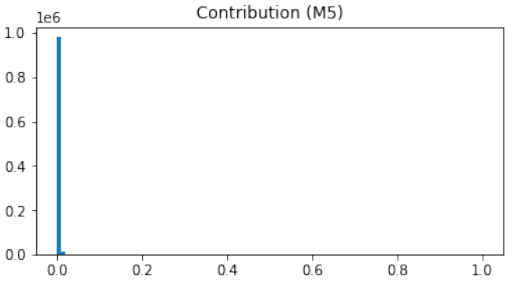
(b) Description Similarity (M2)



(c) Integrated Ownership (M3)

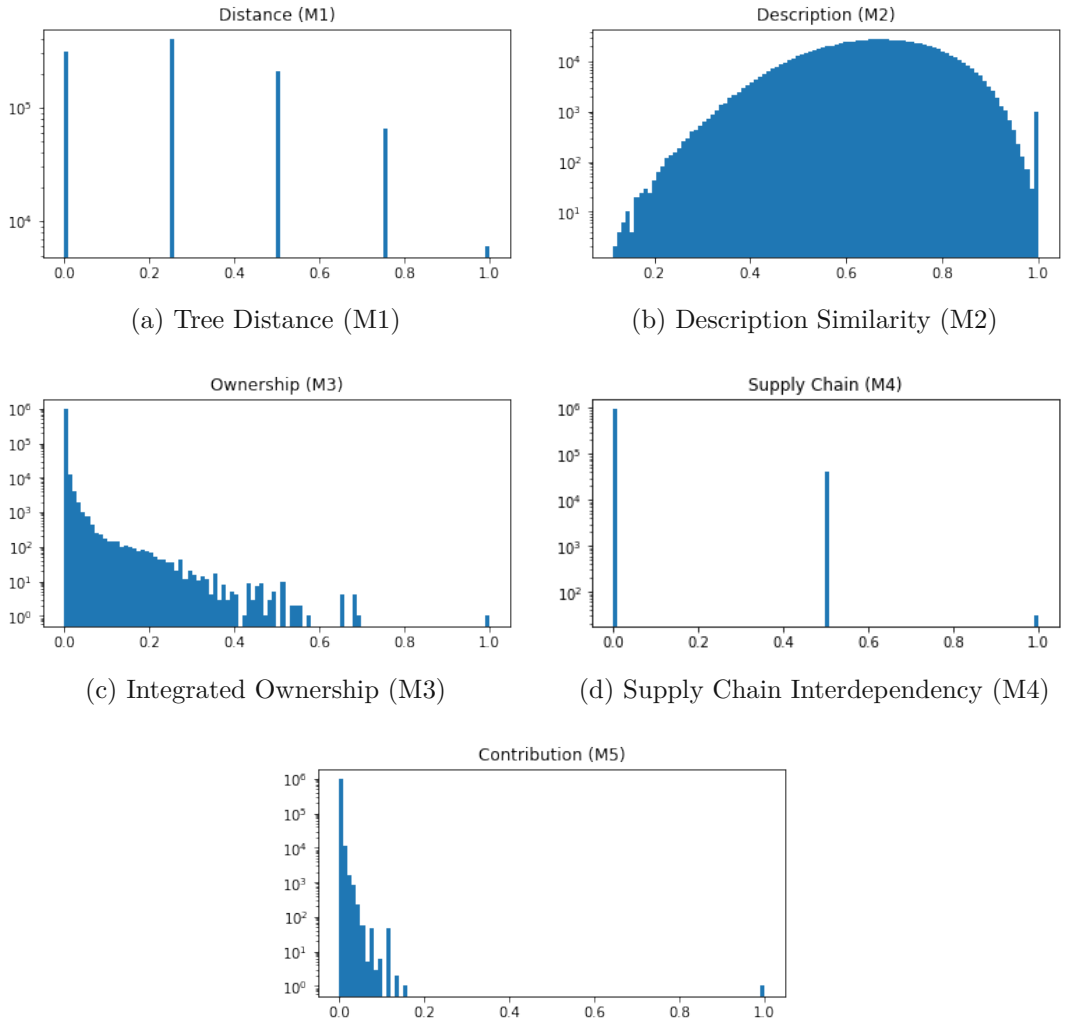


(d) Supply Chain Interdependency (M4)



(e) Economic Contribution (M5)

Figure 6.1: Histograms of the output values of the implemented industry metrics



(a) Tree Distance (M1)

(b) Description Similarity (M2)

(c) Integrated Ownership (M3)

(d) Supply Chain Interdependency (M4)

(e) Economic Contribution (M5)

Figure 6.2: Log-scaled histograms of the output values of the implemented industry metrics

Continuity

Table 6.1 shows the number of distinct values per metric both as an absolute value as well as relative to the maximum possible number of distinct values (MD). The latter of the two variables is computed as follows:

$$RangeUtilization(metric) = \frac{|distinctValues(metric)|}{MD}$$
$$MD = \frac{|Industries|^2}{2}$$

where $|distinctValues(metric)|$ is the number of distinct values of the respective metric and $|Industries|$ is the number of industries of a given industry taxonomy. Note that by our definition, $Similarity(x, y) = Similarity(y, x)$ holds, which means MD is only half the number of industry pairs. For NACE specifically, $MD = 496,008$.

Metric	Distinct values	Range Utilization
Tree Distance (M1)	5	< 0.01 %
Description Similarity (M2)	490,519	98.89 %
Integrated Ownership (M3)	135,975	27.20 %
Supply Chain Interdependency (M4)	3	< 0.01 %
Economic Contribution (M5)	3,007	0.61 %

Table 6.1: Absolute and relative number of distinct values per metric

Floor and Ceiling Effects

Table 6.2 shows the five most common values for each metric and their share of all of its output values.

M1			M2			M3		
Rank	Value	Share	Rank	Value	Share	Rank	Value	Share
1	0.25	40.47%	1	1	0.10%	1	0	48.34%
2	0	31.32%	2	0.7654	$< 0.01\%$	2	$1 * 10^{-7}$	0.02%
3	0.5	21.00%	3	0.5756	$< 0.01\%$	3	$5 * 10^{-6}$	0.01%
4	0.75	6.61%	4	0.6595	$< 0.01\%$	4	0.0008	0.01%
5	1	0.60%	5	0.5585	$< 0.01\%$	5	0.0018	0.01%

M4			M5		
Rank	Value	Share	Rank	Value	Share
1	0	95.93%	1	0	2.82%
2	0.5	4.07%	2	0.000937	0.54%
3	1	$< 0.01\%$	3	0.000941	0.54%
			4	0.001253	0.45%
			5	0.000392	0.45%

Table 6.2: The five most common output values per metric

6.1.3 Discussion

To evaluate the plausibility of the proposed industry similarity metrics, we compared and assessed the statistical properties of their output values.

Continuity

As already described, a high continuity increases the practicality of an industry similarity metric, since it facilitates its use in tasks that are based on comparisons or rankings. We measured the continuity of all proposed metrics by counting their distinct output values. Additionally, we calculated their *range utilization*, which is the same value but relative to the maximum possible number of distinct values. Ideally, a metric has a range utilization of 100%, which would mean that no two industry pairs yield the exact same similarity. In the following paragraphs, we discuss the observed continuity of each metric in detail.

As can be inferred from inspecting the histograms of M1 and M4 (Figure 6.1), they have a very low number of distinct values, since their output values are concentrated exclusively at a few spike-like spots with none in between. Our impression is reinforced by the precise computation of the number of distinct values, which yield only five and three for M1 and M4, respectively. Relatively speaking, both metrics utilize less than 0.01% of their possible range. These low figures can be explained when taking their initial approach and implementation's input data into account.

