Discovering Latent Knowledge about Individuals, Enterprises and Technologies

A thesis submitted in fulfilment of the requirements
for the degree of

Doctor of Philosophy
in
Analytics

by

Xian Gong

Supervisor: Marian-Andrei Rizoiu
Co-supervisor: Adriana-Simona Mihaita
External Supervisor: Paul X. McCarthy

School of Computer Science
Faculty of Engineering and Information Technology
University of Technology Sydney
NSW, Australia

March 2026

CERTIFICATE OF ORIGINAL AUTHORSHIP

I, *Xian Gong*, declare that this thesis is submitted in fulfilment of the requirements for the award of Doctor of Philosophy, in the *School of Computer Science, Faculty of Engineering and Information Technology* at the University of Technology Sydney, Australia. This thesis is wholly my own work unless otherwise referenced or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis. This document has not been submitted for qualifications at any other academic institution. This research was supported by an Australian Government Research Training Program (RTP) Scholarship doi.org/10.82133/C42F-K220.

SIGNATURE: Production Note:
Signature removed prior to publication.

Xian Gong

DATE: 6th March, 2026

PLACE: Sydney, Australia

ACKNOWLEDGMENTS

I would like to express my sincere appreciation to everyone who has supported and encouraged me throughout my PhD journey.

I am deeply thankful to my principal supervisor, Associate Professor Marian-Andrei Rizoiu, for his steadfast guidance, thoughtful feedback, and ongoing support. His mentorship has been instrumental in shaping my academic growth and professional path.

My special thanks go to my external supervisor, Paul X. McCarthy, whose insightful advice and consistent encouragement have significantly shaped my research direction and career development.

I would also like to thank all my collaborators—Richard Holden, Fabian Braesemann, Paolo Boldi, Claire McFarland, Colin Griffith, Margaret L. Kern, Fabian Stephany, John A. Johnson, and Marieth Coetzer—for their contributions to the various projects that shaped this thesis. Their intellectual generosity, constructive feedback, and shared expertise have greatly enriched the quality and scope of my research. Working alongside such talented and committed researchers has been both inspiring and rewarding, and I am deeply appreciative of the opportunities to learn and grow through these collaborations.

I am also grateful to the University of Technology Sydney for providing the institutional backing and technical resources essential for conducting this research.

I extend my appreciation to my colleagues at the Behavioural Data Science Lab–Pio, Lin, Frankie, Jooyoung, Philipp, and –for their collaboration, engaging conversations, and unwavering support.

Lastly, I am profoundly grateful to my family and friends. To my parents and siblings, thank you for your unconditional love and belief in me. To my friends in Sydney and my support network in China, your presence has been a continual source of strength and encouragement. Thank you all.

PUBLICATIONS

RELATED TO THE THESIS

1. McCarthy, P. X., **Gong, Xian**, Braesemann, F., Stephany, F., Rizoiu, M.-A., & Kern, M. L. (2023). The impact of founder personalities on startup success. *Scientific Reports*, *13*(1), 17200

2. **Gong, Xian**, McFarland, C., McCarthy, P. X., Griffith, C., & Rizoiu, M.-A. (2023). Informing innovation management: Linking leading r&d firms and emerging technologies. *Proceedings of the International Society for Professional Innovation Management (ISPIM)*, 1–16

3. **Gong, Xian**, McCarthy, P. X., Rizoiu, M.-A., & Boldi, P. (2024). Harmony in the australian domain space. *Proceedings of the 16th ACM Web Science Conference*, 92–102

4. **Gong, Xian**, McCarthy, P. X., Griffith, C., McFarland, C., & Rizoiu, M.-A. (2025). Cosmos 1.0: A multidimensional map of the emerging technology frontier. *Scientific Data* ***(accepted)***

OTHERS

5. McCarthy, P. X., **Gong, Xian**, Eghbal, S., Falster, D. S., & Rizoiu, M.-A. (2021). Evolution of diversity and dominance of companies in online activity. *Plos one*, *16*(4), e0249993

6. McCarthy, P. X., McFarland, C., **Gong, Xian**, & Griffith, C. (2023). The atlas of australia online 2023

7. McCarthy, P. X., **Gong, Xian**, Coetzer, M., Rizoiu, M.-A., Kern, M. L., Johnson, J. A., Holden, R., & Braesemann, F. (2025). The economics of global personality diversity. *arXiv preprint arXiv:2503.19388*

TABLE OF CONTENTS

LIST OF FIGURES

FIGURE **Page**

LIST OF TABLES

ABSTRACT

Innovation, productivity, and long-term economic resilience depend on the effective interaction of individuals, enterprises, and technologies. Yet, despite their central role in shaping economic and societal progress, we lack systematic, scalable ways to detect and quantify how these domains are connected and how their interactions drive performance. This gap persists because such relationships are complex, multidimensional, and embedded within vast, unstructured, and heterogeneous datasets. To tackle these challenges, this research leverages machine learning algorithms, natural language processing techniques, and large language models to discover and map latent knowledge in these domains. The objectives of this thesis are to develop novel computational methods to (i) quantify the influence of individual-level characteristics, such as personality traits, on organisational outcomes, (ii) reveal enterprise-level online network structures and their economic implications, (iii) construct multidimensional representations of the global technology frontier, and (iv) link emerging technologies with leading enterprises. Chapter 1 outlines the motivation, background, and research questions that frame the work. Chapter 2 examines how founders' personality traits, derived from large-scale linguistic analysis, impact startup success, illustrating how NLP and machine learning can extract latent individual-level attributes associated with entrepreneurial outcomes. This study highlights the role of individual psychological factors in shaping the trajectories of start-ups and their market performance. Chapter 3 introduces the first systematic application of harmonic centrality at the web-domain level, combining structural web-graph metrics with firm-level non-structural data to uncover latent relationships between enterprises' online positions, economic diversity, and performance. This work demonstrates the potential of network science to reveal hidden drivers of enterprise competitiveness in the digital economy. Chapter 4 presents Cosmos 1.0, a multidimensional map of the emerging technology frontier constructed from Wikipedia entity embeddings and temporal dynamics to uncover structural patterns, temporal trends, and thematic clusters. This map provides a valuable resource for policymakers, investors, and researchers to monitor, anticipate, and respond to technological change. Chapter 5 develops an NLP-based methodology that uses Wikipedia-derived entity embeddings to link 100 emerging technologies with the world's leading R&D spending companies. This contribution provides a dynamic and scalable framework for connecting corporate innovation strategies with the evolving technology landscape. Overall, this thesis advances computational social science by developing integrated methods to reveal hidden dynamics among individuals, enterprises, and technologies, informing innovation management, economic policy, and strategic decision-making.

CHAPTER 1

INTRODUCTION

Within the framework of the production function in economics, individuals, enterprises, and technologies constitute fundamental inputs that collectively drive productive capacity and economic outcomes. Complex and interdependent relationships exist between individuals, enterprises, and technologies, shaping innovation ecosystems. Understanding these latent connections is crucial for advancing scientific discovery, informing strategic decision-making, and driving economic growth. This thesis presents a series of studies that integrate machine learning algorithms, natural language processing, large language models, and network science to uncover and map these hidden relationships at scale. The research is structured around four interconnected publications, each addressing a distinct yet complementary dimension of the innovation landscape. The first focuses on the human dimension, examining how founder personality traits influence startup success using large-scale, unobtrusive data. The second explores firm level through harmonic centrality in the Australian web domain space to assess the structural and non-structural features of organisational influence in the digital economy. The third concentrates on the technology level by constructing Cosmos 1.0, a multidimensional map of the emerging technology frontier. The fourth investigates how emerging technologies can be systematically linked to leading R&D firms using Wikipedia entity embeddings. Overall, these studies develop scalable, data-driven approaches to reveal latent knowledge across multiple levels of analysis, offering methodological innovations and actionable insights for research, policy, and practice in technology and innovation management.

1.1 Background and Motivation

The human dimension is critical to innovation performance. Startups, often at the forefront of technological disruption, play a disproportionately large role in job creation, economic renewal, and addressing societal challenges. However, the majority fail within a few years, raising essential questions about the determinants of entrepreneurial success. While external factors such as market conditions, industry trends, and access to capital remain important, a growing body of research suggests that the personality traits and composition of founding teams are also significant predictors of venture outcomes (Freiberg & Matz, 2023; Kerr et al., 2018). Understanding how founder personality influences strategic decision-making, adaptability, and team dynamics can provide valuable insights for investors, policymakers, and entrepreneurial support programs aiming to improve startup survival rates and growth potential.

The structure of the digital economy itself equally offers valuable insights into innovation and competitiveness. The Web, as a large-scale and continuously evolving system, serves as a rich data source for understanding economic diversity and organisational influence. Recent studies reveal a steady decline in global link diversity, driven by the increasing dominance of a small number of firms (McCarthy, 2015). Network science approaches, particularly harmonic centrality (Boldi et al., 2022), can capture both structural and non-structural attributes of firms in the digital domain. Analyses of the Australian web domain space show that harmonic centrality distributions remain relatively stable over time, correlate with firm-level economic features and highlight shifts in competitive positioning (**Gong, Xian** et al., 2024).

Moreover, national and corporate competitiveness increasingly relies on the identification and adoption of emerging technologies. Technologies, ranging from artificial intelligence and advanced materials to robotics and renewable energy systems, can redefine industries and shape long-term economic strategy. Their rapid evolution across diverse knowledge domains makes it challenging for policymakers, investors, and business leaders to track, classify, and prioritise them effectively. Traditional "top-down" expert-driven approaches can struggle to capture the full breadth and dynamism of the technology frontier, creating a demand for systematic, data-rich, and multidimensional mapping methods that can detect emerging trends early and support evidence-based decision-making.

Taking a further step, understanding how emerging technologies are being developed, adopted, and deployed by companies is of increasing interest to stakeholders, including investors, policymakers, customers, suppliers, and researchers. Yet, traditional data sources,

taxonomies, and methodologies struggle to capture the dynamic and nascent nature of emerging technologies, as well as the distinctive capabilities of leading R&D spending firms. Patent analysis, for example, remains an important tool in industries such as pharmaceuticals, but performs less well in rapidly evolving domains like software. Recent advances in natural language processing and word embedding models have demonstrated the ability to extract detailed technical information and latent knowledge from unstructured text (Tshitoyan et al., 2019). Leveraging these advances, this research develops a methodology using Wikipedia2vec (Yamada et al., 2020) to map emerging technologies to the world's top R&D spending companies.

As innovation ecosystems grow increasingly complex and interconnected, the ability to capture and analyse such hidden linkages among individuals, enterprises and technologies offers significant potential to inform technology forecasting, guide investment strategies, shape public policy, and strengthen entrepreneurial ecosystems.

1.2 Research Objectives

This thesis develops and applies scalable, data-driven methodologies that leverage machine learning, natural language processing, large language models, and network science to uncover latent knowledge about individuals, enterprises, and technologies. The objectives are to: 1) investigate the latent relationships between founder personality and startup performance, 2) examine the latent relationships between the digital economy and the fundamental economic indicators at firm-level, 3) create a technology dataset, called Cosmos 1.0, that integrates semantic, temporal, and academic signals from diverse sources, and 4) explore the latent relationships between the world's top R&D spending firms and emerging technologies. The novelty lies in its integration of diverse large-scale datasets, advanced computational techniques, and cross-domain perspectives to address interlinked challenges in innovation mapping, economic network analysis, technology foresight, and entrepreneurship research. In the future, we aim to continuously develop a unified embedding space that integrates individuals, enterprises, and technologies, enabling the systematic exploration of their interrelationships and their connections to real-world economic performance.

1.3 Research Questions

Methodological limitations, including small and non-representative samples, reliance on self-reported personality assessments, and a narrow focus on a limited set of trait dimensions, have often constrained research examining the relationship between founder personality and startup success. These constraints hinder the ability to detect nuanced patterns in how personality influences entrepreneurial decision-making, team dynamics, and venture performance. In particular, prior research examining the relationship between personality and entrepreneurship has largely relied on modest sample sizes, survey-based designs, and aggregate Big Five domain measures, thereby limiting the ability to capture fine-grained personality diversity within founding teams at scale. Meta-analytic evidence demonstrates systematic differences between entrepreneurs and non-entrepreneurs in the Big Five traits (Zhao & Seibert, 2006), while subsequent work has shown that entrepreneurial personality constructs are strongly embedded within these broad domains (Leutner et al., 2014). At the macro level, regional analyses further reveal that aggregated Big Five profiles correlate with regional entrepreneurship rates and activity types(Obschonka et al., 2013). However, these studies primarily focus on domain-level traits and individual or regional comparisons, offering limited insight into how nuanced personality configurations within founding teams interact to influence venture resilience, adaptability, and long-term performance. This research addresses these gaps by applying a large-scale, unobtrusively collected dataset of inferred founder personality profiles (drawn from social media text) to analyse their relationship with startup performance across a broad and diverse set of ventures. Using detailed, facet-level personality data, the study identifies distinct trait clusters among founders and examines their correlation with various venture outcomes, including acquisition, public listing, or continued private operation. The first research question is: **Do distinct personality trait clusters exist among startup founders, do these clusters correlate with different rates of venture success or failure, and which specific traits are most predictive of positive outcomes?**

While harmonic centrality is a theoretically well-grounded distance-based centrality measure for directed graphs (Boldi & Vigna, 2014; Freeman, 1979), most large-scale web-graph analyses have historically emphasised structural properties (e.g., degree distributions and distances) across page/host/domain aggregations rather than systematically using harmonic centrality at the web-domain level (Albert et al., 1999; Broder et al., 2000; Meusel et al., 2015), or have relied primarily on PageRank-style prestige measures rather than distance-based centralities (Brin & Page, 1998; Langville & Meyer, 2006). Consequently,

despite its strong theoretical foundations, empirical applications of harmonic centrality at aggregated web-domain levels remain relatively limited, particularly in relation to their economic interpretation. In the Australian domain space, the extent to which harmonic centrality reflects meaningful patterns of economic structure, diversity, and firm-level performance is not well understood. Existing research on web-based economic indicators often focuses on global or cross-national datasets, overlooking country-specific characteristics and temporal dynamics. This research addresses this gap by systematically analysing harmonic centrality distributions within the Australian (.au) domain network from 2018 to 2023. It investigates the persistence and stability of centrality-based patterns over time, examines their relationships with firm-level economic indicators, and evaluates whether non-structural attributes, such as industry sector or organisational type, can predict characteristics in centrality-derived clusters. By linking web-based network measures with economic performance indicators, the study aims to advance the use of web graph analysis as a scalable, reproducible tool for economic insight and policy assessment. The second research question is: **What do harmonic centrality distributions in the Australian web space look like, how do these patterns relate to firm-level economic features, and can non-structural attributes predict cluster membership in centrality-based classifications?**

Existing frameworks for identifying and classifying emerging technologies are often limited in scope, rely on static or infrequently updated taxonomies, and lack the multidimensional integration needed to reflect the complex, evolving nature of the technology landscape. Such limitations hinder the ability of policymakers, investors, researchers, and industry leaders to monitor technological change, identify strategic opportunities, and anticipate disruption. Traditional expert-driven approaches, while valuable, may overlook nascent technologies with low visibility or those emerging from unexpected domains. This study addresses these challenges through the development of Cosmos 1.0, a data-driven, multidimensional dataset of emerging technologies that integrates semantic, temporal, and academic signals from multiple knowledge sources. By leveraging Wikipedia, scholarly publications, patents, and pageviews, and applying entity embeddings with clustering techniques, the approach aims to capture both established and emerging interconnections across science, technology, and industry. The resulting framework provides a scalable and dynamic tool for systematically identifying, categorising, and mapping emerging technologies in a way that supports evidence-based decision-making. The third research question is: **Can the integration of multiple knowledge sources produce a comprehensive, validated dataset of technologies and their interconnections?**

Despite the growing importance of understanding how emerging technologies are connected to the world's most innovative firms, there remains no robust, scalable, and dynamic approach for systematically mapping and analysing the relationships between top global R&D spending companies and a comprehensive set of emerging technologies. Existing methods, such as patent co-classification or manual expert reviews, often lack coverage across diverse industries, struggle to reflect rapidly evolving technology landscapes, and fail to capture latent, multidimensional relationships. This research addresses these limitations by developing and validating a novel methodology that leverages Wikipedia entity embeddings to quantify the proximity between the top 500 R&D spending firms and an established set of emerging technologies. By embedding both firms and technologies in the same semantic space, the approach enables the detection of meaningful associations, identification of thematic clusters, and recognition of strategic white spaces. Such a mapping framework offers significant potential for informing innovation strategy, policy design, and investment prioritisation. The fourth research question is: **Which emerging technologies are most closely linked to the top 500 R&D-spending firms, which firms are most strongly associated with each technology, and do the resulting firm-technology maps reveal coherent clusters or outliers that align with independent validation sources, such as patent data?**

The four research questions collectively structure the progression of this thesis; however, they do not represent its ultimate objective. The broader aim is to integrate individuals, enterprises, and technologies into a unified embedding space, as outlined in the fourth research question, thereby enabling a systematic exploration of their interrelationships, mutual influences, and connections to real-world economic performance. In this regard, the fourth research question serves as a preliminary demonstration of feasibility, offering an initial step toward realising this integrative vision.