The number of distinct output values the Tree Distance metric (M1) can produce is directly dependent on how many different path lengths can be found between any two industries within the underlying industry taxonomy. Furthermore, we implemented the path length calculation in a way that compensates for differences between the compared

industries' hierarchy levels. This restricts the number of possible path lengths within the NACE taxonomy to just five (0, 2, 4, 6, and 8), which matches the observed number of distinct output values of our M1 implementation.

For the Supply Chain Interdependency metric (M4), the reason for its low continuity can be found in its input data. Upon analyzing the supply chain report the implementation is based on, we can deduce that there are no two industries which operate in more than two supply chains together. In other words, there are just three different states an industry pair can be in: "No common supply chain", "A single common supply chain", and "Two common supply chains". This explains the low number of just three distinct industry similarity scores that our M4 implementation produces.

Compared to M1 and M4, the Economic Contribution metric (M5) has a higher continuity, but still utilizes only 0.61% of the maximum possible number of distinct values. The reason for this becomes evident upon taking a look at the structure of the input data the M5 implementation is based on. The report discerns only between level-2 NACE codes and gives no detailed information regarding the contributions of industries at the more granular levels of the taxonomy. Moreover, for some industries like "`T` - Activities of households as employers; undifferentiated goods- and services-producing activities of households for own use" or "`U` - Activities of extraterritorial organisations and bodies" there are no recorded cash flows at all, which further reduces the variety of the metric's output values.

With a range utilization of 27.20%, the Integrated Ownership metric (M3) has a significantly higher continuity than the metrics discussed before. Unlike them, the approach of M3 is not based on pre-processed reports or the relatively simple structure of an industry classification scheme. Instead, it uses organic input data, in particular the real-life relationships between the more than three million companies included in the Banca d'Italia company ownership graph. The variety in this data is reflected in the comparatively high variety of output values of the M3 metric.

The Description Similarity metric (M2) achieves an almost full range utilization of 98.89%. Similar to M1, it is based on the existing NACE industry taxonomy. However, it uses the textual descriptions of the industries to compute its output values. These were compiled and formalized by the macro economy researchers of the European Statistical Office (Eurostat) and therefore encapsulate a lot of valuable qualitative information. Since no two text descriptions are the same, the variety of output values of the M2 metric is correspondingly high.

To summarize, the M2 metric is clearly the preferable industry similarity metric from a continuity perspective and when it comes to representing nuances between industries. M3 has a significantly lower number of distinct values but still yields a satisfactory range utilization. M1, M4, and M5, however, are barely usable for tasks that require to compare and rank industry similarities since most industry pairs will yield the exact same similarity score.

Floor and Ceiling Effects

Statistical floor or ceiling effects indicate that a metric is unable to accurately model the whole range of possible industry similarities. In these cases, a considerable portion of similarities is mapped to either 0 or 1, which degrades the practical usability of the respective metric similarly to a low continuity. We measured the existence of floor and ceiling effects for each metric by computing its five most common output values and how often they appear. This reveals whether there are unusually many values at the top or bottom of the output scale. In the following paragraphs, we discuss the observed results in detail.

As can be inferred from inspecting the visualizations of M3, M4, and M5 (Figure 6.1), all of these metrics show very strong floor effects. This is especially noticeable in the linearly scaled histograms, where bins other than the lowest one are barely visible. The remaining bins only become apparent when using logarithmic scaling (Figure 6.2). Our rankings of the metrics' most common output values support this impression.

The Supply Chain Interdependency metric (M4) has by far the most noticeable floor effect. Over 95% of output values are 0, which means that for almost all possible industries, the metric cannot identify any differences in similarity and just considers them maximally dissimilar. This makes the metric virtually unusable for practical applications, since it conveys hardly any meaningful information. The reason for this is its underlying approach in conjunction with the content of the supply chain report that the M4 implementation is based on. Most industries appear only once in the entire report and their similarities to other industries are therefore considered 0. This issue could only be resolved by switching to a more extensive input dataset.

The Integrated Ownership metric (M3) also shows a strong floor effect with 48.34% of its output values being 0. This indicates that the algorithm it is based on is unable to find ownership relationships between companies of about half of all NACE industries. Using a more comprehensive input dataset could potentially resolve this issue, but since the one used for our implementation already covers the largest part of Italy's national economy, this is likely not a generally viable solution.

The Tree Distance metric (M1) is a special case regarding floor and ceiling effects. Although almost a third (31.32%) of output values are 0, it is not the metric's most common value overall, since 0.25 occurs even more often (40.47%). When considering our assessment of the continuity of M1, it becomes apparent that the extremely high share of zeroes is a consequence of the very low number of distinct output values. Nonetheless, its histogram also reveals that the output values resemble a normal distribution whose left side is cut off, which indicates that the metric is not completely free of floor effects. Since the value distribution of M1 is a direct result of the utilized industry taxonomy, the only way of resolving this issue is by switching to a differently structured classification scheme, which is not feasible for most applications.

With 2.82% of its output values being 0, the floor effect of the Economic Contribution metric (M5) is less pronounced than the ones of the metrics discussed before. Similarly

to M3, the reason for its floor effect is mostly the partial lack of input data. For certain industry pairs, the report, that the M5 implementation is based on, simply does not provide any mutual cash flow data and the similarity between the respective industries is therefore 0.

Unlike the metrics discussed before, the Description Similarity metric (M2) does not show any floor effect at all. Its most common output value is 1, which means it shows a potential ceiling effect instead. However, due to its low share of 0.10%, this does not negatively affect the metric's quality. In fact, it is a direct consequence of $Similarity(X, X) = 1$, which holds for all the proposed metrics and means that the similarity between an industry and itself is always considered to be 1. Because of this, the number of output values equal to 1 will at least be 996 for all NACE-based metrics, which is roughly 0.1% of all output values and matches the share observed in the evaluation. We therefore conclude that the slight ceiling effect of M2 can safely be disregarded.

To summarize, the M2 metric is clearly the preferable industry similarity metric with regard to unwanted statistical floor and ceiling effects. All other metrics show floor effects to a certain degree, with M1, M3, and M4 exhibiting the most noticeable ones. This decreases their practical usability because they are unable to convey any information about large numbers of industry pairs.

6.2 Comparison to Human Judgements

RQ 2: How does the proposed solution compare to the human notion of industry similarity?

In this section, we describe the methodology used to answer RQ 2, which includes the acquisition of a test dataset and its comparison to our implemented metrics. Afterwards, we present our results and discuss them.

6.2.1 Methodology

In order to evaluate how closely the implemented metrics match the human intuition of industry similarity, we compare them to a "gold standard" test dataset. However, as we have described in Section 3.1, there is no universally agreed upon definition of industry similarity and consequently no respective dataset available.

Therefore, the first step in our evaluation is to gather a collection of human judgements that we can later use to assess the quality of the proposed metrics. Each instance of the test dataset is a 4-tuple $(reference, option_1, option_2, judgement)$ whose elements are NACE codes and either $judgement = option_1$ or $judgement = option_2$ holds. The value of *judgement* represents the answer to the following question:

> "Is $option_1$ or $option_2$ more similar to *reference*?"

The set of potential questions consists of 100 industry triples $(reference, option_1, option_2)$ that are randomly selected and aim at being a representative sample of all NACE industries.

The judgements making up the test dataset are collected by asking a group of academics and economics experts to state their intuitive opinion on each question. The participants mostly consist of researchers at the Knowledge Graph Lab[1] and the Banca d'Italia.

After the data acquisition, we aggregate the answers for each question and compute the majority vote. Questions that were not answered unambiguously are removed. We assess this property by using a two-sided binomial test and checking whether the resulting p-value is lower than 0.1. In other words, the probability that the majority vote of a question is a result of users randomly selecting options must be 10% or less in order for the respective question to remain in the test dataset.

To evaluate the proposed metrics, we compare the majority vote of each question to the decision of the respective metric, which is computed as follows:

$$Decision(ref, opt_1, opt_2) = \begin{cases} opt_1 & \text{if } Similarity(ref, opt_1) > Similarity(ref, opt_2) \\ opt_2 & \text{if } Similarity(ref, opt_1) < Similarity(ref, opt_2) \\ null & \text{else} \end{cases}$$

The more often the metric decision matches the majority vote, the higher the *validity* of the respective metric.

Additionally, we compute the *coverage* of each metric, which we define as the number of questions the metric give a definitive answer for (i.e., where $Decision(ref, opt_1, opt_2) \neq null$). Metrics with a high coverage are generally preferable.

The purpose of the evaluation is to gain a good sense of the quality of each metric and discuss which of them has the most favorable combination of validity and coverage.

6.2.2 Data Acquisition Tool

To facilitate the collection of the test dataset, we developed a simple survey tool and made it available for the participants to use. It is implemented as a web application that asks users to answer up to 100 questions, where each one is formulated as follows:

Which industry is more similar to X?

where X may be any NACE industry description. The user is then presented with options in the form of two different NACE industry descriptions that he or she needs to choose

[1] `https://kg.dbai.tuwien.ac.at/` (last accessed 15.03.2022)

from. After selecting an option, the next question is presented until all of them have been answered. Figure 6.3 and Figure 6.4 show screenshots of the tool's introductory page and an exemplary question page.

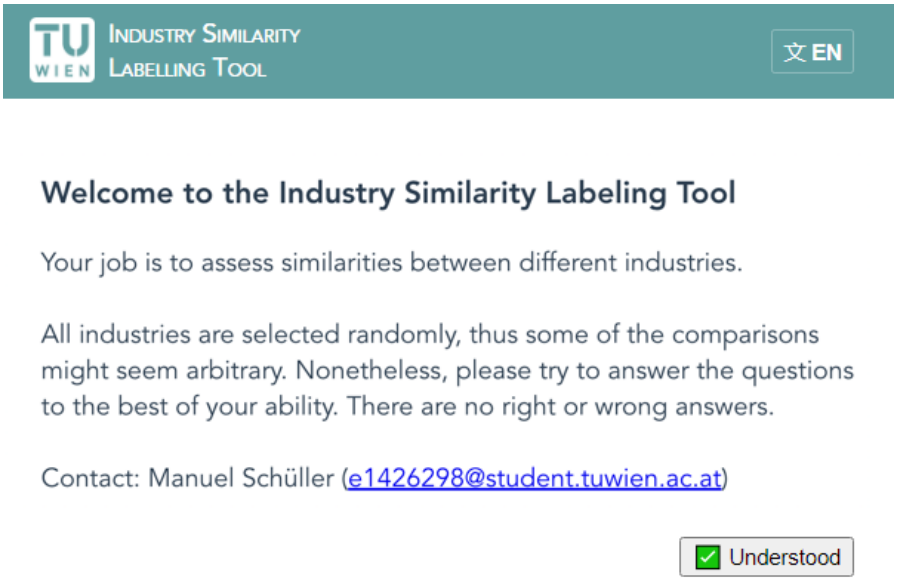


Figure 6.3: The introductory page of the data aquisition tool

As already described, the potential questions are a set of 100 industry triples that are a random sample of all NACE industries. However, descriptions containing "of other" and "n.e.c." are deliberately excluded, as they are mostly not meaningful without further context.

To reduce bias and the effect of background variables, the order of questions and options is randomized for each participant.

In order to minimize comprehension issues caused by language barriers, the application allows the user to display the questions either in German, English, or Italian. For the German and Italian translations, we used the ÖNACE[2] and ATECO[3] industry descriptions, which are the Austrian and Italian equivalents of the NACE standard, respectively.

[2] https://www.data.gv.at/katalog/dataset/stat_onace-2008 (last accessed 15.03.2022)

[3] https://www.istat.it/it/files/2011/03/metenorme09_40classificazione_attivita_economiche_2007.pdf (last accessed 15.03.2022)

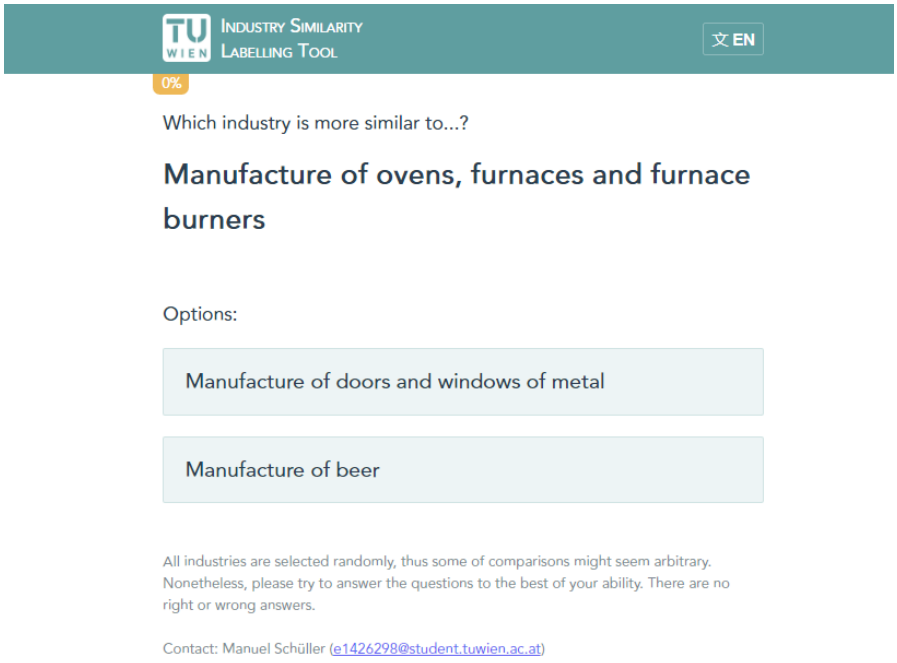


Figure 6.4: An exemplary question the participants are presented with

6.2.3 Acquired Dataset

In total, we collected 1,556 answers during the data acquisition process. Each question received between 12 and 20 answers. The full dataset can be seen in Table A.1.

To obtain a meaningful test dataset, we conducted a two-sided binomial test for each question in order to assess whether there was a strong consent within the participant's answers. Questions with a p-value greater than 0.1 were removed.

The final test dataset consists of 56 questions with a median of 16 answers per questions.

6.2.4 Results

Table 6.3 shows the performance figures of each metric. "Valid" and "Invalid" refer to the number of questions where the majority vote of the human judgements and the decision of the metric do or do not match, respectively. The calculated coverage is stated both as a ratio as well as a percentage. Figure 6.3 visualizes the same results using stacked bar charts. The heights of the green and red bars correspond to the number of valid and invalid decisions of the respective metric, respectively. The height of the whole bar corresponds to its coverage.

Metric	Valid	Invalid	Coverage
Tree Distance (M1)	14 (82%)	3 (18%)	17/56 (30%)
Description Similarity (M2)	47 (84%)	9 (16%)	56/56 (100%)
Integrated Ownership (M3)	24 (75%)	8 (25%)	32/56 (57%)
Supply Chain Interdependency (M4)	4 (100%)	0 (0%)	4/56 (7%)
Economic Contribution (M5)	36 (68%)	17 (32%)	53/56 (95%)

Table 6.3: Results of the comparison of the industry similarity metrics to human judgements

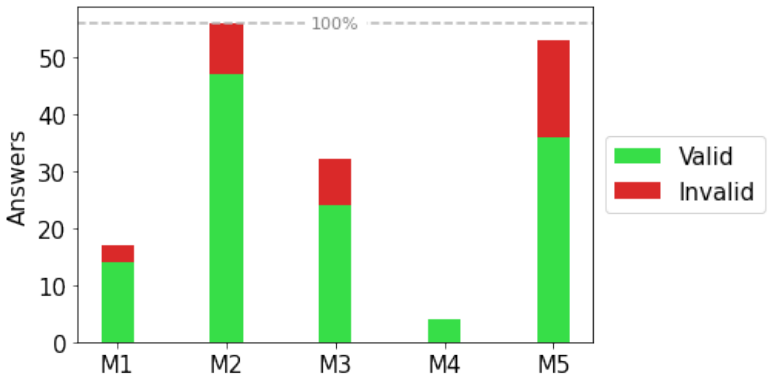


Figure 6.5: Visualized results of the comparison of the industry similarity metrics to human judgements

6.2.5 Discussion

To evaluate the quality of the proposed industry similarity metrics, we tested them against a self-created dataset comprised of human judgements. In particular, we measured two key figures per metric: First, the *validity* denotes the percentage of questions where the metric's decision matches the dataset's majority vote. Second, the *coverage* refers to the percentage of questions the metric can give a definitive answer for. Both values should be as high as possible. In the following paragraphs, we discuss the results of our evaluation.