1.4 Thesis Structure

This thesis examines the research questions outlined below in detail.

In Chapter 2, we demonstrate that founder personality traits play a significant role in determining a startup's ultimate success. Using a large-scale global dataset of 21,187 startups, we compare the Big Five personality traits of founders, measured from social media text, with those of the employees, revealing substantial differences. Successful entrepreneurs are more likely to exhibit personality traits such as openness to adventure (a preference for variety and novelty), lower modesty (enjoying being the centre of attention), and higher

activity levels (exuberance and energy). Rather than a single "founder type," we identify six distinct personality profiles among successful founders using a clustering algorithm. Furthermore, our results highlight that larger, personality-diverse founding teams have a higher likelihood of success, underscoring the value of cognitive and behavioural diversity in entrepreneurial outcomes. These findings position personality diversity as a novel and measurable dimension of team composition, with important implications for investors, accelerators, and policymakers seeking to improve venture performance.

In Chapter 3, we present the first systematic investigation of harmonic centrality at the web domain level, offering new insights into the structure and dynamics of the Australian companies. Focusing on the relationship between firm-level economic diversity and the structural properties of the Australian domain space, harmonic centrality is employed as the primary structural metric. Temporal analysis reveals that the distribution of harmonic centrality values remains consistent across multiple years. This distribution can be effectively partitioned into six clusters, with the movement of domains between clusters over time providing valuable perspectives on the evolution of the Australian web and its association with non-structural attributes. At the global scale, a significant correlation is observed between the median harmonic centrality of domains in OECD (The Organisation for Economic Co-operation and Development) countries and the World Justice Project Rule of Law Index. Furthermore, 35 OECD countries share notably similar harmonic centrality distributions, suggesting promising opportunities for uncovering patterns in corporate, regional, and national characteristics.

In Chapter 4, we create the Cosmos 1.0 dataset and a novel methodology for mapping the global landscape of emerging technologies. Drawing from diverse and comprehensive contemporary knowledge sources, the dataset comprises 23,544 technology-adjacent entities (TA23k) organised into a hierarchical structure with eight categories of external indices. Each technology is classified into three meta tech-clusters (TC3) and seven thematic tech-clusters (TC7), represented by 100-dimensional embedding vectors. A manually validated subset of 100 emerging technologies (ET100) is also provided. The dataset is further enriched with custom indices designed to characterise the emerging technology landscape, including the Technology Awareness Index, Generality Index, Deeptech classification, and Age of Tech Index. Extensive metadata from Wikipedia and linked data from third-party sources, such as Crunchbase, Google Books, OpenAlex, and Google Scholar, are integrated to validate the relevance and accuracy of these indices. Additionally, a classifier is trained to distinguish between fully developed technologies and technology-related terms, enhancing the dataset's precision and utility.

In Chapter 5, we integrate novel data sources and analytical tools with an innovative methodology to model the relationships between a defined set of emerging technologies and the world's leading R&D spending companies. The resulting map provides a structured representation of the innovation landscape, capturing the proximity between technologies and firms as inferred from the knowledge embedded in their respective Wikipedia profiles. This proximity analysis enables the identification of the most closely associated company–technology pairs, offering insights into strategic positioning and potential collaboration opportunities. A significant positive correlation with relevant patent data supports the validity of the approach. As a demonstration, a set of Circular Economy emerging technologies is mapped to their closest leading R&D spending companies. This application illustrates the model's potential for broader or more targeted analyses, including thematic technology domains, competitive landscapes within specific industries, and areas of strategic national interest.

Lastly, Chapter 6 synthesises the key findings, discusses their implications, and outlines potential avenues for future research.

AUTHOR DECLARATION

The following chapter contains content from the following publication.

McCarthy, P. X., **Gong, Xian**, Braesemann, F., Stephany, F., Rizoiu, M.-A., & Kern, M. L. (2023). The impact of founder personalities on startup success. *Scientific Reports, 13*(1), 17200

Author Contributions: Xian Gong collected and tabulated the data, created the final artwork, and contributed substantially to the creation of figures, data analysis, and methodological implementation. She also played a leading role in the founder/employee prediction component, working alongside P.X. McCarthy and M.-A. Rizoiu was actively involved in designing the research and undertaking the investigation. F. Braesemann and F. Stephany led the multi-factor analysis and, together with Xian Gong, contributed to figure development. M.L. Kern provided leadership on personality insights, while all authors contributed to research design, data analysis, and investigation. The manuscript was written collaboratively, with contributions from all authors.

Production Note:
Signature removed prior to publication.

Paul X. McCarthy

Production Note:
Signature removed prior to publication.

Xian Gong

Production Note:
Signature removed prior to publication.

Fabian Braesemann

Production Note:
Signature removed prior to publication.

Fabian Stephany

Production Note:
Signature removed prior to publication.

Marian-Andrei Rizoiu

Production Note:
Signature removed prior to publication.

Margaret L. Kern

CHAPTER 2

THE IMPACT OF FOUNDER PERSONALITIES ON STARTUP SUCCESS

Startup companies solve many of today's most challenging problems, such as the decarbonisation of the economy or the development of novel life-saving vaccines. Startups are a vital source of innovation, yet the most innovative are also the least likely to survive. The probability of success of startups has been shown to relate to several firm-level factors such as industry, location and the economy of the day. Still, attention has increasingly considered internal factors relating to the firm's founding team, including their previous experiences and failures, their centrality in a global network of other founders and investors, as well as the team's size. The effects of founders' personalities on the success of new ventures are, however, mainly unknown. Here, we show that founder personality traits are a significant feature of a firm's ultimate success. We draw upon detailed data about the success of a large-scale global sample of startups (n = 21,187). We find that the Big Five personality traits of startup founders across 30 dimensions significantly differ from that of the population at large. Key personality facets that distinguish successful entrepreneurs include a preference for variety, novelty and starting new things (openness to adventure), like being the centre of attention (lower levels of modesty) and being exuberant (higher activity levels). We do not find one 'Founder-type' personality; instead, six different personality types appear. Our results also demonstrate the benefits of larger, personality-diverse teams in startups, which show an increased likelihood of success. The findings emphasise the role of the diversity of personality types as a novel dimension of team diversity that influences performance and success.

2.1 Introduction

The success of startups is vital to economic growth and renewal, with a small number of young, high-growth firms creating a disproportionately large share of all new jobs (Davila et al., 2015; Henrekson & Johansson, 2010). Startups create jobs and drive economic growth, and they are also an essential vehicle for solving some of society's most pressing challenges.

As a poignant example, six centuries ago, the German city of Mainz was abuzz as the birthplace of the world's first moveable-type press created by Johannes Gutenberg. However, in the early part of this century, it faced several economic challenges, including rising unemployment and a significant and growing municipal debt. Then in 2008, two Turkish immigrants formed the company BioNTech in Mainz with another university research colleague. Together they pioneered new mRNA-based technologies. In 2020, BioNTech partnered with US pharmaceutical giant Pfizer to create one of only a handful of vaccines worldwide for Covid-19, saving an estimated six million lives (The Economist, 2022). The economic benefit to Europe and, in particular, the German city where the vaccine was developed has been significant, with windfall tax receipts to the government clearing Mainz's €1.3bn debt and enabling tax rates to be reduced, attracting other businesses to the region as well as inspiring a whole new generation of startups (Oltermann, 2021).

While stories such as the success of BioNTech are often retold and remembered, their success is the exception rather than the rule. The overwhelming majority of startups ultimately fail. One study of 775 startups in Canada that successfully attracted external investment found only 35% were still operating seven years later (Grant et al., 2019).

But what determines the success of these 'lucky few'? When assessing the success factors of startups, especially in the early-stage unproven phase, venture capitalists and other investors offer valuable insights. Three different schools of thought characterise their perspectives: first, *supply-side* or *product investors*: those who prioritise investing in firms they consider to have novel and superior products and services, investing in companies with intellectual property such as patents and trademarks. Secondly, *demand-side* or *market-based investors*: those who prioritise investing in areas of highest market interest, such as in hot areas of technology like quantum computing or recurrent or emerging large-scale social and economic challenges such as the decarbonisation of the economy. Thirdly, *talent investors*: those who prioritise the foundation team above the startup's initial products or what industry or problem it is looking to address.

Investors who adopt the third perspective and prioritise talent often recognise that a good team can overcome many challenges in the lead-up to product-market fit. And while

the initial products of a startup may or may not work, a successful and well-functioning team has the potential to pivot to new markets and new products, even if the initial ones prove untenable. Not surprisingly, an industry 'autopsy' into 101 tech startup failures found 23% were due to not having the right team—the number three cause of failure ahead of running out of cash or not having a product that meets the market need (Insights, 2019).

Accordingly, early entrepreneurship research was focused on the personality of founders, but the focus shifted away in the mid-1980s onwards towards more environmental factors such as venture capital financing (Davila et al., 2003; Fracassi et al., 2016; Hochberg et al., 2007), networks (Nann et al., 2010), location (Guzman & Stern, 2015) and due to a range of issues and challenges identified with the early entrepreneurship personality research (Aldrich & Wiedenmayer, 2019; Gartner, 1988). At the turn of the 21st century, some scholars began exploring ways to combine context and personality and reconcile entrepreneurs' individual traits with features of their environment. In her influential work 'The Sociology of Entrepreneurship', Patricia H. Thornton (Thornton, 1999) discusses two perspectives on entrepreneurship: the supply-side perspective (personality theory) and the demand-side perspective (environmental approach). The supply-side perspective focuses on the individual traits of entrepreneurs. In contrast, the demand-side perspective focuses on the context in which entrepreneurship occurs, with factors such as finance, industry and geography each playing their part. In the past two decades, there has been a revival of interest and research that explores how entrepreneurs' personality relates to the success of their ventures. This new and growing body of research includes several reviews and meta-studies, which show that personality traits play an important role in both career success and entrepreneurship (Eikelboom et al., 2018; Freiberg & Matz, 2023; B. H. Hamilton et al., 2019; Kerr et al., 2018; Salmony & Kanbach, 2022), that there is heterogeneity in definitions and samples used in research on entrepreneurship (Kerr et al., 2018; Salmony & Kanbach, 2022), and that founder personality plays an important role in overall startup outcomes (Freiberg & Matz, 2023; B. H. Hamilton et al., 2019).

To study personality rigorously and comparably across large samples, this research adopts the Big Five framework, the dominant and most extensively validated taxonomy of personality structure in contemporary psychology (Goldberg, 1990; John & Srivastava, 1999; McCrae & John, 1992). Emerging from lexical and factor-analytic traditions, the Big Five model provides a comprehensive representation of broad personality, such as Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism, that have demonstrated substantial predictive validity across occupational, organisational, and leadership contexts (Barrick & Mount, 1991; Bono & Judge, 2004). Its hierarchical architecture further enables the

decomposition of broad domains into narrower facet-level traits, offering a theoretically grounded structure for more fine-grained behavioural analysis (Costa Jr & McCrae, 1995; DeYoung et al., 2007). Empirical evidence suggests that these lower-order facets often provide incremental predictive power beyond domain-level scores, particularly when examining specific behavioural outcomes (Paunonen & Ashton, 2001). Facet-level distinctions are especially valuable in entrepreneurial research (Zhao & Seibert, 2006) as domain-level effects may mask meaningful heterogeneity in the specific traits that shape venture formation and performance (Leutner et al., 2014).

Motivated by the pivotal role of the personality of founders on startup success outlined in these recent contributions, we investigate two main research questions:

1. Which personality features characterise founders?

2. Do their personalities, particularly the diversity of personality types in founder teams, play a role in startup success?

We aim to understand whether certain founder personalities and their combinations relate to startup success, defined as whether their company has been acquired, acquired another company or listed on a public stock exchange. For the quantitative analysis, we draw on a previously published methodology (Kern et al., 2019), which matches people to their 'ideal' jobs based on social media-inferred personality traits.

We find that personality traits matter for startup success. In addition to firm-level factors of location, industry and company age, we show that founders' specific Big Five personality traits, such as adventurousness and openness, are significantly more widespread among successful startups. As we find that companies with multi-founder teams are more likely to succeed, we cluster founders in six different and distinct personality groups to underline the relevance of the complementarity in personality traits among founder teams. Startups with diverse and specific combinations of founder types (e.g., an adventurous 'Leader', a conscientious 'Accomplisher', and an extroverted 'Developer') have significantly higher odds of success.

We organise the rest of this paper as follows. In the Section "Results", we introduce the data used and the methods applied to relate founders' psychological traits with their startups' success. We introduce the natural language processing method to derive individual and team personality characteristics and the clustering technique to identify personality groups. Then, we present the result for multi-variate regression analysis that allows us to relate firm success with external and personality features. Subsequently, the Section Discussion mentions limitations and opportunities for future research in this domain. In the

Section Methods, we describe the data, the variables in use, and the clustering in greater detail. Robustness checks and additional analyses can be found in the Supplementary Information.

The remainder of this paper is organised into three sections: Results, Discussion, and Methods. The Results section outlines the data and success definition, presents the inference of founders' Big Five traits, identifies six founder types via clustering, and reports multifactor analyses linking team-level personality characteristics to startup success. The Discussion interprets the findings and their broader implications, while the Methods section details the data construction, modelling procedures, and robustness analyses. This structure prioritises conceptual clarity and empirical transparency by presenting the substantive findings upfront.

2.2 Results

2.2.1 Data

Our analysis relies on two datasets. We inferred individual-level personality facets from Twitter user profiles using an open-vocabulary (Kern et al., 2019) supervised machine-learning approach operationalised via IBM Watson Personality Insights, as detailed in the Methods section. Here, we restrict our analysis to founders with a Crunchbase profile. Crunchbase is the world's largest directory on startups. It provides information about more than one million companies, primarily focused on funding and investors. A company's public Crunchbase profile can be considered a digital business card of an early-stage venture. As such, the founding teams tend to provide information about themselves, including their educational background or a link to their Twitter account.

IBM Watson Personality Insights was used to infer the personality profiles of founders of early-stage ventures from their publicly available Twitter posts using (see Methods for more details). Then, we correlate this information to data from Crunchbase to determine whether particular combinations of personality traits correspond to the success of early-stage ventures. The final dataset used in the success prediction model contains n = 21,187 startup companies (for more details on the data see the Section Methods and Appendix section A.5).

Revisions of Crunchbase as a data source for investigations on a firm and industry level confirm the platform to be a useful and valuable source of data for startups research, as comparisons with other sources at micro-level, e.g., VentureXpert or PwC, also suggest that

the platform's coverage is very comprehensive, especially for start-ups located in the United States (den Besten, 2017). Moreover, aggregate statistics on funding rounds by country and year are quite similar to those produced with other established sources, going to validate the use of Crunchbase as a reliable source in terms of coverage of funded ventures. For instance, Crunchbase covers about the same number of investment rounds in the analogous sectors as collected by the National Venture Capital Association (Block & Sandner, 2009). However, we acknowledge that the data source might suffer from registration latency (a certain delay between the foundation of the company and its actual registration on Crunchbase) and success bias in company status (the likeliness that failed companies decide to delete their profile from the database).

2.2.2 The definition of startup success

The success of startups is uncertain, dependent on many factors and can be measured in various ways. Due to the likelihood of failure in startups, some large-scale studies have looked at which features predict startup survival rates (Antretter et al., 2018), and others focus on fundraising from external investors at various stages (Dworak, 2022). Success for startups can be measured in multiple ways, such as the amount of external investment attracted, the number of new products shipped or the annual growth in revenue. But sometimes external investments are misguided, revenue growth can be short-lived, and new products may fail to find traction.