With a validity of 68%, the Economic Contribution metric (M5) is the least valid one out of all the metrics. This indicates that for the participants of our survey the mutual economic contribution does not seem to be a crucial factor when judging industry similarities. Despite its low continuity of just 3,007 distinct values, the metric achieves a rather high coverage of 95%, which is likely due to the comparatively low floor effect of its output values.

The Integrated Ownership metric (M3) exhibits a validity of 75%. Similarly to M5, the reason for this might be that the participants do not take the ownership structures of real-life companies into account when assessing industry similarity. The coverage of the

M3 metric is 57%, which is likely caused by the combination of its satisfactory continuity and its unfavorable floor effect.

The Tree Distance metric (M1) shows a relatively high validity of 82%. When considering the approach and implementation of M1, this result seems plausible: Since its similarity scores merely reflect the way that NACE classifies and structures industries, this simply means that the NACE standard itself does in fact model the human intuition of industries adequately. The high validity of M1 is countered by its rather poor coverage of just 30%, which is a direct consequence of its low continuity. This reinforces our impression of the limited practical usability of M1, which we have already touched upon in Section 6.1.3.

The Description Similarity metric (M2) achieves excellent results in regards to both validity and coverage. With 84% of questions answered correctly, it has the second highest validity of all metrics. The reason for this is that by employing methods that quantify textual semantics, our M2 implementation is able to utilize all the valuable information that is already contained in the NACE industry descriptions. In addition to its high validity, the metric has a flawless coverage of 100% due to its high continuity and lack of floor or ceiling effects. Overall, M2 meets our expectations of a valid and usable industry similarity metric.

The Supply Chain Interdependency metric (M4) is rather special regarding its validity and coverage. On the one hand, it has a flawless validity of 100%, which is by far the highest one out of all proposed metrics. On the other hand, its coverage is just 7%, which in turn is by far the lowest observed value. The low coverage, which makes the metric virtually unusable in practice, is caused by its exceptionally low continuity and strong floor effect. It is noteworthy though that for the very few question it is able to answer its decisions seem to fully match the human intuition of industry similarity.

Interestingly, the two metrics that are based on empirical economic data (M3 and M5) are also the ones exhibiting the lowest validity. Exploring the reason for this disparity between real-world data and human judgements might be an attractive subject for future work.

In summary, the M2 metric is clearly the preferable metric when it comes to matching the human notion of industry similarity as it shows the most favorable combination of validity and coverage. M1 has a similarly high validity but a significantly worse coverage whereas M5 has similarly high coverage but a worse validity. For M3, neither of the two key figures are exceptionally high or low. Lastly, our impression of the M4 metric being of no practical use has been reinforced due to its poor coverage, regardless of its high apparent validity.

6.3 Takeover Prediction - A Case Study

RQ 3: Which applications can industry similarity metrics be used for in practice?

In this section, we describe the methodology used to answer RQ 3, in particular the conduction of a case study. Afterwards, we present our results and discuss them.

6.3.1 Methodology

In Section 4.2, we have already discussed the concept and implications of hostile foreign company takeovers. In order to demonstrate the practical applicability of industry similarity metrics, we conduct a case study concerning an actual takeover attempt. The purpose of this evaluation is to assess whether the proposed metrics in combination with automated reasoning could have supported authorities in the prediction of that particular takeover.

To do this, we will first describe the case itself, the conglomerate in question, and how its companies are related to each other. Then, we apply the takeover criteria proposed in Section 5.2 on the Banca d'Italia company ownership graph, which contains the data of the involved companies. More specifically, we compute the *TC1 Score* and *TC2 Score* of each subsidiary and rank them accordingly. If a subsidiary is ranked high, its probability to be sold off in the future is also considered high. As we have discussed in Section 6.1.3 and 6.2.5, the Description Similarity metric (M2) is generally the most preferable out of our industry similarity metrics. Therefore, we choose it to be the basis of the TC1 and TC2 implementations used in this evaluation.

Based on our results, we discuss how the company which was actually attempted to be sold was ranked by both TC1 and TC2. The higher its position in the ranking, the more applicable the respective takeover criterion.

Due to compliance reasons, the companies involved in the case study and their specific industry codes had to be anonymized.

6.3.2 Case Description

The Italian company F is a financial service institution and the parent company of a conglomerate containing 40 other Italian companies. One of them is company P, which is a manufacturer of pharmaceutical preparations. At one point in time during the COVID-19 crisis, P received a voluntary tender offer by an international company, which, if accepted, would have lead to P falling under foreign control. According to media reports, F agreed to the offer. However, the Italian authorities exercised their right to prevent the transaction from happening until certain conditions were met since it considered P to be of strategic value to the nation.

Figure 6.6 visualizes the conglomerate as a company ownership graph. Each node represents a company with an anonymized ID as its label and each edge represents either a `CONTROL` or `SHARE` relationship. The edges of the latter of the two are labeled with the percentage of shares the respective company owns. The blue node in the middle represents the conglomerate's parent company F and the green node to its right represents company P. Subsidiaries other than P are colored orange.

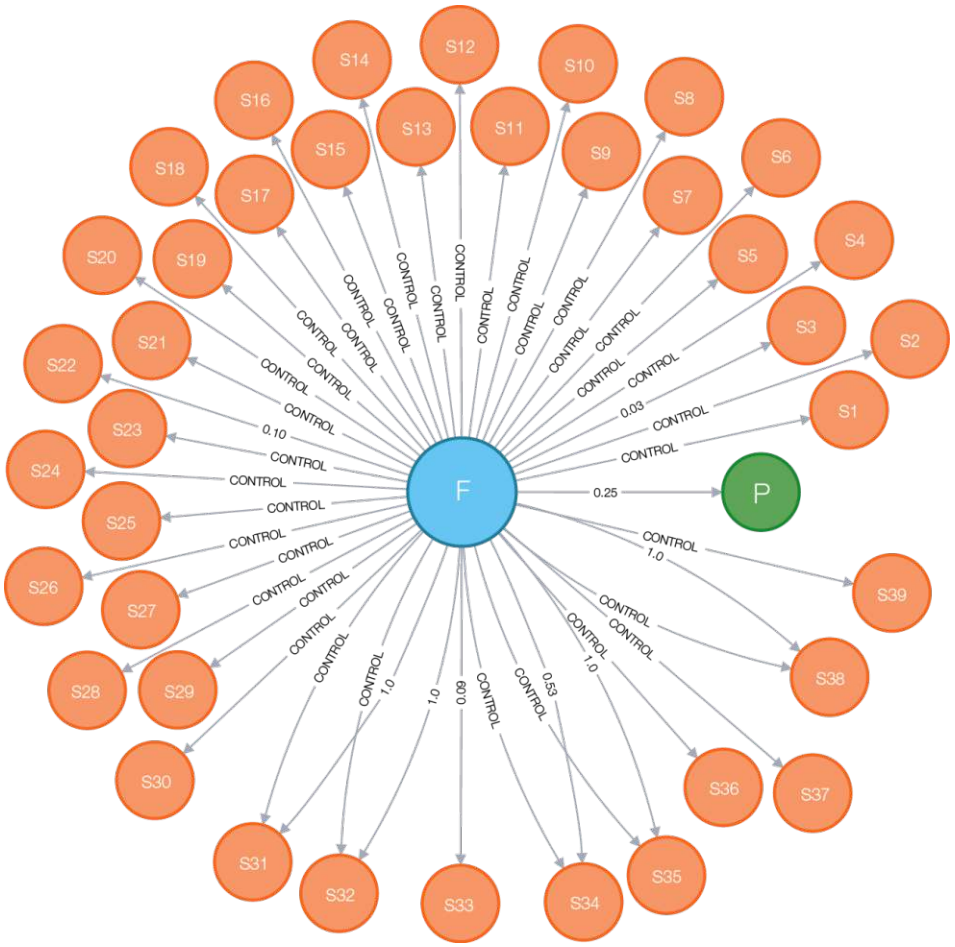


Figure 6.6: The conglomerate that is subject to the case study. The parent company F is centered between its subsidiaries. The company P, which was actually attempted to be sold off, is positioned to its right.

6.3.3 Results

Table 6.4 shows the ten subsidiaries of F with the lowest *TC1 Scores*. Since P is not among these subsidiaries, a dedicated row was added to show its position in the ranking. Note that since there are multiple companies with the same NACE code, they receive the same *TC1 Score* and their order in the table itself is arbitrary. This fact is emphasized by their equal *TC1 Rank*.

	Company	TC1 Rank	TC1 Score
1	S4	1	0.398472
2	S5	1	0.398472
3	S6	1	0.398472
4	S7	1	0.398472
5	S8	1	0.398472
6	S9	1	0.398472
7	S13	2	0.518941
8	S10	3	0.598803
9	S33	4	0.600029
10	S32	4	0.600029
...			
21	**P**	**8**	**0.690814**
...			

Table 6.4: The ten subsidiaries of F with the lowest *TC1 Scores*. The position of company P is added separately.

Table 6.5 shows the ten subsidiaries of F with the lowest *TC2 Scores*. The position of P is emphasized with bold letters. Just like in the TC1 ranking, subsidiaries with the same NACE code receive the same *TC2 Score* and their rank should therefore be considered equal.

6.3.4 Discussion

In order to evaluate the practical applicability of industry similarity metrics, we conducted a case study concerning a hostile foreign company takeover attempt. Our goal was to assess whether the proposed metrics in combination with automated reasoning could have supported the prediction of said takeover. This was done by using the takeover criteria presented in Section 5.2 and ranking the conglomerate's subsidiary according to their *TC1 Scores* and *TC2 Scores*, respectively. In the following section, we discuss the observed results.

As can be seen in the Tables 6.4 and 6.5, both criteria result in largely different rankings. According to TC1, there are 20 other sub companies of F that are considered more at risk of a takeover than P. We can therefore safely assume that an analyst utilizing TC1

	Company	TC2 Rank	TC2 Score
1	**P**	**1**	**0.494436**
2	S37	2	0.516453
3	S13	3	0.522913
4	S38	4	0.548609
5	S4	5	0.559223
6	S5	5	0.559223
7	S6	5	0.559223
8	S7	5	0.559223
9	S8	5	0.559223
10	S9	5	0.559223
...			

Table 6.5: The ten subsidiaries of F with the lowest *TC2 Scores*. The position of company P is emphasized with bold letters.

to automatically predict unwanted company sell-offs would not be able to detect P as being the most vulnerable subsidiary.

According to TC2, P is actually the most likely takeover candidate as it takes the top spot in the respective ranking. Therefore, the results of TC2 would have successfully lead analysts towards considering P as a subsidiary prone to be sold off. Any subsequent actual takeover attempts could then be watched out for deliberately by the respective authorities.

The superiority of TC2 over TC1 in the given case can be explained by inspecting the industry sectors of the parent company F and its subsidiaries. F is an umbrella company providing financial service activities, which is rather unrelated to any specific operational domain. Comparing its industry directly to those of its sub companies, as implemented by TC1, does not yield informative results because F alone simply does not represent the whole conglomerate adequately. On the contrary, TC2 largely disregards the parent company and instead compares the subsidiaries among themselves, which is why it is able to detect outliers within the conglomerate more accurately. The fact that TC2 confidently detected the one company that had actually been attempted to be sold off out of 40 subsidiaries requiring no information apart from their industry classifications is very promising.

In summary, the application of industry similarity metrics in practice has certainly proved its potential. We showed that by utilizing criteria based on our proposed metrics we can support analysts and authorities in automatically detecting companies that are at risk of hostile foreign takeover attempts. Of course, it is difficult to derive generalizable knowledge from a single case study and the question remains whether the proposed criteria yield similarly impressive results for other test cases. Therefore, further research is needed to substantiate our positive impression.

6.4 Limitations

In this section, we discuss the limitations of both our solution and the methodology of our evaluation.

6.4.1 Limitations of the Proposed Solution

Focus on NACE

The proposed solution and implementation focus exclusively on NACE. In order to provide more generalizable findings, it would be necessary to implement and evaluate the industry similarity metrics using different classification systems. Candidates of interest could be the national counterparts of NACE, such as ATECO and ÖNACE, as well as internationally recognized industry classification schemes like *Standard Industrial Classification (SIC)* [US87] and NAICS.

Coverage of Similarity Aspects

In this thesis, we selected five distinct ways for quantifying industry similarity in order to conceptualize our industry similarity metrics. However, as mentioned in Section 3.1, the human perception of similarity is manifold and hard to formalize. We can therefore safely assume that our selection of metrics is non-exhaustive and that there are numerous other and possibly better ways to quantify industry similarity. More research regarding similarity and especially its psychological foundations would certainly benefit and expand our findings.

Selection of Input Data

As we have shown, the quality and applicability of the proposed industry similarity metrics is largely dependent on their input data. This is most notable with our implementation of M4 (Supply Chain Interdependency), whose assessment revealed its high potential but also its poor practical usability caused by its low continuity and strong floor effect. These properties are a direct consequence of the economic report the metric is based on. Using different input data could drastically improve its quality. Similar statements can be made about almost all of the proposed metrics, which is why acquiring and using more extensive input data would most likely benefit the whole solution.

Topicality of Input Data

The content of the reports used as the foundation of M4 and M5 were last updated in 2012 and 2015, respectively. Therefore, any recent changes to the economic landscape are not accurately reflected in the data, which might affect the validity of our results.

Quantification of Text Descriptions

The implementation of the Description Similarity metric (M2) requires a way to quantify the semantics of text. We use word embeddings based on the Word2Vec to achieve this, but our approach leaves room for improvement:

- Since we use a pre-trained word embeddings dataset, some words of the NACE industry descriptions have no corresponding vector representation (they are *out-of-vocabulary*) because they are so rare that they did not occur in the corpus the dataset was trained on. This decrements the quality of the affected description embeddings, since these rare words are potentially very expressive and therefore important for representing the respective industry. Using either a more comprehensive dataset or one created specifically for our purpose could mitigate this problem.

- When mapping a word of an industry description to its respective vector representation, the word's context is not taken into account. For example, the word "rice" always corresponds to the same vector representation. In cases where it is preceded by the word "except" however, the representation should be different to reflect the difference in meaning. This circumstance is currently not considered in our implementation and more sophisticated language models need to be employed to mitigate it.

- In our current solution, we obtain the description embeddings by calculating the mean vector of the description's separate word embeddings. Although this approach is simple and transparent, it is also a rather naive one and alternatives are certainly worth investigating. New and better methods in this field are constantly emerging and some of them, such as ones based on Doc2Vec and BERT, might become feasible for our use case in the future.