Success in a startup is typically staged and can appear in different forms and times. For example, a startup may be seen to be successful when it finds a clear solution to a widely recognised problem, such as developing a successful vaccine. On the other hand, it could be achieving some measure of commercial success, such as rapidly accelerating sales or becoming profitable or at least cash positive. Or it could be reaching an exit for foundation investors via a trade sale, acquisition or listing of its shares for sale on a public stock exchange via an Initial Public Offering (IPO).

For our study, we focused on the startup's extrinsic success rather than the founders' intrinsic success per se, as its more visible, objective and measurable. A frequently considered measure of success is the attraction of external investment by venture capitalists (Hsu, 2006). However, this is not in and of itself a good measure of clear, incontrovertible success, particularly for early-stage ventures. This is because it reflects investors' expectations of a startup's success potential rather than actual business success. Similarly, we considered other measures like revenue growth (Blank et al., 2013), liquidity events (Gompers & Lerner, 2004; Hallen & Eisenhardt, 2012; Kaplan & Lerner, 2010), profitability (Shane &

Venkataraman, 2000) and social impact (Zahra & Wright, 2016), all of which have benefits as they capture incremental success, but each also comes with operational measurement challenges.

Therefore, we apply the success definition initially introduced by Bonaventura et al., 2020, namely that a startup is acquired, acquires another company or has an initial public offering (IPO). We consider any of these major capital liquidation events as a clear threshold signal that the company has matured from an early-stage venture to becoming or is on its way to becoming a mature company with clear and often significant business growth prospects. Together these three major liquidity events capture the primary forms of exit for external investors (an acquisition or trade sale and an IPO). For companies with a longer autonomous growth runway, acquiring another company marks a similar milestone of scale, maturity and capability.

Using multifactor analysis and a binary classification prediction model of startup success, we looked at many variables together and their relative influence on the probability of the success of startups. We looked at seven categories of factors through three lenses of firm-level factors: (1) location, (2) industry, (3) age of the startup; founder-level factors: (4) number of founders, (5) gender of founders, (6) personality characteristics of founders and; lastly team-level factors: (7) founder-team personality combinations. The model performance and relative impacts on the probability of startup success of each of these categories of founders are illustrated in more detail in Appendix section A.6 of Appendix (in particular Extended Data Figure A.19 and Extended Data Figure A.20). In total, we considered over three hundred variables (n = 323) and their relative significant associations with success.

2.2.3 The personality of founders

Besides product-market, industry, and firm-level factors (see Appendix section A.1), research suggests that the personalities of founders play a crucial role in startup success- (Freiberg & Matz, 2023). Therefore, we examine the personality characteristics of individual startup founders and teams of founders in relationship to their firm's success by applying the success definition used by Bonaventura et al., 2020.

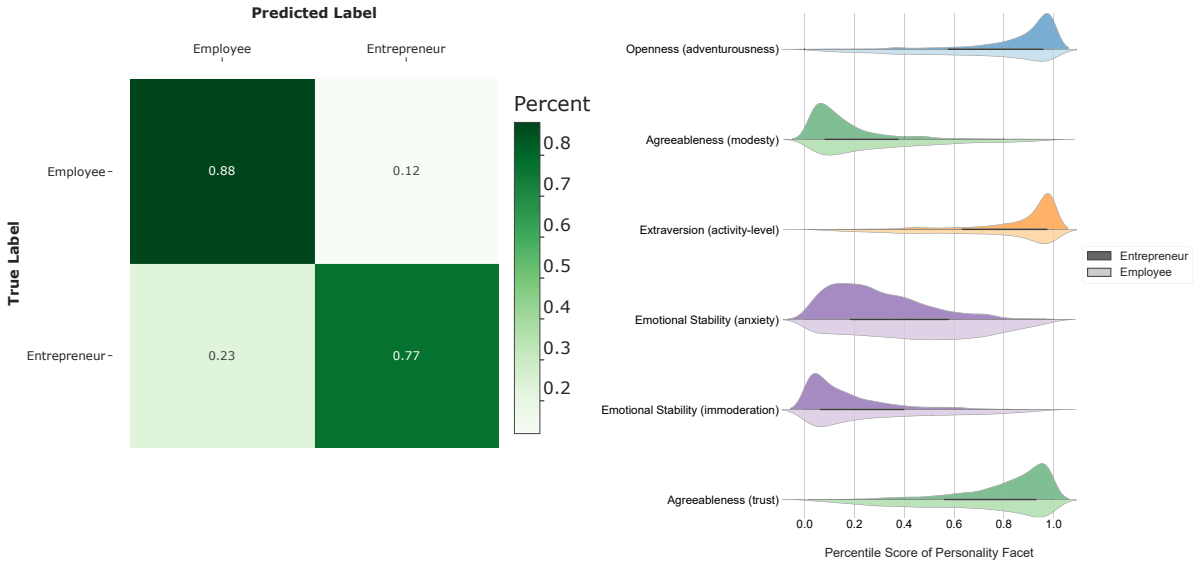
Proved on established computational personality inference approaches validated in prior research (Arnoux et al., 2017; Plank & Hovy, 2015; Schwartz et al., 2013), IBM Watson Personality Insights help us estimate Big Five personality traits across 30 facet-level dimensions for a large global sample of startup founders (see Section Methods for more details).

The startup founders cohort was created from a subset of founders from the global startup industry directory Crunchbase, who are also active on the social media platform Twitter.

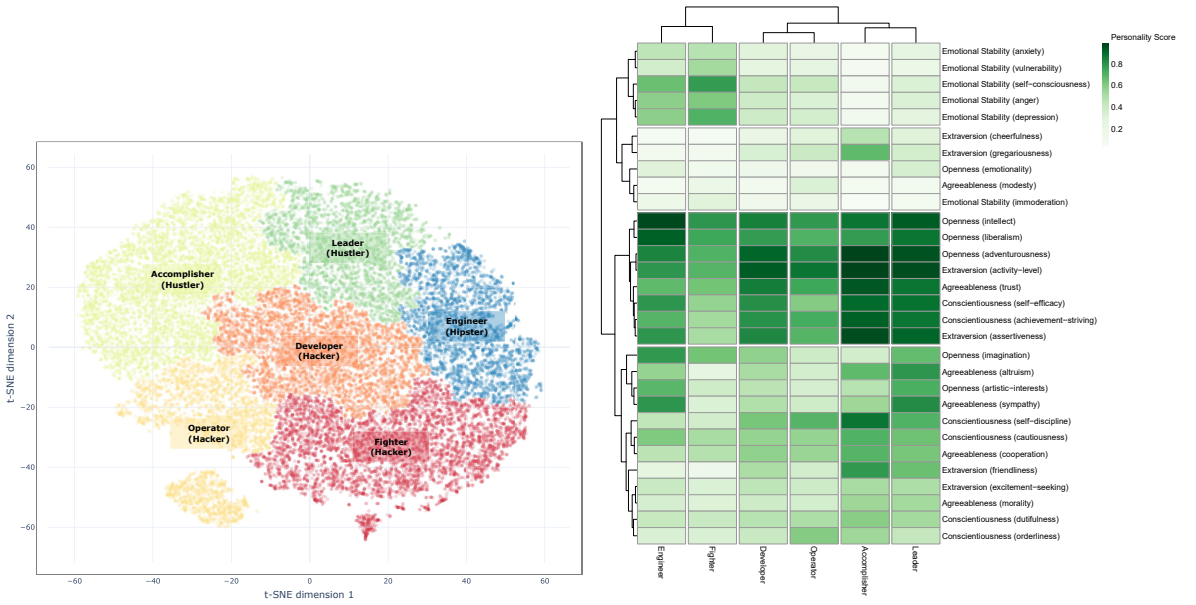
To measure the personality of the founders, we used the Big Five, a popular model of personality which includes five core traits: Openness to Experience, Conscientiousness, Extraversion, Agreeableness, and Emotional stability. Each of these traits can be further broken down into thirty distinct facets. Studies have shown that the Big Five personality traits prospectively predict consequential life outcomes across social, occupational, health, and psychological domains (Ozer & Benet-Martínez, 2006; Roberts et al., 2007), including job performance (Barrick & Mount, 1991), mental and physical health symptoms (Goodwin & Friedman, 2006; Malouff et al., 2005), relationship quality (Watson et al., 2000), and biomarkers of longevity (Friedman & Kern, 2014), often at levels of effect comparable to intelligence and socioeconomic status. Using machine learning to infer personality traits by analysing the use of language and activity on social media has been shown to be more accurate than predictions of coworkers, friends and family and similar in accuracy to the judgement of spouses (Youyou et al., 2015). Further, as other research has shown, we assume that personality traits remain stable in adulthood even through significant life events (Damian et al., 2019; Rantanen et al., 2007; Soldz & Vaillant, 1999). Personality traits have been shown to emerge continuously from those already evident in adolescence (Roberts et al., 2001) and are not significantly influenced by external life events such as becoming divorced or unemployed (Cobb-Clark & Schurer, 2012). This suggests that the direction of any measurable effect goes from founder personalities to startup success and not vice versa.

As a first investigation to what extent personality traits might relate to entrepreneurship, we use the personality characteristics of individuals to predict whether they were an entrepreneur or an employee. We trained and tested a machine-learning random forest classifier to distinguish and classify entrepreneurs from employees and vice-versa using inferred personality vectors alone. As a result, we found we could correctly predict entrepreneurs with 77% accuracy and employees with 88% accuracy (Figure 2.1A). Thus, based on personality information alone, we correctly predict all unseen new samples with 82.5% accuracy (See Appendix section A.2 for more details on this analysis, the classification modelling and prediction accuracy).

We explored in greater detail which personality features are most prominent among entrepreneurs. We found that the subdomain or facet of Adventurousness within the Big Five Domain of Openness was significant and had the largest effect size. The facet of Modesty within the Big Five Domain of Agreeableness and Activity Level within the Big Five Domain of Extraversion was the subsequent most considerable effect (Figure 2.1B). Adventurous-



A **B**



C **D**

Figure 2.1: Founder-Level Factors of Startup Success. (**A**), Successful entrepreneurs differ from successful employees. They can be accurately distinguished using a classifier with personality information alone. (**B**), Successful entrepreneurs have different Big 5 facet distributions, especially on adventurousness, modesty and activity level. (**C**), Founders come in six different types: Fighters, Operators, Accomplishers, Leaders, Engineers and Developers (FOALED). (**D**), Each founder Personality-Type has its distinct facet footprint.

ness in the Big Five framework is defined as the preference for variety, novelty and starting new things—which are consistent with the role of a startup founder whose role, especially in the early life of the company, is to explore things that do not scale easily (Graham, 2013) and is about developing and testing new products, services and business models with the market.

Once we derived and tested the Big Five personality features for each entrepreneur in our data set, we examined whether there is evidence indicating that startup founders naturally cluster according to their personality features using a Hopkins test (see Extended Data Figure A.6). We discovered clear clustering tendencies in the data compared with other renowned reference data sets known to have clusters. Then, once we established the founder data clusters, we used agglomerative hierarchical clustering. This 'bottom-up' clustering technique initially treats each observation as an individual cluster. Then it merges them to create a hierarchy of possible cluster schemes with differing numbers of groups (See Extended Data Figure A.7). And lastly, we identified the optimum number of clusters based on the outcome of four different clustering performance measurements: Davies-Bouldin Index, Silhouette coefficients, Calinski-Harabas Index and Dunn Index (see Extended Data Figure A.8). We find that the optimum number of clusters of startup founders based on their personality features is six (labelled #0 through to #5), as shown in Figure 2.1C.

To better understand the context of different founder types, we positioned each of the six types of founders within an occupation-personality matrix established from previous research (McCarthy et al., 2022). This research showed that 'each job has its own personality' using a substantial sample of employees across various jobs. Utilising the methodology employed in this study, we assigned labels to the cluster names #0 to #5, which correspond to the identified occupation tribes that best describe the personality facets represented by the clusters (see Extended Data Figure A.9 for an overview of these tribes, as identified by McCarthy et al., 2022).

Utilising this approach, we identify three 'purebred' clusters: #0, #2 and #5, whose members are dominated by a single tribe (larger than 60% of all individuals in each cluster are characterised by one tribe). Thus, these clusters represent and share personality attributes of these previously identified occupation-personality tribes (McCarthy et al., 2022), which have the following known distinctive personality attributes (see also Table 2.1):

- *Accomplishers (#0)* — Organised & outgoing. confident, down-to-earth, content, accommodating, mild-tempered & self-assured.

- *Leaders (#2)* — Adventurous, persistent, dispassionate, assertive, self-controlled, calm under pressure, philosophical, excitement-seeking & confident.

- *Fighters (#5)* — Spontaneous and impulsive, tough, sceptical, and uncompromising.

We labelled these clusters with the tribe names, acknowledging that labels are somewhat arbitrary, based on our best interpretation of the data (See Appendix section A.3 for more details).

For the remaining three clusters #1, #3 and #4, we can see they are 'hybrids', meaning that the founders within them come from a mix of different tribes, with no one tribe representing more than 50% of the members of that cluster. However, the tribes with the largest share were noted as #1 Experts/Engineers, #3 Fighters, and #4 Operators.

To label these three hybrid clusters, we examined the closest occupations to the median personality features of each cluster. We selected a name that reflected the common themes of these occupations, namely:

- *Experts/Engineers (#1)* as the closest roles included Materials Engineers and Chemical Engineers. This is consistent with this cluster's personality footprint, which is highest in openness in the facets of imagination and intellect.

- *Developers (#3)* as the closest roles include Application Developers and related technology roles such as Business Systems Analysts and Product Managers.

- *Operators (#4)* as the closest roles include service, maintenance and operations functions, including Bicycle Mechanic, Mechanic and Service Manager. This is also consistent with one of the key personality traits of high conscientiousness in the facet of orderliness and high agreeableness in the facet of humility for founders in this cluster.

Together, these six different types of startup founders (Figure 2.1C) represent a framework we call the FOALED model of founder types—an acronym of Fighters, Operators, Accomplishers, Leaders, Engineers and Developers.

Each founder's personality type has its distinct facet footprint (for more details, see Extended Data Figure A.10 in Appendix section A.3). Also, we observe a central core of correlated features that are high for all types of entrepreneurs, including intellect, adventurousness and activity level (Figure 2.1D).To test the robustness of the clustering of the personality facets, we compare the mean scores of the individual facets per cluster with a 20-fold

Table 2.1: Typology of Founders by Personality. Six different types of founders are revealed by clustering founders (n = 32 k) by their Big Five personality facets. Each type—Fighter, Operator, Accomplisher, Leader, Engineer and Developer (FOALED)—has its distinctive personality footprint.

Founder type		Distinctive Personality Traits
Clustered by personality (Cluster code)		Personality traits of founders in this cluster (Big Five facets)
Fighter	(#5)	Emotional stability (anger, anxiety, depression, immoderation, self-consciousness, vulnerability).
Operator	(#4)	Highest in conscientiousness in the facet of orderliness and high agreeableness in the facet of humility for founders in this cluster.
Accomplisher	(#0)	Highly extraverted (all facets) and conscientious (five facets).
Leader	(#2)	Highest in openness in the facets of artistic interests and emotionality also highest in agreeableness in facets of altruism and sympathy.
Expert/Engineer	(#1)	Highest in openness in the facets of imagination and intellect.
Developer	(#3)	'Middle child' cluster—no facets are maximums or minimums, but it shares characteristics similar to fighters but higher in extraversion.

resampling of the data and find that the clusters are, overall, largely robust against resampling (see Extended Data Figure A.11 in Appendix section A.3 for more details).

We also find that the clusters accord with the distribution of founders' roles in their startups. For example, Accomplishers are often Chief Executive Officers, Chief Financial Officers, or Chief Operating Officers, while Fighters tend to be Chief Technical Officers, Chief Product Officers, or Chief Commercial Officers (see Extended Data Figure A.12 in Appendix section A.4 for more details).

2.2.4 The ensemble theory of success

While founders' individual personality traits, such as Adventurousness or Openness, show to be related to their firms' success, we also hypothesise that the combination, or ensemble, of personality characteristics of a founding team impacts the chances of success. The logic behind this reasoning is complementarity, which is proposed by contemporary research on the functional roles of founder teams. Examples of these clear functional roles have evolved in established industries such as film and television, construction, and advertising (Pratt, 2006). When we subsequently explored the combinations of personality types among founders and their relationship to the probability of startup success, adjusted for a range of other factors in a multi-factorial analysis, we found significantly increased chances of success for mixed foundation teams.

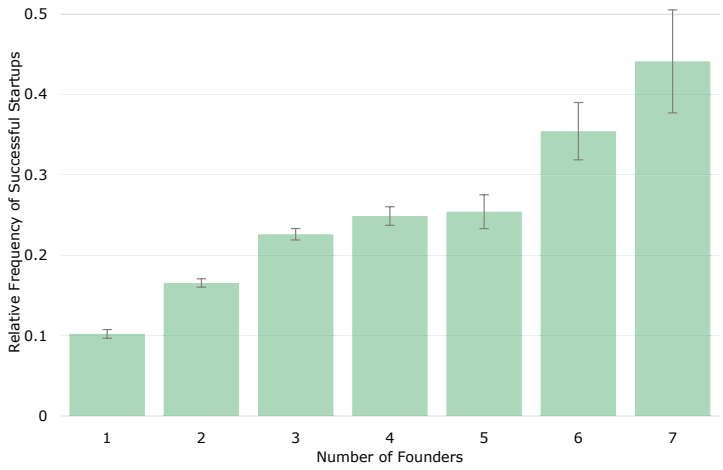
Initially, we find that firms with multiple founders are more likely to succeed, as illustrated in Figure 2.2A, which shows firms with three or more founders are more than twice as likely to succeed than solo-founded startups. This finding is consistent with investors' advice to founders and previous studies (Klotz et al., 2014). We also noted that some personality types of founders increase the probability of success more than others, as shown in Appendix section A.6 (Extended Data Figures A.A.16 and A.A.17). Also, we note that gender differences play out in the distribution of personality facets: successful female founders and successful male founders show facet scores that are more similar to each other than are non-successful female founders to non-successful male founders (see Extended Data Figure A.18).

Access to more extensive networks and capital could explain the benefits of having more founders. Still, as we find here, it also offers a greater diversity of combined personalities, naturally providing a broader range of maximum traits. So, for example, one founder may be more open and adventurous, and another could be highly agreeable and trustworthy, thus, potentially complementing each other's particular strengths associated with startup success.

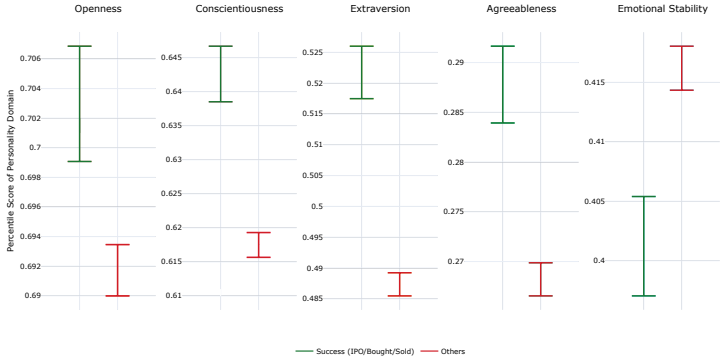
The benefits of larger and more personality-diverse foundation teams can be seen in the apparent differences between successful and unsuccessful firms based on their combined Big Five personality team footprints, as illustrated in Figure 2.2B. Here, maximum values for each Big Five trait of a startup's co-founders are mapped; stratified by successful and non-successful companies. Founder teams of successful startups tend to score higher on Openness, Conscientiousness, Extraversion, and Agreeableness.

When examining the combinations of founders with different personality types, we find that some ensembles of personalities were significantly correlated with greater chances of startup success—while controlling for other variables in the model—as shown in Figure 2.2C (for more details on the modelling, the predictive performance and the coefficient estimates of the final model, see Extended Data Figures A.A.19, A.A.20, and A.A.21 in Appendix section A.6).

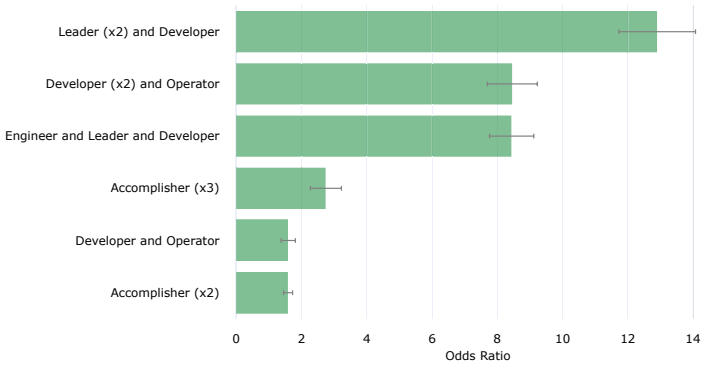
Three combinations of trio-founder companies were more than twice as likely to succeed than other combinations, namely teams with (1) a Leader and two Developers, (2) an Operator and two Developers, and (3) an Expert/Engineer, Leader and Developer. To illustrate the potential mechanisms on how personality traits might influence the success of startups, we provide some examples of well-known, successful startup founders and their characteristic personality traits in Extended Data Figure A.22.



A



B



C

Figure 2.2: The Ensemble Theory of Team-Level Factors of Startup Success. (**A**), Having a larger founder team elevates the chances of success. This can be due to multiple reasons, e.g., a more extensive network or knowledge base but also personality diversity. (**B**), We show that joint personality combinations of founders are significantly related to higher chances of success. This is because it takes more than one founder to cover all beneficial personality traits that "breed" success. (**C**), In our multifactor model, we show that firms with diverse and specific combinations of types of founders, including (Hipster, Hustler, and Hacker) have significantly higher odds of success.

2.3 Discussion

Startups are one of the key mechanisms for brilliant ideas to become solutions to some of the world's most challenging economic and social problems. Examples include the Google search algorithm, disability technology startup Fingerwork's touchscreen technology that became the basis of the Apple iPhone, or the Biontech mRNA technology that powered Pfizer's COVID-19 vaccine.

We have shown that founders' personalities and the combination of personalities in the founding team of a startup have a material and significant impact on its likelihood of success. We have also shown that successful startup founders' personality traits are significantly different from those of successful employees—so much so that a simple predictor can be trained to distinguish between employees and entrepreneurs with more than 80% accuracy using personality trait data alone.

Just as occupation-personality maps derived from data can provide career guidance tools, so too can data on successful entrepreneurs' personality traits help people decide whether becoming a founder may be a good choice for them.

We have learnt through this research that there is not one type of ideal 'entrepreneurial' personality but six different types. Many successful startups have multiple co-founders with a combination of these different personality types.

To a large extent, founding a startup is a team sport; therefore, diversity and complementarity of personalities matter in the foundation team. It has an outsized impact on the company's likelihood of success. While all startups are high risk, the risk becomes lower with more founders, particularly if they have distinct personality traits.

Our work demonstrates the benefits of personality diversity among the founding team of startups. Greater awareness of this novel form of diversity may help create more resilient startups capable of more significant innovation and impact.

The data-driven research approach presented here comes with certain methodological limitations. The principal data sources of this study—Crunchbase and Twitter—are extensive and comprehensive, but there are characterised by some known and likely sample biases.

Crunchbase is the principal public chronicle of venture capital funding. So, there is some likely sample bias toward: (1) Startup companies that are funded externally: self-funded or bootstrapped companies are less likely to be represented in Crunchbase; (2) technology companies, as that is Crunchbase's roots; (3) multi-founder companies; (4) male founders: while the representation of female founders is now double that of the mid-2000s,

women still represent less than 25% of the sample; (5) companies that succeed: companies that fail, especially those that fail early, are likely to be less represented in the data.

Samples were also limited to those founders who are active on Twitter, which adds additional selection biases. For example, Twitter users typically are younger, more educated and have a higher median income (Duggan et al., 2015). Another limitation of our approach is the potentially biased presentation of a person's digital identity on social media, which is the basis for identifying personality traits. For example, recent research suggests that the language and emotional tone used by entrepreneurs in social media can be affected by events such as business failure (Fisch & Block, 2021), which might complicate the personality trait inference.

In addition to sampling biases within the data, there are also significant historical biases in startup culture. For many aspects of the entrepreneurship ecosystem, women, for example, are at a disadvantage (Brush et al., 2019). Male-founded companies have historically dominated most startup ecosystems worldwide, representing the majority of founders and the overwhelming majority of venture capital investors. As a result, startups with women have historically attracted significantly fewer funds (Kanze et al., 2018), in part due to the male bias among venture investors, although this is now changing, albeit slowly (Fan, 2022).

The research presented here provides quantitative evidence for the relevance of personality types and the diversity of personalities in startups. At the same time, it brings up other questions on how personality traits are related to other factors associated with success, such as:

1. Will the recent growing focus on promoting and investing in female founders change the nature, composition and dynamics of startups and their personalities leading to a more diverse personality landscape in startups?

2. Will the growth of startups outside of the United States change what success looks like to investors and hence the role of different personality traits and their association to diverse success metrics?

3. Many of today's most renowned entrepreneurs are either Baby Boomers (such as Gates, Branson, Bloomberg) or Generation Xers (such as Benioff, Cannon-Brookes, Musk). However, as we can see, personality is both a predictor and driver of success in entrepreneurship. Will generation-wide differences in personality and outlook affect startups and their success?

Table 2.2: Summary of the basic information of the Entrepreneurs Only (EO) dataset. The number of founders and associated startups in population, how many countries those startups are across, and the time span the data collected covers, the number of features included.

Founders with Personality Data	Associated Startups	Countries	Date Range	Founders Individual Features
32,732	23,292	215	2008-2021	100

Moreover, the findings shown here have natural extensions and applications beyond startups, such as for new projects within large established companies. While not technically startups, many large enterprises and industries such as construction, engineering and the film industry rely on forming new project-based, cross-functional teams that are often new ventures and share many characteristics of startups.

There is also potential for extending this research in other settings in government, NGOs, and within the research community. In scientific research, for example, team diversity in terms of age, ethnicity and gender has been shown to be predictive of impact, and personality diversity may be another critical dimension (AlShebli et al., 2018).

Another extension of the study could investigate the development of the language used by startup founders on social media over time. Such an extension could investigate whether the language (and inferred psychological characteristics) change as the entrepreneurs' ventures go through major business events such as foundation, funding, or exit.

Overall, this study demonstrates, first, that startup founders have significantly different personalities than employees. Secondly, besides *firm-level* factors, which are known to influence firm success, we show that a range of *founder-level* factors, notably the character traits of its founders, significantly impact a startup's likelihood of success. Lastly, we looked at *team-level* factors. We discovered in a multifactor analysis that personality-diverse teams have the most considerable impact on the probability of a startup's success, underlining the importance of personality diversity as a relevant factor of team performance and success.

2.4 Methods

2.4.1 Data sources

2.4.1.1 Entrepreneurs dataset

Data about the founders of startups were collected from Crunchbase (Table 2.2), an open reference platform for business information about private and public companies, primarily

early-stage startups. It is one of the largest and most comprehensive data sets of its kind and has been used in over 100 peer-reviewed research articles about economic and managerial research.

Crunchbase contains data on over two million companies - mainly startup companies and the companies who partner with them, acquire them and invest in them, as well as profiles on well over one million individuals active in the entrepreneurial ecosystem worldwide from over 200 countries and spans. Crunchbase started in the technology startup space, and it now covers all sectors, specifically focusing on entrepreneurship, investment and high-growth companies.

While Crunchbase contains data on over one million individuals in the entrepreneurial ecosystem, some are not entrepreneurs or startup founders but play other roles, such as investors, lawyers or executives at companies that acquire startups. To create a subset of only entrepreneurs, we selected a subset of 32,732 who self-identify as founders and co-founders (by job title) and who are also publicly active on the social media platform Twitter. We also removed those who also are venture capitalists to distinguish between investors and founders.

We selected founders active on Twitter to be able to use natural language processing to infer their Big Five personality features using an open-vocabulary approach shown to be accurate in the previous research by analysing users' unstructured text, such as Twitter posts in our case. For this project, we employed a commercial service, IBM Watson Personality Insight, to infer personality facets. IBM Watson Personality Insights operationalises an open-vocabulary (Kern et al., 2019), supervised machine learning approach consistent with prior computational personality research (Arnoux et al., 2017; Plank & Hovy, 2015; Schwartz et al., 2013), which has demonstrated that linguistic patterns in social media text can reliably predict Big Five personality traits. This service provides raw scores and percentile scores of Big Five Domains (Openness, Conscientiousness, Extraversion, Agreeableness and Emotional Stability) and the corresponding 30 subdomains or facets. In addition, the public content of Twitter posts was collected, and there are 32,732 profiles that each had enough Twitter posts (more than 150 words) to get relatively accurate personality scores (less than 12.7% Average Mean Absolute Error).

The entrepreneurs' dataset is analysed in combination with other data about the companies they founded to explore questions about the nature and patterns of personality traits of entrepreneurs and the relationships between these patterns and company success.

For the multifactor analysis, we further filtered the data in several preparatory steps for the success prediction modelling (for more details, see Appendix section A.5). In particu-

lar, we removed data points with missing values (Extended Data Figure A.13) and kept only companies in the data that were founded from 1990 onward to ensure consistency with previous research (Bonaventura et al., 2020) (see Extended Data Figure A.14). After cleaning, filtering and pre-processing the data, we ended up with data from 25,214 founders who founded 21,187 startup companies to be used in the multifactor analysis. Of those, 3442 startups in the data were successful, 2362 in the first seven years after they were founded (see Extended Data Figure A.15 for more details).

2.4.1.2 Entrepreneurs and employees dataset

To investigate whether startup founders show personality traits that are similar or different from the population at large (i. e. the entrepreneurs vs employees sub-analysis shown in Figure 2.1A and B), we filtered the entrepreneurs' data further: we reduced the sample to those founders of companies, which attracted more than US$100k in investment to create a reference set of successful entrepreneurs (n=4400).

To create a control group of employees who are not also entrepreneurs or very unlikely to be of have been entrepreneurs, we leveraged the fact that while some occupational titles like CEO, CTO and Public Speaker are commonly shared by founders and co-founders, some others such as Cashier, Zoologist and Detective very rarely co-occur seem to be founders or co-founders. To illustrate, many company founders also adopt regular occupation titles such as CEO or CTO. Many founders will be Founder and CEO or Co-founder and CTO. While founders are often CEOs or CTOs, the reverse is not necessarily true, as many CEOs are professional executives that were not involved in the establishment or ownership of the firm.

Using data from LinkedIn, we created an Entrepreneurial Occupation Index (EOI) based on the ratio of entrepreneurs for each of the 624 occupations used in a previous study of occupation-personality fit (McCarthy et al., 2022). It was calculated based on the percentage of all people working in the occupation from LinkedIn compared to those who shared the title Founder or Co-founder (See Appendix section A.2 for more details). A reference set of employees (n=6685) was then selected across the 112 different occupations with the lowest propensity for entrepreneurship (less than 0.5% EOI) from a large corpus of Twitter users with known occupations, which is also drawn from the previous occupational-personality fit study (McCarthy et al., 2022).

2.4.2 Hierarchical clustering

We applied several clustering techniques and tests to the personality vectors of the entrepreneurs' data set to determine if there are natural clusters and, if so, how many are the optimum number.

Firstly, to determine if there is a natural typology to founder personalities, we applied the Hopkins statistic—a statistical test we used to answer whether the entrepreneurs' dataset contains inherent clusters. It measures the clustering tendency based on the ratio of the sum of distances of real points within a sample of the entrepreneurs' dataset to their nearest neighbours and the sum of distances of randomly selected artificial points from a simulated uniform distribution to their nearest neighbours in the real entrepreneurs' dataset. The ratio measures the difference between the entrepreneurs' data distribution and the simulated uniform distribution, which tests the randomness of the data. The range of Hopkins statistics is from 0 to 1. The scores are close to 0, 0.5 and 1, respectively, indicating whether the dataset is uniformly distributed, randomly distributed or highly clustered.

To cluster the founders by personality facets, we used Agglomerative Hierarchical Clustering (AHC)—a bottom-up approach that treats an individual data point as a singleton cluster and then iteratively merges pairs of clusters until all data points are included in the single big collection. Ward's linkage method is used to choose the pair of groups for minimising the increase in the within-cluster variance after combining. AHC was widely applied to clustering analysis since a tree hierarchy output is more informative and interpretable than K-means. Dendrograms were used to visualise the hierarchy to provide the perspective of the optimal number of clusters. The heights of the dendrogram represent the distance between groups, with lower heights representing more similar groups of observations. A horizontal line through the dendrogram was drawn to distinguish the number of significantly different clusters with higher heights. However, as it is not possible to determine the optimum number of clusters from the dendrogram, we applied other clustering performance metrics to analyse the optimal number of groups.