6.4.2 Limitations of the Evaluation

Data Acquisition Tool

In order to be able to compare our proposed solution to human judgements, we first had to implement a data acquisition tool that facilitates the collection of said judgements. Although we took measures to reduce potential bias, such as by randomizing the order of questions and options, there are still limitations that have to be considered. The way we measured the participants' opinions is entirely text-based, since the only information we provide are three short industry descriptions. This potentially skews the outcome of the evaluation and might benefit the metrics that are also based on analyzing textual descriptions, such as M2. Finding and employing further evaluation methods is necessary to reveal to which extent this actually affects our results.

Human Judgements Test Dataset

Due to limitations in our resources, the scope of the data acquisition had to be limited as well. Most importantly, the number of potential questions the participants were faced with was rather small and the number of those that made up the final test dataset was even smaller. Eventually, we tested our metrics against only 56 questions, which limits the generalizability of our findings. To improve this aspect, a future test dataset should contain at least as many questions as are necessary to reach a sufficiently large sample size for the respective industry code set. As our post-processing step of the acquired data shows, about half of the initial questions have to be filtered out because the participants' answers do not show a strong consent. The initial set of potential questions should therefore actually contain at least twice as many questions.

Case Study

In order to demonstrate the practical usability of our proposed solution, we conducted a case study. Although our results show the high potential of applying industry similarity metrics, assessing a single case is not sufficient to draw generalizable conclusions from. To confirm our positive impression, it is necessary to execute and evaluate the same tests on further known hostile takeover cases. Additionally, a long-term study in which analysts use our solution for predicting future takeovers could yield valuable insights. Applications other than takeover prediction are certainly worth to be explored as well.

CHAPTER 7

Conclusion and Future Work

In this chapter, we summarize the content and findings of this thesis, express final thoughts, and make suggestions for future work.

7.1 Conclusion

The conceptualization, implementation, and evaluation of ways to quantify the similarity between industry sectors is a challenging process. There is no universal definition of similarity that respective metrics could immediately be derived from and the lack of "gold standard" test datasets makes it difficult to assess their validity. However, there is significant real-life demand for such metrics that so far has not been met. In this thesis, our goal was to close this gap by proposing and evaluating different ways to quantify industry similarity so that existing classifications can be used for artificial intelligence tasks.

As we showed in Chapter 3, previously existing solutions are not satisfactory. Some of them fail to convey nuances between industries while others are non-academic and depend on closed data, which makes them difficult to assess.

The evaluation of our implemented prototype metrics covered their statistical properties, how they compare to the human notion of similarity, and their applicability in a real-life use case. It showed that certain metrics, in particular the Description Similarity metric (M2), meet our expectations of a valid and highly usable industry similarity metric and indeed close the aforementioned gap. Also, the case study revealed the high potential of our solution, as it was able to detect a hostile company takeover through automated reasoning with no information about the involved companies except for their respective industries.

However, our findings also reveal the limitations of our proposed solution and methodology. The quality of our metrics is particularly constrained by their input data, as this is the

most important influence factor to their validity and usability. Also, the significance of our evaluation results is limited by the small size of our test dataset and the focus on a single case study.

7.2 Future Work

As mentioned above, acquiring more extensive and topical input data could drastically enhance the quality of our metric prototypes. The landscape of openly accessible economic data is constantly changing, so a suggestion for future work is to keep adjusting the proposed solution to use the latest and most comprehensive data available.

For the M2 metric (Description Similarity) specifically, there are multiple ways to improve upon our way to quantify the semantics of textual industry descriptions. Although approaches based on Doc2Vec and BERT are not feasible for our use case at the moment, methods of Natural Language Processing and language models in particular are evolving at a rapid pace and applying improvements in these fields could benefit future work.

Our capability of evaluating the validity of industry similarity metrics is still very limited. We tested our metrics against a self-acquired test dataset based on about 1,500 data points, which showed promising results but was ultimately too small of a sample size to make confident statements regarding their significance. Future research should pursue extending our test dataset and also finding more and better ways of assessing whether a respective metric matches the human intuition of industry similarity.

Our case study focuses on a single application of industry similarity metrics, namely the prediction of hostile company takeovers. Although we achieved impressive results for our particular case, the question remains whether our proposed criteria yield similar results for other situations. Also, applications other than company takeovers are certainly worth investigating as well.

List of Figures

List of Tables

List of Algorithms

Bibliography

[ABI+20] Paolo Atzeni, Luigi Bellomarini, Michela Iezzi, Emanuel Sallinger, and Adriano Vlad. Weaving Enterprise Knowledge Graphs: The Case of Company Ownership Graphs. In *Proceedings of the 23rd International Conference on Extending Database Technology, EDBT*, pages 555–566, 2020.

[ARTL19] Ashutosh Adhikari, Achyudh Ram, Raphael Tang, and Jimmy Lin. DocBERT: BERT for Document Classification. *ArXiv*, abs/1904.0, 2019.

[BBC+20] Luigi Bellomarini, Marco Benedetti, Stefano Ceri, Andrea Gentili, Rosario Laurendi, Davide Magnanimi, Markus Nissl, and Emanuel Sallinger. Reasoning on Company Takeovers during the COVID-19 Crisis with Knowledge Graphs. In *RuleML+RR 2020 - 4th International Joint Conference on Rules and Reasoning*, pages 145–156, 2020.

[BBG+20] Luigi Bellomarini, Marco Benedetti, Andrea Gentili, Rosario Laurendi, Davide Magnanimi, Antonio Muci, and Emanuel Sallinger. COVID-19 and Company Knowledge Graphs: Assessing Golden Powers and Economic Impact of Selective Lockdown via AI Reasoning. *CoRR*, abs/2004.1, 2020.

[BBS11] Luca Berchicci, Joern Hendrich Block, and Philipp G. Sandner. The Influence of Geographical Proximity and Industry Similarity in a Business Angel's Investment Choice. *SSRN Electronic Journal*, 2011.

[BFF+20] Julia Bachtrögler, Matthias Firgo, Oliver Fritz, Michael Klien, Peter Mayerhofer, Philipp Piribauer, and Gerhard Streicher. Kurzanalyse zur relativen Betroffenheit der Wiener Wirtschaft von der aktuellen COVID-19-Krise. In *WIFO Studies*, 2020.

[BGK+18] Anuradha Bhamidipaty, Daniel Gruen, Jeffrey O. Kephart, Siva Sankalp Patel, Justin Platz, Danny Soroker, John Vergo, and Alan Webb. Towards a Generalized Similarity Service. *IDA@ KDD*, 20, 2018.

[BGPS17] Luigi Bellomarini, Georg Gottlob, Andreas Pieris, and Emanuel Sallinger. Swift Logic for Big Data and Knowledge Graphs. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pages 2–10, 2017.

[BGPS18] Luigi Bellomarini, Georg Gottlob, Andreas Pieris, and Emanuel Sallinger. The VADALOG System: Swift Logic for Big Data and Enterprise Knowledge Graphs. In *Proceedings of the 12th Alberto Mendelzon International Workshop on Foundations of Data Management*, 2018.

[BLO03] Sanjeev Bhojraj, Charles M. C. Lee, and Derek Oler. What's My Line? A Comparison of Industry Classification Schemes for Capital Market Research. *Journal of Accounting Research*, 41(5):745–774, 2003.

[BM14] Parama Barai and Pitabas Mohanty. Role of industry relatedness in performance of Indian acquirers - Long and short run effects. *Asia Pacific Journal of Management*, 31(4):1045–1073, 2014.

[BR94] Massimo Boninsegna and Matteo Rossi. Similarity measures in computer vision. *Pattern Recognition Letters*, 15(12):1255–1260, 1994.

[BRG05] Mary R. Brooks, Philip J. Rosson, and Horand I. Gassmann. After the M&A: Influences on Corporate Visual Identity Choice. *Corporate Reputation Review*, 8(2):136–144, 2005.