A range of Clustering performance metrics were used to help determine the optimal number of clusters in the dataset after an apparent clustering tendency was confirmed. The following metrics were implemented to evaluate the differences between within-cluster and between-cluster distances comprehensively: Dunn Index, Calinski-Harabasz Index, Davies-Bouldin Index and Silhouette Index. The Dunn Index measures the ratio of the minimum inter-cluster separation and the maximum intra-cluster diameter. At the same time, the Calinski-Harabasz Index improves the measurement of the Dunn Index by calculating the ratio of the average sum of squared dispersion of inter-cluster and intra-cluster. The

Davies-Bouldin Index simplifies the process by treating each cluster individually. It compares the sum of the average distance among intra-cluster data points to the cluster centre of two separate groups with the distance between their centre points. Finally, the Silhouette Index is the overall average of the silhouette coefficients for each sample. The coefficient measures the similarity of the data point to its cluster compared with the other groups. Higher scores of the Dunn, Calinski-Harabasz and Silhouette Index and a lower score of the Davies-Bouldin Index indicate better clustering configuration.

2.4.3 Classification modelling

2.4.3.1 Classification algorithms

To obtain a comprehensive and robust conclusion in the analysis predicting whether a given set of personality traits corresponds to an entrepreneur or an employee, we explored the following classifiers: Naïve Bayes, Elastic Net regularisation, Support Vector Machine, Random Forest, Gradient Boosting and Stacked Ensemble. The Naïve Bayes classifier is a probabilistic algorithm based on Bayes' theorem with assumptions of independent features and equiprobable classes. Compared with other more complex classifiers, it saves computing time for large datasets and performs better if the assumptions hold. However, in the real world, those assumptions are generally violated. Elastic Net regularisation combines the penalties of Lasso and Ridge to regularise the Logistic classifier. It eliminates the limitation of multicollinearity in the Lasso method and improves the limitation of feature selection in the Ridge method. Even though Elastic Net is as simple as the Naïve Bayes classifier, it is more time-consuming. The Support Vector Machine (SVM) aims to find the ideal line or hyperplane to separate successful entrepreneurs and employees in this study. The dividing line can be non-linear based on a non-linear kernel, such as the Radial Basis Function Kernel. Therefore, it performs well on high-dimensional data while the 'right' kernel selection needs to be tuned. Random Forest (RF) and Gradient Boosting Trees (GBT) are ensembles of decision trees. All trees are trained independently and simultaneously in RF, while a new tree is trained each time and corrected by previously trained trees in GBT. RF is a more robust and straightforward model since it does not have many hyperparameters to tune. GBT optimises the objective function and learns a more accurate model since there is a successive learning and correction process. Stacked Ensemble combines all existing classifiers through a Logistic Regression. Better than bagging with only variance reduction and boosting with only bias reduction, the ensemble leverages the benefit of model diversity with both lower variance and bias. All the above classification algorithms distinguish

successful entrepreneurs and employees based on the personality matrix.

2.4.3.2 Evaluation metrics

A range of evaluation metrics comprehensively explains the performance of a classification prediction. The most straightforward metric is accuracy, which measures the overall portion of correct predictions. It will mislead the performance of an imbalanced dataset. The F1 score is better than accuracy by combining precision and recall and considering the False Negatives and False Positives. Specificity measures the proportion of detecting the true negative rate that correctly identifies employees, while Positive Predictive Value (PPV) calculates the probability of accurately predicting successful entrepreneurs. Area Under the Receiver Operating Characteristic Curve (AUROC) determines the capability of the algorithm to distinguish between successful entrepreneurs and employees. A higher value means the classifier performs better on separating the classes.

2.4.3.3 Feature importance

To further understand and interpret the classifier, it is critical to identify variables with significant predictive power on the target. Feature importance of tree-based models measures Gini importance scores for all predictors, which evaluate the overall impact of the model after cutting off the specific feature. The measurements consider all interactions among features. However, it does not provide insights into the directions of impacts since the importance only indicates the ability to distinguish different classes.

2.4.3.4 Statistical analysis

T-test, Cohen's D and two-sample Kolmogorov-Smirnov test are introduced to explore how the mean values and distributions of personality facets between entrepreneurs and employees differ. The T-test is applied to determine whether the mean of personality facets of two group samples are significantly different from one another or not. The facets with significant differences detected by the hypothesis testing are critical to separate the two groups. Cohen's d is to measure the effect size of the results of the previous t-test, which is the ratio of the mean difference to the pooled standard deviation. A larger Cohen's d score indicates that the mean difference is greater than the variability of the whole sample. Moreover, it is interesting to check whether the two groups' personality facets' probability distributions are from the same distribution through the two-sample Kolmogorov-Smirnov

test. There is no assumption about the distributions, but the test is sensitive to deviations near the centre rather than the tail.

AUTHOR DECLARATION

The following chapter contains content from the following publication.

Gong, Xian, McCarthy, P. X., Rizoiu, M.-A., & Boldi, P. (2024). Harmony in the australian domain space. *Proceedings of the 16th ACM Web Science Conference*, 92–102

Author Contributions: Xian Gong led the data collection, preprocessing, and analysis, developed the methodology, generated the main visualisations and results, and drafted the initial manuscript. P.X. McCarthy contributed to the conceptual framing of the study, provided guidance on linking enterprise networks to broader economic implications, and assisted in refining the manuscript. M.-A. Rizoiu supervised the research and contributed to the interpretation of results. P. Boldi contributed to the theoretical grounding of network analysis, advised on methodological robustness, and assisted with interpretation and manuscript refinement.

Production Note:
Signature removed prior to publication.

Xian Gong

Production Note:
Signature removed prior to publication.

Paul X. McCarthy

Production Note:
Signature removed prior to publication.

Marian-Andrei Rizoiu

Production Note:
Signature removed prior to publication.

Paolo Boldi

CHAPTER 3

HARMONY IN THE AUSTRALIAN DOMAIN SPACE

In this paper we use for the first time a systematic approach in the study of harmonic centrality at a Web domain level, and gather a number of significant new findings about the Australian web. In particular, we explore the relationship between economic diversity at the firm level and the structure of the Web within the Australian domain space, using harmonic centrality as the main structural feature. The distribution of harmonic centrality values is analysed over time, and we find that the distributions exhibit a consistent pattern across the different years. The observed distribution is well captured by a partition of the domain space into six clusters; the temporal movement of domain names across these six positions yields insights into the Australian Domain Space and exhibits correlations with other non-structural characteristics. From a more global perspective, we find a significant correlation between the median harmonic centrality of all domains in each OECD country and one measure of global trust, the WJP Rule of Law Index. Further investigation demonstrates that 35 countries in OECD share similar harmonic centrality distributions. The observed homogeneity in distribution presents a compelling avenue for exploration, potentially unveiling critical corporate, regional, or national insights.

3.1 Introduction

The broad field of computational social science aims to use collected data of various kinds (most notably, online data) to explore different aspects of economics, history, and psychology. The type of analysis that can be performed on the collected data may have different

purposes and the tools used are also varied, ranging from simple statistics to complex machine learning models. The Web is a particularly interesting source of data, as it is a large-scale, real-world, and continuously evolving system. Data about the Web can be collected in various ways, including crawling, scraping, and mining. There are many public collections of web crawls, but one that is known for being very reliable and quite wide in scope is the Common Crawl[1]. Common Crawl's measurements are preferred for web and network analysis due to their extensive coverage, regular updates, and large-scale, publicly accessible datasets, which reduces the need for resource-intensive data collection and is applicable across various research in a reproducible way.

One important and extremely relevant phenomenon that social science is considering recently is the decline in economic diversity. Economic diversity refers to the variety and distribution of different types of economic activities and industries within a region or economy, reflecting its breadth and resilience against market fluctuations. Research shows that more diversified economies are associated with higher levels of gross domestic product, lower unemployment rates and less instability (Malizia & Ke, 1993; Smith, 2001).

The exploration of web graph data is emerging as a powerful tool for investigating economic diversity. A large-scale study that looked at over 10 years of web data by McCarthy et al. (McCarthy et al., 2021) reveals a clear trend: a steady long-term decrease in global link diversity. This decline is attributed to the growing market hegemony of a few firms, as captured by the Herfindahl-Hirschman Index. This trend signifies an increasing monopolization of the market by a small number of corporations, raising critical questions about the future of economic diversity.

In addition to its value for macroeconomic analysis, web graph data at a domain or page level proves instrumental in investigating firm-level trends, with various measures of network centrality serving as essential tools. Comprehending the patterns in the centrality of domains within the web graph not only facilitates firm-level analysis but also potentially elucidates broader trends across the economy.

In Web analysis, various measures of centrality can be used (both at a page or at a domain level of granularity), with PageRank (Page et al., 1998) having a major role; nonetheless, recently other centrality measurements, and especially harmonic centrality (Boldi & Vigna, 2014; Boldi et al., 2022; Ortega & Eballe, 2022), were suggested as crucial for identifying influential web pages and understanding their role in the information flow and structure of the internet network. Previous research (Page et al., 1998) already showed that harmonic centrality can be used to analyze economic networks, identifying key intercon-

[1] https://commoncrawl.org/

nected industries or regions, thereby providing insights into the structure and resilience of an economy's diversity (Hussain et al., 2019).

In this paper, we use for the first time a systematic approach in the study of harmonic centrality at a Web domain level, and gather a number of significant new findings. In particular, we explore the relationship between economic diversity and the structure of the Web within the Australian domain space, using harmonic centrality as the main structural feature. From a more global perspective, harmonic centrality can be used at a country level to look at the relative global centrality of all domains within a top-level country domain space (McFarland et al., 2023) and we find a significant correlation (coefficient = 0.45 at a 5% significance level) between the median harmonic centrality of all domains in each OECD country and one measure of global trust, the WJP Rule of Law Index. Further investigation demonstrates that 35 countries in OECD share similar harmonic centrality distributions (Figure 3.1). The observed homogeneity in distribution presents a compelling avenue for exploration, potentially unveiling critical corporate, regional, or national insights.

Our focus here is mainly on the relationship between firm-level characteristics and network structure within the Australian domain space. At the beginning of our journey, we were interested in whether the multi-peaked distribution of harmonic centrality observed in the most recent snapshot of the Australian web space held true at different time points, which turned out to be true. The next step was decomposing the entire harmonic centrality distribution and analysing the quantitative and qualitative characteristics between different peaks. Finally, we use statistical analysis and machine learning algorithms to explore whether the harmonic centrality measurements contain information not explicitly stated in the web graph, such as the company's revenue, expenses, and social media links. These non-structural features are closely tied to its economic diversity, influencing financial stability, cost management, and market reach.

This study seeks to better understand what harmonic centrality captures in large-scale web domain networks and why it may matter beyond purely structural graph properties. In our analysis, we found that harmonic centrality scores are complex measures that combine information related to the network graph structure, such as the size of domains and the number of inbound links, while potentially also reflecting offline unstructured characteristics such as the size, scope, influence, and impact of the organisation. Accordingly, the aim of this paper is to systematically examine the structural and non-structural features associated with harmonic centrality, identify statistically significant correlations, and clarify the interpretive meaning of this metric in an economic context. We illustrate that it not only measures the centrality of a node in a network, providing insight into its influence and con-

nectivity within the network, but also contains other implicit information embedded in the web graph that may assist in classifying firms and anticipating patterns of economic performance. Our findings indicate that Its interconnectedness and influence within a network can reflect economic diversity, as companies with a wide range of products or services often have diverse market connections and are more resilient to market changes. Such insights may inform policy analysis concerned with digital infrastructure and industrial diversity, as well as organisational strategy related to partnership formation, market expansion, and competitive positioning.

The remainder of this paper is organised into four sections: Background and Related Work (Section 3.2), Data and Experiments (Section 3.3), Results (Section 3.4), and Discussion and Conclusion (Section 3.5). Section 2 reviews the theoretical foundations of harmonic centrality and related research on firm-level networks. Section 3 describes the datasets, feature construction, clustering methodology, statistical analyses, and classification design. Section 4 presents the quantitative and qualitative empirical findings derived from the Australian domain space. Section 5 discusses the broader economic and policy implications of the results and concludes the study. This structure, from theory to data, empirical analysis, and broader interpretation, is a typical format for a proceeding paper that ensures methodological transparency while clearly linking technical findings to their economic significance.

3.2 Background and Related Work

3.2.1 Harmonic Centrality

Harmonic centrality (Beauchamp, 1965) is a measure of centrality of nodes in a graph which was proposed to solve the issue of unreachable vertices in closeness centrality (Boldi & Vigna, 2014). In particular, we can define the *harmonic centrality* of a node u in a directed graph G as

$$h(u) = \sum_{v \in N_G \setminus \{u\}} \frac{1}{d_{uv}}$$

where N_G is the set of nodes of G, d_{uv} is the length of the shortest path from u to v and we assume $1/\infty = 0$. Harmonic centrality has a number of theoretical and practical qualities (see, for instance, (Boldi & Vigna, 2014; Boldi et al., 2023)) that make it quite promising as a measure of importance in general directed networks.

How harmonic centrality values are distributed in certain real-world networks is an underexplored area of research. Since we are using harmonic centrality distribution within

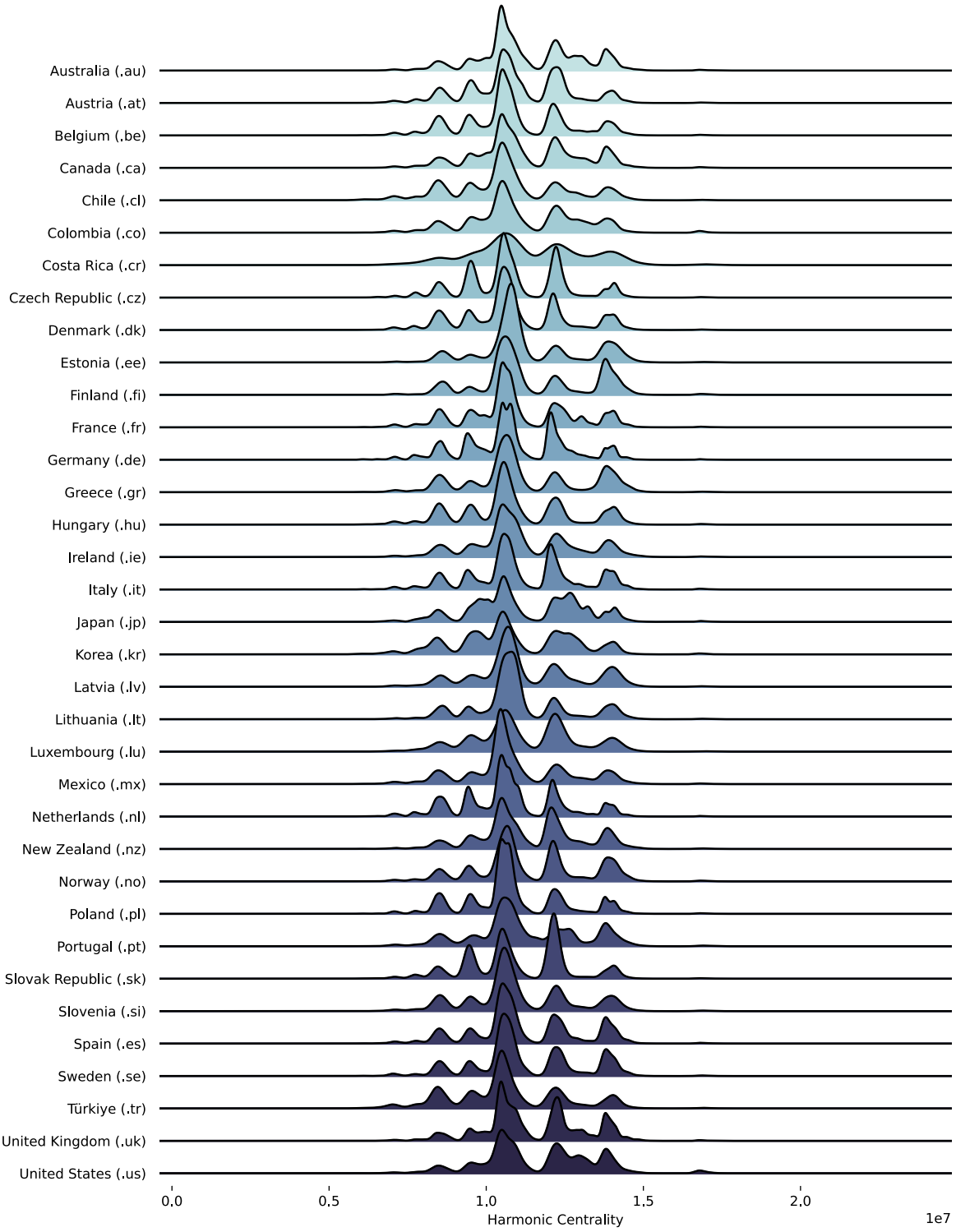


Figure 3.1: Empirical Harmonic Centrality Distributions of Web Domains among 35 Countries in 2023.

domain graphs as a way to measure some underlying phenomena, we wonder what the *natural* distribution we can expect is.

In general, understanding the reasons behind (or even just reproducing) certain observed behaviors and distributions of some characteristics of real graphs are challenging, even in the case of very simple features such as in- and out-degrees—this is, in fact, the main quest of the theory of random graph models of complex networks (Albert & Barabási, 2002; Watts & Strogatz, 1998).

A well-known scale-free model of directed networks which is expected to represent faithfully web graphs and their offspring is described in (Bollobás et al., 2003): it is an evolving network model, where the network grows one node/arc at every step, based on four parameters called α, β, δ^- and δ^+: with probability α, a new node v is added along with an arc *to* an existing node, chosen based on its in-degree; with probability β, a new arc *between* two existing nodes is added, whose source and target are chosen based on their out- and in-degree, respectively; with probability $1-(\alpha+\beta)$, a new node v is added with an arc coming *from* an existing node, chosen based on its out-degree. When choosing a node based on its in- and out-degrees we, in fact, adopt a bias, and $\delta^\pm$ represent the two biases (see (Bollobás et al., 2003) for details).

Now, using the same analysis as in Section 4 of (Bollobás et al., 2003), we assume that $\delta^+ = 0$ (i.e., that outdegrees are unbiased), obtaining the following relations between in-degree exponent X^-, outdegree exponent X^+ and the other parameters:

$$\alpha = \frac{X^+ - 2}{X^+ - 1}$$

$$\alpha = \frac{1 + \delta^-(1-\beta)}{\alpha + \beta}$$

Based on the measurements of (Meusel et al., 2015), in- and out-degrees of host graphs fit well a power-law distribution with in-degree exponent $X^- = 2.12$ and out-degree exponent $X^+ = 2.14$, which gives

$$\alpha \approx 0.12281$$

and values of β in the range $[0.77, 0.87]$. Taking β in the midpoint of the range (i.e., $\beta = 0.82$), we obtain $\delta^- = 0.31078$. For these values of the parameters, the distribution of harmonic centrality values is shown in Figure 3.2 (on a sample with 700 000 nodes).

The large mode on 0 is an artifact of the model, which produces many isolated nodes. However, the distribution corresponds reasonably well to the observed multi-modality, although in our datasets there are more modes and the corresponding densities are more similar to one another.

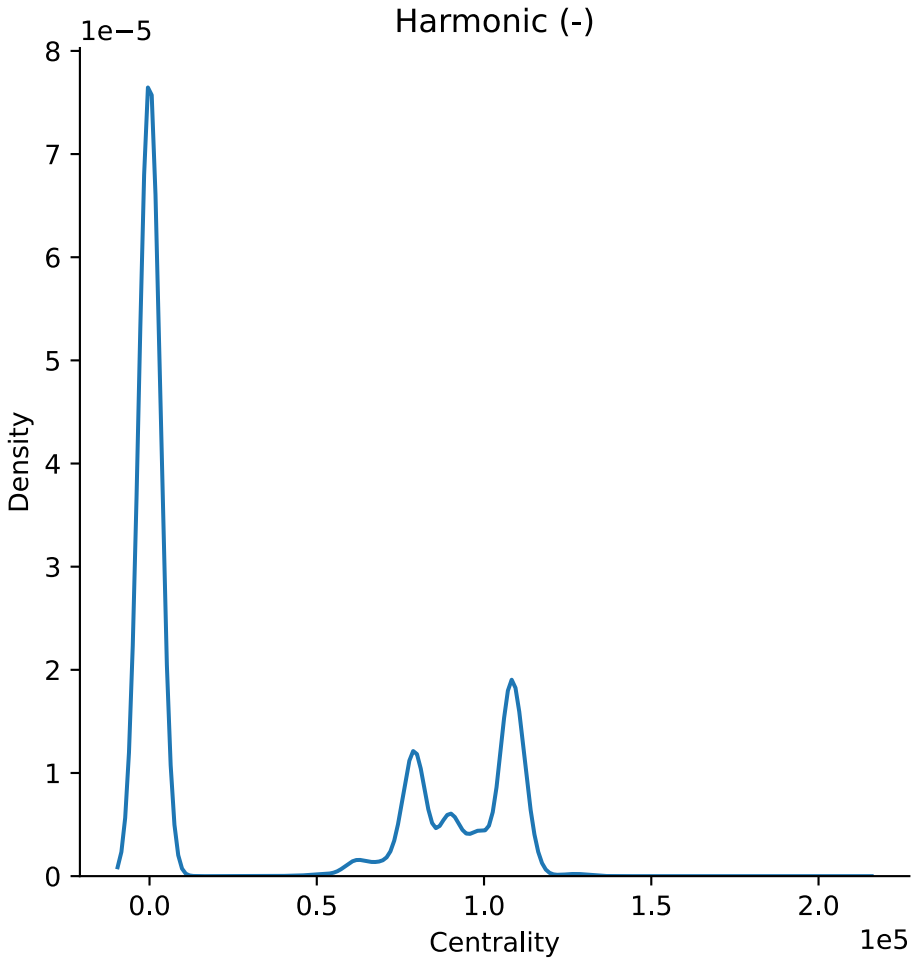


Figure 3.2: Expected distribution of harmonic centrality values according to a theoretical model and a sample network graph with 700 000 nodes (Bollobás et al., 2003).

3.2.2 Applications on Firm-Level

In this study, we apply the concept of harmonic centrality at the domain (website) level, with the aim of providing valuable insights into the structure and dynamics of the web at a firm level. It can be a useful tool for understanding firm-level economic diversity.

One previous study used harmonic centrality to analyse structures of other possible networks, such as global supply chains (Lavassani & Movahedi, 2021). Here researchers mined data from financial records to construct the global supply chain network of the auto manufacturing sector and explored the effect of these supply-chain centrality measures on firms' financial performance, investment risk, and market value volatility.

In another study (Riccaboni et al., 2021) authors introduced a new normalized firm centrality measure independent of corporate group size. This centrality notion was computed

on a set of global corporate networks encompassing 17.8 million firms. They found a positive relationship between firm centrality and firm performance, but the significance of the relationship decreases as corporate group size increases.

Recent research, grounded in the concept of harmonic centrality, introduces two proximate metrics to examine the connection between senior executives' social connections and the firm's value (Rao et al., 2022). This study demonstrates the significant impact of top executives' social networks on the firm's market capitalisation by correlating the Chief Marketing Officer's network centrality with the company's market value.

While various centrality metrics are employed to scrutinise distinct networks, the underlying objective remains consistent. The aim is to unearth effective measures for evaluating firm-level data to assist in decision-making processes. Though not focused on localised networks, the extensive and all-encompassing web graph from Common Crawl utilised in our study enables the extraction of implicit information on a broader scale. For businesses, harmonic centrality stands out as a more comprehensible and readily accessible metric.

3.3 Data and Experiments

This section describes the sources of datasets used for this study. We describe the structural and non-structural features used for the experiments in Table 3.1 (Section 3.3.1), the design of the clustering analysis (Section 3.3.2), and the statistical analysis (Section 3.3.3) followed by classification prediction (Section 3.3.4).

3.3.1 Dataset Description

We utilise the datasets mainly from Common Crawl and BuiltWith in our study. Common Crawl provides the domain-level graph that encapsulates the relationships among over 88 million domains (websites) found on the World Wide Web across different periods. It is a network representation of domain interconnections, offering insights into the web's link structure, domain attributes and the prominence of websites within the web ecosystem. Harmonic centrality is one of the structural features of the domain-level graph. As explained above, it is based on the reciprocal distance between nodes, emphasising the importance of nodes that are closer to other nodes in terms of network traversal. Moreover, it excels at identifying nodes with strong local influence within tightly-knit communities or clusters. Unlike PageRank, it is less susceptible to manipulation and provides an intuitive interpretation, measuring the average reciprocal distance between a node and all oth-

Table 3.1: Descriptions of Structural and Non-Structural Features of Australian Domains

	Features	Description
Structural	Harmonic Centrality	Node importance in a network depends on close connections, especially shorter distances.
	Domain Size	The number of web pages or websites hosted under that specific domain name.
	inbound	The number of incoming links or connections to a specific web page or website.
	outbound	The number of links or connections originating from a specific web page or website.
	Tranco	Website popularity and ranking determined by web traffic and domain authority assessment.
Non-Structural	Tech Spend	Monthly expenditure estimate on tech services and tools for websites or online businesses.
	Sales Revenue	An estimate of the monthly or annual sales revenue of eCommerce websites.
	Social	Estimation of social media links, encompassing Facebook, Twitter, Google+, Pinterest, GitHub, and LinkedIn.
	Employees	An estimate of amount of employees for websites.
	Age	Years since this website was first indexed until now.

ers, which is particularly useful for detecting community structures and analysing dynamic networks.

BuiltWith is a web technology profiling and analysis tool that focuses on providing insights into the technological infrastructure of websites. Its core functionality involves detecting and reporting on the web technologies, frameworks, content management systems, and software tools utilised by websites. BuiltWith offers users the ability to ascertain the structural elements of web development, including programming languages, web servers, and security measures. While it primarily addresses the structural aspects, it provides basic information about non-structural features of domains, such as technology spending, sales revenues, and employees, which offer insights into domains and economic indicators.

Common Crawl and BuiltWith have different focuses and capture different types of data, and their approaches and frequencies to tracking dynamic changes of features over time vary. Common Crawl primarily focuses on capturing the static content of web pages and typically conducts large-scale crawls every few months to update its datasets. BuiltWith specialises in profiling the web technologies and software tools used by websites. It does not capture dynamic changes in website content or features over time. Instead, it provides a snapshot of the technologies in use at the time of analysis. Even though there is a time mismatch between Common Crawl and BuiltWith, we assume that the features of domains will remain stable for a short period. Therefore, our further analysis uses the most recent domain ranks from Common Crawl and the current snapshot of basic information from

BuiltWith.

We obtained the domain name and its Harmonic Centrality from the Common Crawl website for the last quarter of each of the six years from 2018-2023. Since this article only focuses on the Australian domain name space, we filter the entire Common Crawl websites by domain names ending in .au. The number of Australian domain names in these six years ranged from 744k to 977k. The number of domain names in Australia remained above 900k in the first five years but will drop to 744k in 2023. The sharp decline in Australian domain names between 2022 and 2023 is primarily due to the introduction of .au direct domains, which allows entities to register shorter .au domain names. This resulted in a restructuring of domain registration in Australia, probably reducing the use of .com.au, .net.au, and .org.au domains as people moved to the shorter .au domain. We use Australian domains in 2023 as a benchmark to obtain features from BuildWith due to its snapshot characters. We found 744,058 Australian domain names in 2023, and based on this we collected the features of Table 3.1.

3.3.2 Clusters Movements over time

In our investigation of the Australian domain space, we conducted an extensive analysis of harmonic centrality data obtained from the Common Crawl dataset spanning the years 2018 to 2023. Our examination primarily focuses on the harmonic centrality distributions observed in the first quarter of each of these six years. As illustrated in Figure 3.3, it is evident that the harmonic centrality distributions exhibit a consistent pattern across the different years. To gain deeper insights into the distinctive peaks within these distributions, we employed a clustering algorithm, specifically KMeans, to decompose each distribution. Using clustering algorithms to analyse data with multimodal distributions over time is beneficial because these algorithms can effectively identify and separate distinct patterns or trends within the data. This approach facilitates a deeper understanding of the dynamics and variability in the data over time, which is essential for concretely observing the changes in domains over the past six years and analysing the changes in those non-structural features under different movement scenarios. To determine the optimal number of clusters for each year, we utilised the average of the Davies–Bouldin index and the Calinski–Harabasz index. The Calinski-Harabasz index (Caliński & Harabasz, 1974) assesses the ratio of within-cluster dispersion to between-cluster dispersion for all clusters, whereas the Davies-Bouldin index (Davies & Bouldin, 1979) quantifies the similarity between clusters, thereby allowing us to evaluate the separability of categories. To align the Davies–Bouldin index with the Calinski–Harabasz index for ease of interpretation, we took the reciprocal

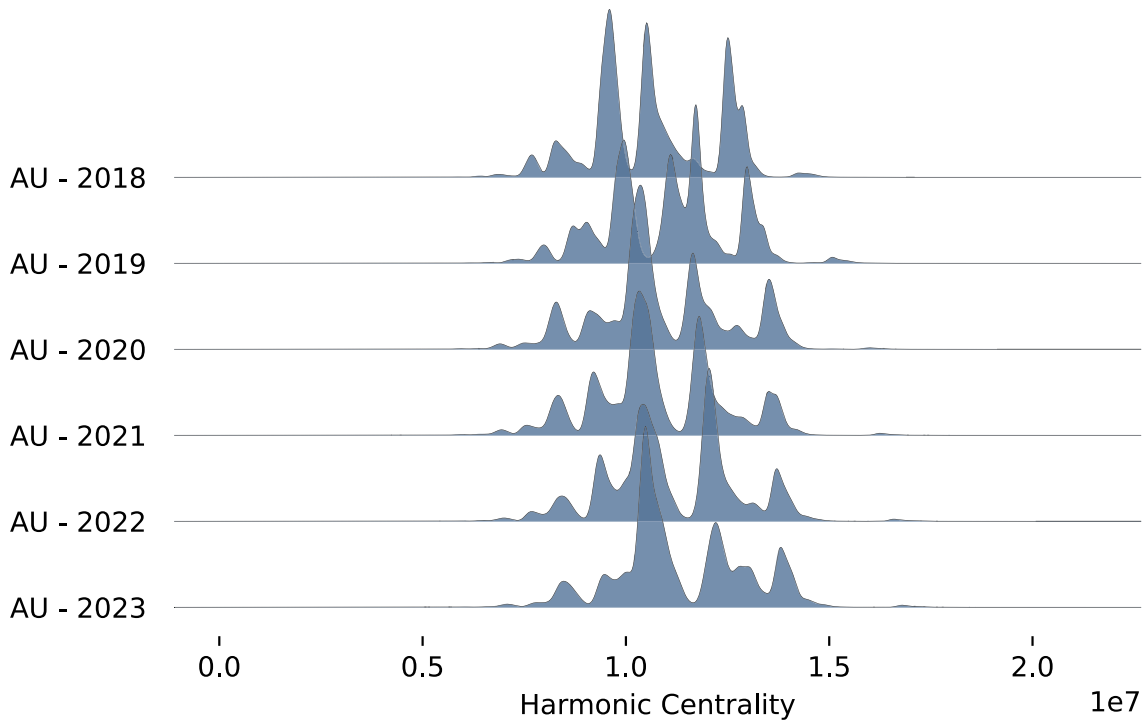


Figure 3.3: Harmonic Centrality Distribution Over Time from 2018 to 2023 within the Australia (.au) Web Domain.

of the former, with a higher value indicating superior clustering performance, and then we combined the two indices by taking their average.

In Figure 3.4, the blue line represents the average combined index along these six years, while the grey area denotes the one-standard deviation interval from the mean. Notably, the score reaches its maximum when selecting six clusters, after which it gradually diminishes. Consequently, we opted to employ six clusters to partition the annual distributions, with the cluster possessing the lowest harmonic centrality value designated as Cluster 0, and the cluster with the highest harmonic centrality value labelled as Cluster 5, following an ascending order. This arrangement results in six clusters with fixed relative positions, and the harmonic centrality values ascend sequentially. With each domain name assigned to its respective clustering category for each year, we were able to analyze the temporal movement of domain names across these six positions. This analysis aims to discern whether such movements yield insights into the Australian Domain Space or exhibit correlations with other characteristics.