[BSV20a] Luigi Bellomarini, Emanuel Sallinger, and Sahar Vahdati. Knowledge Graphs: The Layered Perspective. In *Knowledge Graphs and Big Data Processing*, pages 20–34. Springer International Publishing, 2020.

[BSV20b] Luigi Bellomarini, Emanuel Sallinger, and Sahar Vahdati. Reasoning in Knowledge Graphs: An Embeddings Spotlight. In *Knowledge Graphs and Big Data Processing*, pages 87–101. Springer International Publishing, 2020.

[CGG+09] Yihua Chen, Eric K. Garcia, Maya R. Gupta, Ali Rahimi, and Luca Cazzanti. Similarity-based Classification: Concepts and Algorithms. *Journal of Machine Learning Research*, 10(27):747–776, 2009.

[Cow17] Sam Cowling. Resemblance. *Philosophy Compass*, 12(4), 2017.

[DCLT18] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 1, pages 4171–4186. Association for Computational Linguistics (ACL), 2018.

[DD09] Michel Marie Deza and Elena Deza. *Encyclopedia of Distances*. 2009.

[DG18] Ayushi Dalmia and Manish Gupta. Towards Interpretation of Node Embeddings. In *Companion of the The Web Conference*, pages 945–952, 2018.

[DHP04] Anna Dubois, Kajsa Hulthén, and Ann Charlott Pedersen. Supply chains and interdependence: A theoretical analysis. *Journal of Purchasing and Supply Management*, 10(1):3–9, 2004.

[EU08a] Eurostat European Union. CPA 2008 - Structure and explanatory notes, 2008.

[EU08b] Eurostat European Union. NACE Rev. 2 - Statistical classification of economic activities in the European Community, 2008.

[EW16] Lisa Ehrlinger and Wolfram Wöß. Towards a Definition of Knowledge Graphs. In *SEMANTiCS*, 2016.

[GL16] Aditya Grover and Jure Leskovec. node2vec: Scalable Feature Learning for Networks. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, page 855–864, 2016.

[Gol94] Robert L. Goldstone. The role of similarity in categorization: Providing a groundwork. *Cognition*, 52(2):125–157, 1994.

[HKP12] Jiawei Han, Micheline Kamber, and Jian Pei. Advanced Cluster Analysis. *Data Mining: Concepts and Techniques*, 3:497–541, 2012.

[HOR87] Gailen L. Hite, James E. Owers, and Ronald C. Rogers. The market for interfirm asset sales: Partial sell-offs and total liquidations. *Journal of Financial Economics*, 18:229–252, 1987.

[HVV+06] Angelos Hliaoutakis, Giannis Varelas, Epimenidis Voutsakis, Euripides G. M. Petrakis, and Evangelos Milios. Information Retrieval by Semantic Similarity. *International Journal on Semantic Web and Information Systems*, 2(3):55–73, 2006.

[KKA+17] Nikolaos Konstantinou, Martin Koehler, Edward Abel, Cristina Civili, Bernd Neumayr, Emanuel Sallinger, Alvaro A. A. Fernandes, Georg Gottlob, John A. Keane, Leonid Libkin, and Norman W. Paton. The VADA Architecture for Cost-Effective Data Wrangling. In *Proceedings of the 2017 ACM International Conference on Management of Data*, page 1599–1602, 2017.

[LM14] Quoc Le and Tomas Mikolov. Distributed Representations of Sentences and Documents. In *Proceedings of the 31st International Conference on International Conference on Machine Learning*, volume 32, page II–1188–II–1196, 2014.

[Lov02] Bradley C. Love. Similarity and Categorization: A Review. *AI Magazine*, 23(2):103–105, 2002.

[Mat09] David Matsumoto. *The Cambridge Dictionary of Psychology*. Cambridge University Press, 2009.

[MCCD13] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space. In *1st International Conference on Learning Representations*, 2013.

[MS03] John D. Martin and Akin Sayrak. Corporate diversification and shareholder value: A survey of recent literature. *Journal of Corporate Finance*, 9(1):37–57, 2003.

[MTB+14] Peter Mika, Tania Tudorache, Abraham Bernstein, Chris Welty, Craig Knoblock, Denny Vrandečić, Paul Groth, Natasha Noy, Krzysztof Janowicz, and Carole Goble. The Semantic Web – ISWC 2014. In *13th International Semantic Web Conference*, volume 8797, 2014.

[Nob57] Clyde E. Noble. Psychology and the Logic of Similarity. *Journal of General Psychology*, 57:23–43, 1957.

[PARS14] Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. DeepWalk: Online Learning of Social Representations. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, page 701–710, 2014.

[PLD05] Christopher Pass, Bryan Lowes, and Leslie Davies. *Collins Dictionary of Economics*. HarperCollins Publishers;Collins, 4 edition, 2005.

[PO16] Ryan L. Phillips and Rita Ormsby. Industry classification schemes: An analysis and review. *Journal of Business & Finance Librarianship*, 21:1–25, 2016.

[PSM14] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. GloVe: Global Vectors for Word Representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014.

[Ris06] Edwina L. Rissland. AI and Similarity. *IEEE Intelligent Systems*, 21(3):39–49, 2006.

[RRT15] Andrea Romei, Salvatore Ruggieri, and Franco Turini. The layered structure of company share networks. In *Proceedings of the 2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pages 1–10, 2015.

[SSN09] Istituto nazionale di statistica Sistema Statistico Nazionale. Classificazione delle attività economiche - Ateco 2007, 2009.

[UN08] Economic & Social Affairs United Nations. International Standard Industrial Classification of All Economic Activities (ISIC),, 2008.

[US87] Office of Management & Budget United States. Standard Industrial Classification Manual, 1987.

[US17] Office of Management & Budget United States. North American Industry Classification System, 2017.

[VAS04] Tom Verguts, Eef Ameel, and Gert Storms. Measures of similarity in models of categorization. *Memory and Cognition*, 32(3):379–389, 2004.

[Wag00] John E. Wagner. Regional Economic Diversity: Action, Concept, or State of Confusion. *Journal of Regional Analysis & Policy*, 30(2):1–22, 2000.

[Wal58] Michael A. Wallach. On psychological similarity. *Psychological Review*, 65(2):103–116, 1958.

[WTLB15] Lyndon White, Roberto Togneri, Wei Liu, and Mohammed Bennamoun. How Well Sentence Embeddings Capture Meaning. In *Proceedings of the 20th Australasian Document Computing Symposium*, pages 1–8, 2015.

[Wun92] Bruce D. Wundt. Reevaluating Alternative Measures Of Industrial Diversity As Indicators Of Regional Cyclical Variations. *The Review of Regional Studies*, 22(1):59–73, 1992.

[WXWZ15] Peng Wang, Bao Wen Xu, Yu Rong Wu, and Xiao Yu Zhou. Link prediction in social networks: The state-of-the-art. *Science China Information Sciences*, 58:1–38, 2015.

[ZWSC19] Ruixuan Zhang, Zhuoyu Wei, Yu Shi, and Yining Chen. BERT-AL: BERT for Arbitrarily Long Document Understanding. 2019.

APPENDIX A

Human Judgements

	Reference	Option$_1$	Option$_2$	Votes$_1$	Votes$_2$	Total votes	p-value	$p < .1$
1	A.01.12	G.46.63	O.84.25	15	2	17	0.00235	True
2	A.01.21	B.07.21	M.70.21	11	1	12	0.00634	True
3	A.01.30	B.06.20	H.49.41	12	4	16	0.07681	True
4	A.03.12	C.28.91	F.43.13	3	14	17	0.01272	True
5	B.05.10	G.46.35	G.46.42	12	4	16	0.07681	True
6	B.07.10	B.08.91	C.27.52	16	1	17	0.00027	True
7	B.08.11	G.45.32	G.46.19	3	12	15	0.03515	True
8	B.08.92	G.47.82	M.70.21	16	1	17	0.00027	True
9	C.10.41	H.52.24	P.85.20	13	3	16	0.02127	True
10	C.10.62	A.02.40	K.66.21	14	1	15	0.00097	True
11	C.10.83	A.01.43	M.69.20	16	2	18	0.00131	True
12	C.11.07	C.28.23	S.95.25	13	2	15	0.00738	True
13	C.18.14	C.11.03	C.30.40	11	3	14	0.05737	True
14	C.20.42	C.10.83	F.43.22	13	0	13	0.00024	True
15	C.23.11	C.22.21	F.43.33	14	4	18	0.03088	True
16	C.24.10	F.41.20	O.84.23	14	2	16	0.00418	True
17	C.25.21	C.31.01	G.47.62	15	1	16	0.00051	True
18	C.25.50	C.10.84	J.60.10	11	3	14	0.05737	True
19	C.25.92	C.18.20	E.38.21	3	12	15	0.03515	True
20	C.26.12	C.23.32	C.25.50	4	12	16	0.07681	True
21	C.27.11	C.21.10	C.24.31	3	11	14	0.05737	True
22	C.27.20	M.71.12	P.85.51	14	2	16	0.00418	True
23	C.28.21	C.11.05	C.25.12	2	15	17	0.00235	True
24	C.28.96	A.01.11	C.20.30	5	14	19	0.06356	True
25	C.29.31	A.01.26	O.84.22	4	13	17	0.04904	True
26	C.30.40	F.41.20	I.55.20	15	0	15	0.00006	True

Table A.1 continued from previous page

27	C.30.91	C.25.61	F.43.11	14	4	18	0.03088	True
28	C.33.11	C.23.13	G.47.74	14	2	16	0.00418	True
29	C.33.13	C.17.11	I.55.20	12	3	15	0.03515	True
30	C.33.14	C.10.62	C.10.84	12	3	15	0.03515	True
31	D.35.30	C.23.51	F.43.22	2	14	16	0.00418	True
32	G.46.22	P.85.42	Q.88.10	2	15	17	0.00235	True
33	G.46.51	C.17.21	C.32.91	12	1	13	0.00341	True
34	G.46.73	C.27.51	G.46.24	4	12	16	0.07681	True
35	G.47.11	C.33.12	G.46.14	2	13	15	0.00738	True
36	G.47.42	J.58.12	Q.86.10	13	2	15	0.00738	True
37	G.47.72	A.01.47	C.17.22	4	13	17	0.04904	True
38	G.47.73	G.47.26	G.47.79	12	2	14	0.01293	True
39	H.49.41	C.10.42	F.42.13	0	14	14	0.00012	True
40	H.52.10	C.10.13	G.47.22	3	13	16	0.02127	True
41	J.59.13	G.47.51	J.58.14	1	14	15	0.00097	True
42	J.60.10	C.26.40	M.73.11	4	12	16	0.07681	True
43	J.63.91	C.24.20	H.49.42	3	14	17	0.01272	True
44	K.65.11	I.56.21	S.95.24	12	4	16	0.07681	True
45	L.68.32	C.24.52	I.56.10	2	15	17	0.00235	True
46	M.74.20	C.28.15	M.71.11	1	13	14	0.00183	True
47	M.74.30	C.26.52	G.46.72	11	2	13	0.02246	True
48	N.77.40	A.01.30	J.59.12	3	13	16	0.02127	True
49	N.81.21	C.13.30	E.38.21	2	12	14	0.01293	True
50	N.82.91	G.45.32	G.46.33	12	4	16	0.07681	True
51	P.85.10	J.58.11	M.74.10	12	2	14	0.01293	True
52	P.85.31	C.17.22	C.20.20	12	4	16	0.07681	True
53	P.85.41	N.82.30	O.84.12	3	11	14	0.05737	True
54	S.94.91	G.47.76	M.74.20	3	12	15	0.03515	True
55	S.95.23	C.23.41	S.96.03	17	0	17	0.00001	True
56	S.95.24	C.11.06	C.32.30	2	11	13	0.02246	True
57	A.01.44	C.32.13	G.45.11	6	9	15	0.60723	False
58	A.01.46	F.43.12	G.47.52	9	6	15	0.60723	False
59	A.01.47	C.24.53	Q.86.22	6	7	13	1.00000	False
60	A.01.62	C.28.95	M.73.20	12	7	19	0.35928	False
61	C.10.32	B.09.90	G.47.79	5	10	15	0.30175	False
62	C.10.71	A.01.44	H.51.21	9	6	15	0.60723	False
63	C.13.95	C.24.42	G.47.82	4	11	15	0.11846	False
64	C.15.12	C.30.11	N.82.30	9	6	15	0.60723	False
65	C.17.24	H.53.10	P.85.10	10	4	14	0.17956	False
66	C.18.11	C.28.95	S.94.11	10	5	15	0.30175	False
67	C.20.41	C.26.30	J.59.20	11	6	17	0.33230	False
68	C.22.22	A.01.41	N.77.34	9	7	16	0.80361	False

Table A.1 continued from previous page

69	C.23.14	B.07.10	C.24.46	10	5	15	0.30175	False
70	C.24.32	C.15.20	G.46.24	11	9	20	0.82380	False
71	C.25.11	G.46.46	N.81.10	9	7	16	0.80361	False
72	C.26.80	C.24.46	G.45.20	6	9	15	0.60723	False
73	C.27.31	B.08.12	C.27.52	6	12	18	0.23788	False
74	C.28.41	C.20.60	C.29.31	4	9	13	0.26684	False
75	C.28.91	A.01.24	N.80.20	8	7	15	1.00000	False
76	C.29.20	C.25.71	G.46.61	4	10	14	0.17956	False
77	D.35.11	C.24.32	H.50.10	9	7	16	0.80361	False
78	D.35.22	G.46.15	H.50.40	6	9	15	0.60723	False
79	E.38.32	C.10.71	C.14.11	4	10	14	0.17956	False
80	G.46.12	C.28.92	G.46.44	10	5	15	0.30175	False
81	G.46.14	C.20.41	G.47.71	7	8	15	1.00000	False
82	G.46.16	C.28.94	G.46.90	7	6	13	1.00000	False
83	G.46.41	K.65.20	R.93.13	7	9	16	0.80361	False
84	G.46.46	C.13.20	C.23.31	8	10	18	0.81452	False
85	G.47.52	F.43.34	S.95.22	8	9	17	1.00000	False
86	G.47.65	G.46.33	M.73.12	6	11	17	0.33230	False
87	G.47.71	C.26.60	H.50.30	9	6	15	0.60723	False
88	G.47.74	I.55.30	L.68.32	4	11	15	0.11846	False
89	J.63.11	C.10.41	C.10.72	6	9	15	0.60723	False
90	K.64.11	C.10.91	C.13.93	8	8	16	1.00000	False
91	K.65.30	E.37.00	N.77.22	6	11	17	0.33230	False
92	M.70.22	C.10.42	C.28.96	6	12	18	0.23788	False
93	O.84.21	K.65.30	K.66.30	5	11	16	0.21011	False
94	O.84.22	C.24.43	C.25.12	10	7	17	0.62905	False
95	P.85.52	C.24.41	N.77.12	6	9	15	0.60723	False
96	Q.86.23	G.46.34	J.62.02	4	11	15	0.11846	False
97	R.93.12	C.12.00	C.28.12	7	8	15	1.00000	False
98	S.94.12	G.46.41	H.49.31	4	11	15	0.11846	False
99	S.94.92	C.28.22	N.77.11	7	9	16	0.80361	False
100	S.95.11	F.42.21	K.66.12	6	8	14	0.79052	False

Table A.1: Acquired dataset of human judgements regarding industry similarity