For instance in Figure 3.5, when scrutinizing cluster movements at the two extremes of

Figure 3.4: Six clusters appear as the optimal number within the Harmonic Centrality Distributions. The graph shows the average and standard deviation of combined scores (Calinski-Harabasz and inverse Davies-Bouldin indices) on KMeans performance on harmonic centrality distributions of Australian Web Domain Space from 2018 to 2023.

the Australian domain space timeline between 2018 and 2023, we observed that the sources of most clusters were dispersed with the exception of Cluster 4 (the majority of current domain names in Cluster 4 originated from Cluster 4 six years prior). Additionally, the sizes of high harmonic centrality clusters in 2018 (such as Clusters 4 and 5) were notably larger than their counterparts in 2023, suggesting a trend towards greater concentration of the Australian web space on a selected few domain names. For a more comprehensive statistical summary of our findings, please refer to Section 4.

3.3.3 Statistical Analysis within Clusters

This section presents our process for conducting statistical analyses of harmonic centrality in conjunction with the structural and non-structural attributes outlined in Table 3.1. Our primary objective is to investigate a statistically significant linear association between

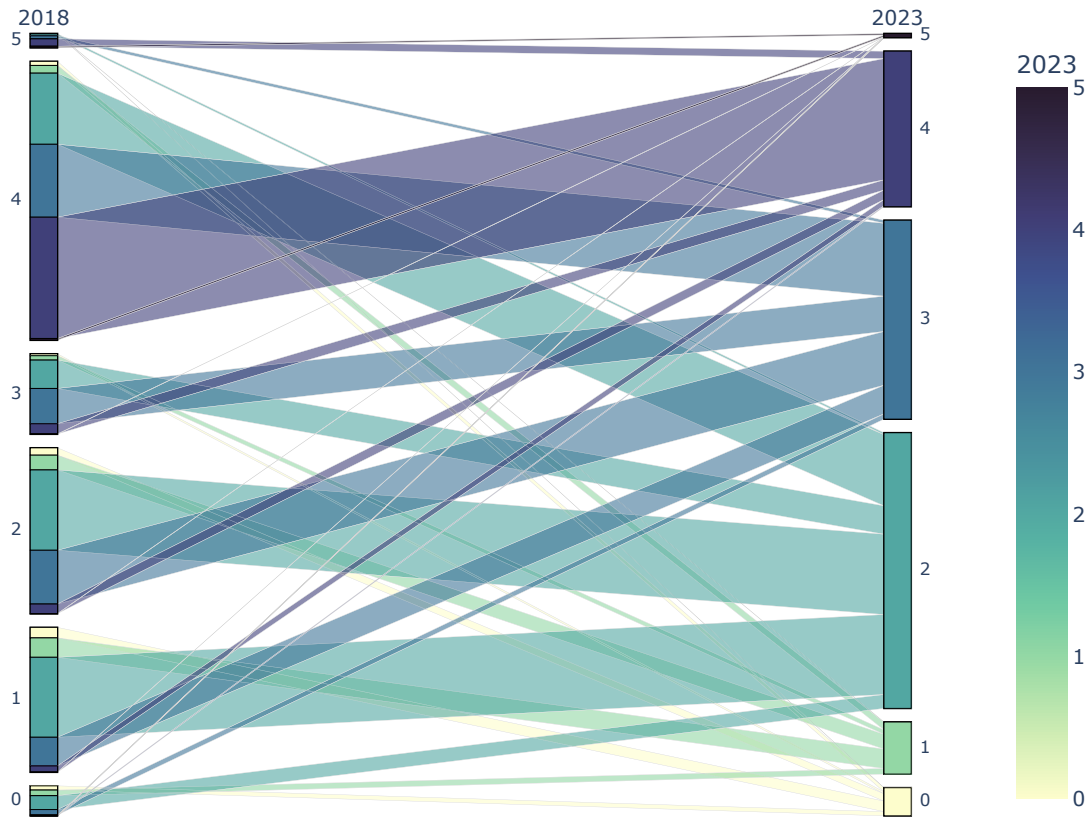


Figure 3.5: Cluster Movements of Australian Domains between 2018 and 2023.

harmonic centrality and these attributes within the six clusters. We employ three distinct statistical tests to evaluate these hypotheses: Pearson correlation, Spearman correlation, and Kendall rank correlation. Pearson Correlation is a statistical metric designed to quantify the strength and direction of the linear relationship between two continuous variables. On the other hand, Spearman Correlation is a non-parametric statistical approach that assesses the magnitude and direction of the monotonic connection between two variables. Kendall Rank Correlation, also referred to as Kendall's Tau, is another non-parametric statistical technique employed to gauge the strength and direction of the association between two variables. These three distinct statistical measures are applied in tandem to analyse the relationships between variables comprehensively. Pearson correlation proves effective in capturing linear relationships, while Spearman and Kendall's correlations are particularly valuable for detecting monotonic relationships, including those that may be non-linear in nature. We broaden our scope by employing all three measures to encompass a wider array of potential connections between variables. This approach enhances the robustness and informativeness of our analyses, especially in scenarios where the nature of the relation-

ship is uncertain or when the data do not adhere to assumptions of normal distribution.

To acquire representative samples from the six sub-distributions for the purpose of hypothesis testing, we opted to select data falling within one standard deviation of the mean as our sample set. It is worth noting that any instances with missing values were excluded from consideration during the testing of individual features. Consequently, the sample sizes for each feature may not be uniform, but we took measures to ensure that the sample sizes remained sufficiently large for statistical analysis.

3.3.4 Clusters Prediction

Our primary focus lies in examining the significance of non-structural attributes in determining the classification of Australian domains into specific clusters. These non-structural attributes contain information not encompassed within the web graph and are intricately linked to a company's economic performance. Furthermore, our statistical analysis has revealed that these non-structural attributes exhibit a non-linear association with harmonic centrality values.

In this segment of our experimentation, we have employed a machine learning classification algorithm to assess the predictive capabilities of these non-structural attributes and establish their relative importance. Specifically, we have utilised the XGBoost classifier and fine-tuned its parameters for this purpose. The XGBoost Classifier (Chen & Guestrin, 2016) is renowned for its exceptional accuracy and efficiency in machine learning tasks. It leverages decision trees to address intricate data relationships, facilitates feature selection, and mitigates overfitting, rendering it a versatile choice for handling classification tasks. Although SVM and Random Forest allow better explainability, XGBoost offers advantages over them in its ability to efficiently handle large and sparse datasets, offer higher performance with features like regularisation and custom optimisation, and generally deliver better accuracy, especially for complex problems. However, the effectiveness of these algorithms can vary based on the specific nature of the data and the problem at hand. In subsequent research, we can introduce other classifiers and compare their performance.

It is evident that the categories within our dataset exhibit a pronounced imbalance. Initially, the classifier is trained to utilise the original dataset. Subsequently, a method of random sampling is employed to extract 2,000 samples from each cluster, thereby facilitating the creation of a new, balanced dataset. This newly constituted dataset is then utilised to retrain the classifier. A comparative analysis is then conducted to evaluate the variations in accuracy between the classifiers trained on the original unbalanced dataset and the newly balanced one.

3.4 Results

In this section, we consolidate the empirical findings from Section 3 into a coherent analytical narrative, encompassing the quantitative and qualitative dimensions.

3.4.1 Quantitative Findings

Basic characteristics of six clusters Our experimental analysis dissected nearly six years of harmonic centrality distributions into six distinct sub-distributions defined by specific clustering criteria. This division categorises the trajectory of domain movements over the six periods into three scenarios: firstly, domains remaining static within their initial cluster (constant); secondly, domains ascending from clusters of lower harmonic centrality to those with higher values (upward moving); and thirdly, domains descending through the cluster hierarchy (downward moving).

Table 3.2 details the mean values or ratios of non-structural features across varied scenarios. The 'Current cluster' in the table presents the count of domain samples per cluster for 2023, along with the average value of each feature. The 'Constant cluster' contains domains of six clusters that have stayed within the same cluster across the six periods. This row still demonstrates the number of samples in each cluster and the corresponding mean values of features.

The latter two scenarios, 'Upward moving' and 'Downward moving', provide a comparative analysis, examining the mean value ratios between 'moving clusters' and 'Constant clusters'. For example, in the 'Upward moving' scenario, there are 331 domains moved from lower-ranked clusters (# 0 - # 4) to # 5 while there are 424 domains kept staying at the highest-ranked cluster # 5 during the six years. A ratio of 0.22 in the Sales Revenue column of upward moving scenario means that the mean value of sales revenue of the 331 upward-moving domains is divided by the mean value of 424 domains in constant cluster # 5. By analogy, the ratio of 1.19 in the same column means that the mean value of 3,518 domains moved from lower-ranked clusters (# 0 - # 3) to # 4 divided by the mean value of 2,279 domains in constant cluster # 4. The 'Downward moving' scenarios work oppositely. It measures the ratios of down-moving domains to the corresponding low-ranked cluster. For instance, the ratio of 2.96 in the Sales Revenue means that the mean value of 1,077 domains moved from the higher-ranked cluster (only # 5 in this case) to # 4 divided by the mean value of 2,279 domains in constant cluster # 4.

Based on the first two scenarios, domains in a higher-ranked cluster exhibit a corresponding increase in average sales revenue and technological expenditure, following an ex-

Table 3.2: Statistics Summary (Mean/Ratio) of Three Scenarios for Six Clusters

Scenarios	**Cluster** (Sample Size)	**Sales Revenue** (Mean)	**Tech Spend** (Mean)	**Social** (Mean)	**Employees** (Mean)
Current	# 5 (n=2,965)	48,323.85	1,226.92	18,864	300
	# 4 (n=108,084)	8,943.13	272.30	5,539	24
	# 3 (n=187,389)	5,067.65	181.38	1,952	6
	# 2 (n=310,480)	3,008.84	101.64	1,169	1
	# 1 (n=81,761)	1,909.26	78.93	303	1
	# 0 (n=53,379)	2,121.93	74.33	616	0
Constant	# 5 (n=424)	120,108.77	3,188.61	48,737	1,153
	# 4 (n=2,279)	10,622.74	344.03	3,735	8
	# 3 (n=3,207)	4,265.89	203.16	791	8
	# 2 (n=4,011)	2,006.97	66.52	457	0
	# 1 (n=502)	1,026.21	32.95	162	0
	# 0 (n=230)	0.00	7.50	0	0

Scenarios	**Cluster** (Sample Size)	**Sales Revenue** (Ratio)	**Tech Spend** (Ratio)	**Social** (Ratio)	**Employees** (Ratio)
Upward Moving	to # 5 (n=331)	0.22	0.29	0.07	0.04
	to # 4 (n=3,518)	1.19	1.17	1.07	0.61
	to # 3 (n=7,540)	2.51	1.61	2.56	0.90
	to # 2 (n=2,851)	4.24	3.51	1.80	2.10
	to # 1 (n=240)	3.22	4.36	0.43	
Downward Moving	to # 4 (n=1,077)	2.96	2.87	1.91	10.57
	to # 3 (n=3,624)	3.02	1.94	3.42	0.99
	to # 2 (n=4,406)	3.66	3.49	12.59	8.05
	to # 1 (n=1,457)	3.31	3.61	2.74	
	to # 0 (n=420)		7.51		

ponential trend. The 2023 data set was employed to highlight the characteristics of these six distributed non-structural attributes. Figure 3.6 graphically represents the average values of each non-structural attribute for domains that maintain their position within the same clusters throughout the observed period. This visualisation demonstrates an exponential increase in these mean values, concomitant with rising harmonic centrality, particularly pronounced within the uppermost clusters (those with the highest harmonic centrality). Such findings underscore the markedly distinct harmonic centrality levels across the six clusters, hinting at different economic performances at the firm level.

Given the limitations of non-structural feature data (only available in 2023), our analytical scope regarding cluster movement scenarios is inherently constrained. Observations derived from these ratios indicate that with the exception of advancements from the lowest-ranking cluster to the uppermost cluster (#5) — which do not exhibit an enhancement in features exceeding a factor of 1 — ascensions through other clusters manifest a discernible impact on feature mean values. Nevertheless, it is noteworthy that the ratio associated with the declining scenarios generally surpasses that observed in ascending scenarios, suggesting a predominant influence exerted by the directionality of cluster movement.

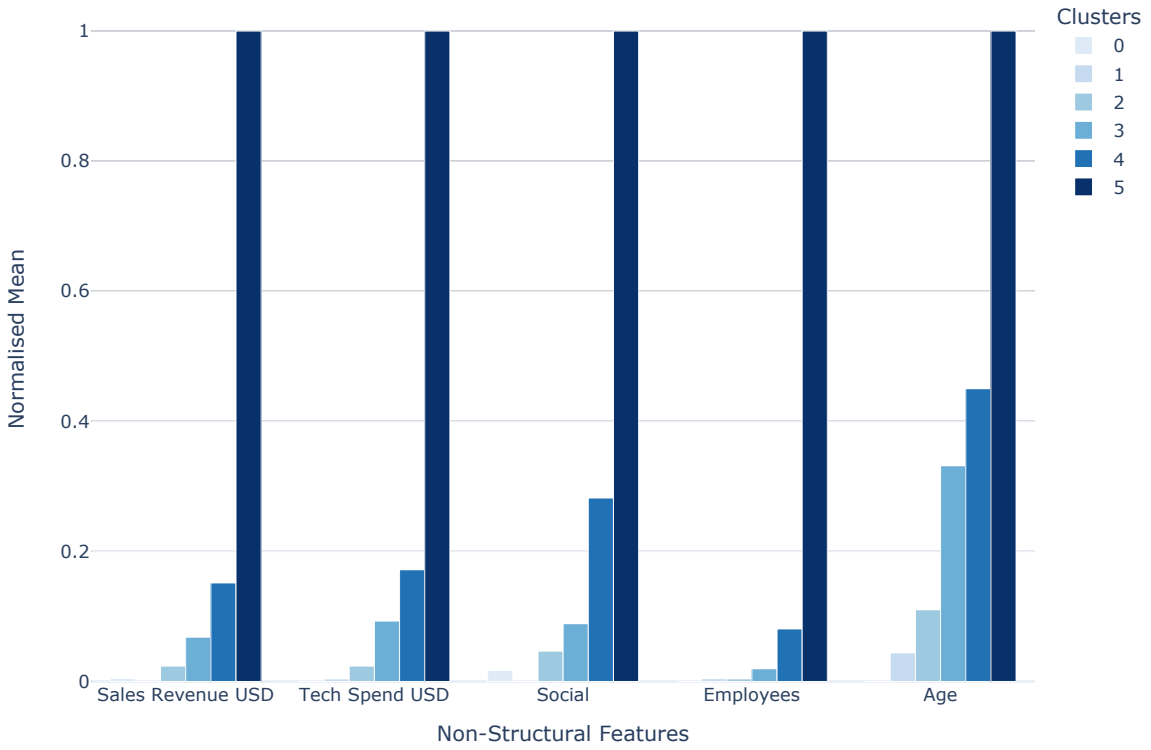


Figure 3.6: Significant different means of non-structural features among six clusters of AU Domain Space in 2023.

Correlations between harmonic centrality and (Structural and Non-structural) Structural features. In addition to comparing the mean values among six clusters, we also explored the linear relationship between harmonic centrality and structural and non-structural features under each cluster. We summarise the results in Table 3.3. In clusters exhibiting large harmonic centrality values, such as clusters 3, 4, and 5, the majority of hypothesis tests display statistical significance. However, numerous missing values within the three characteristics of Tranco, Social, and Employees introduce a degree of uncertainty, potentially making some of these tests non-statistically significant. This observation suggests a need for a cautious interpretation of these results, acknowledging the limitations imposed by the incompleteness of the data for these specific attributes. Interestingly, clusters with low harmonic centrality values, such as clusters 0 and 1, negatively correlate significantly with Sales Revenue. Observations indicate that for certain features, including inbound links, Tech Spend, Sales Revenue, and Social metrics, the Spearman and Kendall

Table 3.3: Correlation Analysis of Structural and Non-Structural Features with Harmonic Centrality. Pearson Correlation (P), Spearman Correlation (S) and Kendall's Tau (K) are used to measure the coefficients. Stars are used to indicate the significance level: 5% (*), 1% (**) and 0.1% (***).

	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
DomainSize (P)			0.02 ***	0.02 ***	0.01 *	0.15 ***
DomainSize (S)	0.01 **		0.05 ***	0.03 ***	0.12 ***	0.52 ***
DomainSize (K)	0.01 **		0.04 ***	0.03 ***	0.09 ***	0.4 ***
inbound (P)	0.04 ***	0.05 ***	-0.02 ***	0.03 ***	0.11 ***	0.52 ***
inbound (S)	0.06 ***	0.05 ***	-0.09 ***	0.15 ***	0.38 ***	0.84 ***
inbound (K)	0.05 ***	0.04 ***	-0.05 ***	0.1 ***	0.29 ***	0.69 ***
outbound (P)			0.04 ***	0.05 ***	0.04 ***	0.2 ***
outbound (S)	0.07 **		0.1 ***	0.04 ***	0.1 ***	0.39 ***
outbound (K)	0.05 **		0.07 ***	0.03 ***	0.07 ***	0.27 ***
Tranco (P)		-0.33 **			-0.17 ***	-0.33 ***
Tranco (S)		-0.34 **	0.08 *	-0.07 *	-0.16 ***	-0.38 ***
Tranco (K)		-0.23 **	0.06 *	-0.05 *	-0.11 ***	-0.26 ***
Tech Spend (P)			0.03 ***	0.05 ***	0.05 ***	0.27 ***
Tech Spend (S)		0.01 *	0.03 ***	0.06 ***	0.15 ***	0.31 ***
Tech Spend (K)		0.01 *	0.02 ***	0.05 ***	0.11 ***	0.23 ***
Sales Revenue (P)			0.02 ***	0.02 ***	0.05 ***	0.16 ***
Sales Revenue (S)	-0.03 *	-0.02 **	0.03 ***	0.02 ***	0.09 ***	0.05 *
Sales Revenue (K)	-0.02 *	-0.01 **	0.02 ***	0.02 ***	0.07 ***	0.04 *
Social (P)			0.01 **			0.1 ***
Social (S)			0.06 ***	0.02 ***	0.11 ***	0.26 ***
Social (K)			0.05 ***	0.01 ***	0.08 ***	0.2 ***
Employees (P)						0.13 ***
Employees (S)			0.02 ***	0.03 ***	0.02 ***	0.1 ***
Employees (K)			0.02 ***	0.03 ***	0.02 ***	0.08 ***
Age (P)		0.03 ***	-0.01 ***	0.04 ***	0.13 ***	0.45 ***
Age (S)		0.03 ***	-0.02 ***	0.05 ***	0.2 ***	0.48 ***
Age (K)		0.02 ***	-0.01 ***	0.03 ***	0.14 ***	0.34 ***

rank correlation coefficients exhibit higher values than the Pearson correlation coefficient. This discrepancy suggests the existence of a robust non-linear relationship between these variables and harmonic centrality, emphasising the complexity of their interdependencies beyond linear associations.

Insights from the classification prediction. Using the XGBoost classification algorithm with tuning and cross-validation, we studied the predictive efficacy of non-structural features in six clusters since we are more interested in the relationship between the harmonic centrality of structural features and those non-structural features. The f1 score of the classifier trained on the original dataset is 0.22, while the f1 score of the classifier trained on the balanced dataset is 0.29. The accuracy has not improved much, but we have very low accuracy for clusters 0 and 5 in the first case. In the second case, the prediction accuracy for clusters 0 and 5 is much improved. Figure 3.7 demonstrates the importance of the feature as ranked by XGBoost. The above two classifiers have the same order of importance for non-structural features. The sparsity of our feature matrix, coupled with the exclusive reliance on non-structural features for model training, has resulted in suboptimal accuracy levels. However, an analysis of feature importance reveals that company information, not encapsulated within the web graph, does exert a discernible influence on the predictive capability regarding the classification of domain names into categories of high or low harmonic centrality. This observation underscores the potential impact of external data sources in enhancing predictive accuracy. The demonstrated predictive capability affords a novel vantage point in evaluating a corporation's economic efficacy. Overlooked structural attributes, such as harmonic centrality, emerge as newly identified determinants influencing corporate performance metrics.

3.4.2 Qualitative Findings

In this section, we delve into a qualitative examination of each cluster's inherent characteristics, drawing upon specific domain examples that typify the respective clusters[2]. Our focus is particularly on domains that have maintained a consistent position within clusters across all periods, as this consistency aligns with one of the previously identified scenarios in our study. To elucidate these findings, we comprehensively summarise their characteristics and prominent domain company instances in Table 3.4.

Domains that have steadfastly remained within the highest harmonic centrality clusters for the whole duration of six years are generally well-established, longstanding organisations. These entities are characterised by their iconic and high-profile branding, a testament to their enduring presence and recognition in their respective fields. Furthermore, these organisations typically command high-traffic websites, which serve as digital hubs attracting large, diverse audiences. This aspect reflects their extensive reach and in-

[2]The interested reader can get more information from the full dataset, anonymously available at this link.

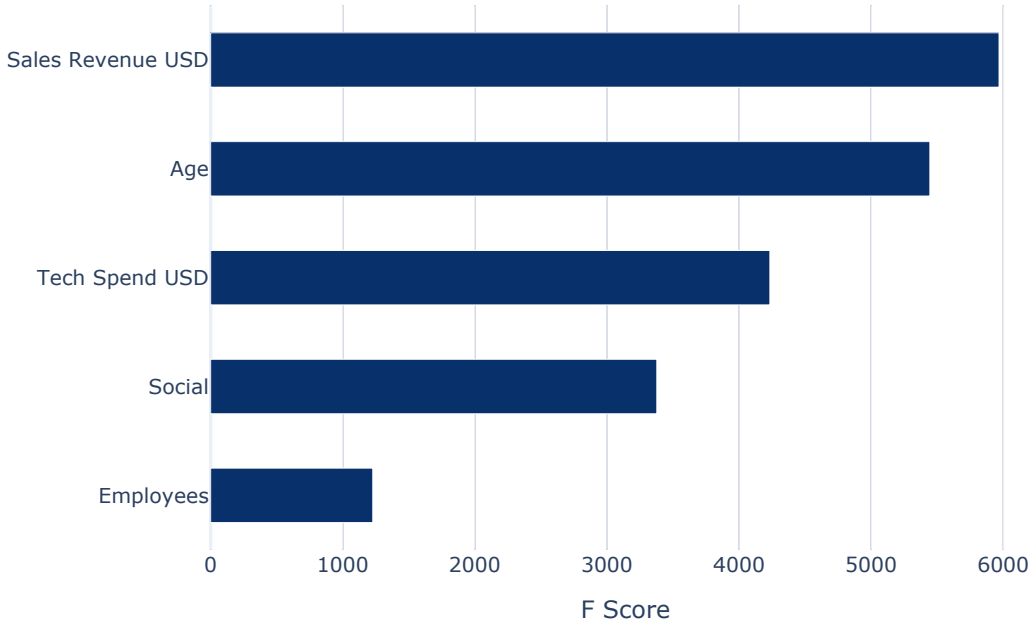


Figure 3.7: Feature importance plot of XGBoost Classifier trained on the balanced dataset. The F-score quantifies the relative significance of each feature within the classifier based on how frequently each feature is used to split the data across all trees in the ensemble.

fluence, both nationally and globally. Often, these enterprises are headquartered in major metropolitan cities, which serves as a nexus for economic, cultural, and technological advancements, further reinforcing their prominence and impact.

Conversely, a contrasting picture emerges when examining domains that have consistently been a part of the lowest harmonic centrality clusters. These domains are commonly associated with younger, newly established enterprises. Their genesis in more recent times often means they are still in the nascent stages of brand development and market penetration. These entities tend to be geographically situated in regional or remote areas, focusing on localised markets and specific niches. The nature of their business often results in smaller-scale firms mirrored in their digital presence through low-traffic websites that cater to a more limited, often specialised audience. This dichotomy between the high and low harmonic centrality clusters underscores the diverse spectrum of enterprises and their varying stages of growth, market presence, and digital influence, revealing the multifaceted

Table 3.4: Characteristics of organisations in whose domain stayed in the same cluster for six years from 2018 to 2023

	Stable Cluster 5	Stable Cluster 4	Stable Cluster 3	Stable Cluster 2	Stable Cluster 1	Stable Cluster 0
Characteristics	Established leading national and global consumer-facing brands, media, universities and Government.	Australian presence of global brands and national challenger brands.	Large suburban businesses in cities and metropolitan areas.	Local CBD businesses and niche speciality businesses.	Local businesses many in regional areas.	Local businesses many in country towns and remote areas.
Examples	Qantas, ASX, ATO, Australia Post, Woolworths, Coles, BoM, CSIRO, ABC, UNSW, University of Sydney	Loreal Australia, Ladbrokes Australia, DHL Australia, Foodworks.	RayWhite Drummoyne (real estate agency), Manly Cycles, Hobart Yachts, Perth Video	Vision blonde (hair salon), Viking Genetics (Cattle breeding)	Katie Smith Solicitor, Olivier Flowers (Central Coast), Northern Rivers Volkswagen (car dealer)	IGA Donnybrook (Donnybrook)

nature of the economic and digital landscape.

In the second scenario, we focus on domains ascending within the harmonic centrality cluster from 2018-2023. Notable instances of this trend include the Australian pharmacy conglomerate Chemist Warehouse, which, according to LinkedIn data, has been expanding at a rate exceeding 8% annually, now boasting a network of over 550 pharmacies. Similarly, Perfect Portal, the UK-based online legal software provider for small practices, has shown a remarkable annual growth of 18% and extended its operations to Australia. Another example is the men's health charity "Mr Perfect", which adopts a community-oriented network model, orchestrating fundraising events like barbeques. Additionally, the global parking application ZipBy exemplifies this upward trajectory with an impressive 67% annual growth rate. These cases are emblematic of organisations typically experiencing rapid expansion in terms of employee count and sales revenue. They often represent young, globally oriented enterprises venturing into new markets, utilising models based on franchising, networking, or app-based services, such as mobile parking applications. This pattern highlights a dynamic where organisations leverage digital platforms and global connectivity to foster significant growth and market penetration.

Our analysis of the final scenario focused on domains that exhibited a downward movement in harmonic centrality clusters during 2018-2023. This category predominantly includes government entities undergoing rebranding, restructuring, or dissolution. The Australian Government's Arts Portfolio (`arts.gov.au`) is a prime example, which transitioned from the top harmonic centrality cluster to the second. Originally part of the Department of Communication and Arts in 2018, encompassing national broadcasters and the postal

service, it underwent a reorganisation, subsequently integrating into the Department of Infrastructure, Transport, and Regional Development. Another case is the Bureau of Tourism Research (BTR) at btr.gov.au, which evolved into Tourism Research Australia (TRA) in 2006. Similarly, the Department of Communications and Arts (`dca.gov.au`) underwent a significant renaming and restructuring process. This scenario is typically characteristic of organisations, including government departments experiencing nomenclature changes, reorganisations, or abolition, such as dca.gov.au and government arts.gov.au. Additionally, it encompasses businesses that have ceased operations, like the Australian Motorcycle Museum (`australianmotorcyclemuseum.com.au`), or those undergoing name changes. This category also extends to entities operating in a seasonal or time-bound capacity, such as expos ((`townsvilleexpo.com.au`) or conferences such as (`sydneycubansalsacongress.com.au`), indicating a dynamic environment where organisational identities and structures are fluid and subject to change due to various internal and external factors.

3.5 Discussion and Conclusion

The health of a competitive economy depends on the existence of many independent firms that challenge each other. Traditionally, this has been ensured by two factors: the diminishing returns that limit the growth of monopolies, and the government regulation that prevents or breaks up excessive market power.

However, in the digital age, these factors are no longer sufficient. Network effects and platform strategies enable some firms to dominate entire industries and markets, without necessarily raising consumer prices. Examples of such firms are Uber, AirBnB and Google, which have gained near-monopoly positions in urban transport, accommodation and online advertising, respectively.

The phenomenon of Online Gravity, as described by Paul X. McCarthy in his book (McCarthy, 2015), explains how digital technologies have rebooted economics and created planet-like super-businesses that attract and retain most of the online attention and value. McCarthy also shows how network science can help understand and measure the evolution of diversity and dominance of companies in online activity (McCarthy et al., 2021), and how this can inform policy and strategy decisions. McCarthy's work also illustrates how in new industries online attention revealed through social media graph metrics can be linked to the growth in enterprise value of innovative companies like Tesla (McCarthy et al., 2021).

To preserve economic diversity and consumer welfare, competition regulators need to adopt new approaches and criteria that go beyond price and market share. How to define

and measure economic diversity is one of the challenges.

One influential advocate of this view is Lina Khan, the chair of the US Federal Trade Commission, who argued in her previous academic work (Khan, 2016) that antitrust law should focus on "the long-term interests of consumers [which] include product quality, variety and innovation — factors best promoted through both a robust competitive process and open markets."

Network science provides some useful tools and metrics to analyse the competitive dynamics, growth and influence of firms, especially in emerging sectors. One such metric is harmonic centrality, which measures the potential reach of a node in a network and is a measure of influence, trust and power within a network. In a competition setting, harmonic centrality may have applications in identifying firms that may have the most power and influence over other firms and thus can pose a potential threat to innovation and competition.

Harmonic centrality, a linkage-based metric, encapsulates the comprehensive web context of a domain along with its associated organisation or brand. This study delved into deciphering domain-level and organisation-level dynamics as illuminated by harmonic centrality. Our findings reveal that harmonic centrality is an intricate indicator, amalgamating data pertaining to the web graph structure at the website level, like domain size and inbound link quantity, with offline, non-structural elements such as organisational scale, reach, influence, and impact.

The detection of both structural and non-structural elements within harmonic centrality at a significant scale paves the way for intriguing future research prospects. One such avenue could involve inverting the inquiry to investigate whether harmonic centrality can predict organisational characteristics. Another promising direction might involve contrasting the harmonic centrality of entities within the same industry, correlating it with known metrics like scale, revenue, and industry influence. This approach aims to unravel additional, more subtle attributes potentially embedded within harmonic centrality, such as trustworthiness and authority.

Moreover, the consistency observed in the distributions of harmonic centrality across various nations suggests that our findings possess a degree of universality. This implication indicates that the insights derived from this study are not confined to the Australian web context but could be effectively generalised and applied to broader international settings. Consequently, this expands the applicability of our conclusions, underscoring the potential of harmonic centrality as a versatile tool in understanding and evaluating web-based and organisational phenomena on a global scale. Such research could significantly contribute

to a deeper understanding of the interplay between web presence and organisational attributes, offering valuable perspectives for businesses and policymakers alike in the digital age.

AUTHOR DECLARATION

The following chapter contains content from the following publication.

Gong, Xian, McCarthy, P. X., Griffith, C., McFarland, C., & Rizoiu, M.-A. (2025). Cosmos 1.0: A multidimensional map of the emerging technology frontier. *Scientific Data* ***(accepted)***

Author Contributions: Xian Gong led the data collection, processing, and analysis, designed and implemented the Cosmos 1.0 methodology, constructed the ET23k and ET100 datasets, produced the visualisations, and drafted the initial manuscript. P.X. McCarthy contributed to the conceptual framing of the project, guided the integration of external validation sources, and assisted with manuscript refinement. C. Griffith supported data validation and robustness checks. C. McFarland contributed to the design of case studies. M.-A. Rizoiu supervised the research, guided the development and validation of the methodology.

Production Note:
Signature removed prior to publication.

Xian Gong

Production Note:
Signature removed prior to publication.

Paul X. McCarthy

Production Note:
Signature removed prior to publication.

Colin Griffith

Production Note:
Signature removed prior to publication.

Claire McFarland

Production Note:
Signature removed prior to publication.

Marian-Andrei Rizoiu

CHAPTER 4

COSMOS 1.0: A MULTIDIMENSIONAL MAP OF THE EMERGING TECHNOLOGY FRONTIER

This paper introduces the Cosmos 1.0 dataset and describes a novel methodology for creating and mapping a universe of technologies, adjacent concepts, and entities. We utilise various source data that contain a rich diversity and breadth of contemporary knowledge. The Cosmos 1.0 dataset comprises 23,544 technology-adjacent entities (TA23k) with a hierarchical structure and eight categories of external indices. Each entity is represented by a 100-dimensional contextual embedding vector, which we use to assign it to seven thematic tech-clusters (TC7) and three meta tech-clusters (TC3). We manually verify 100 emerging technologies (ET100). This dataset is enriched with additional indices specifically developed to assess the landscape of emerging technologies, including the Technology Awareness Index, Generality Index, Deeptech, and Age of Tech Index. The dataset incorporates extensive metadata sourced from Wikipedia and linked data from third-party sources such as Crunchbase, Google Books, OpenAlex and Google Scholar, which are used to validate the relevance and accuracy of the constructed indices.

4.1 Background & Summary

Emerging technologies have been shown to deliver the greatest benefits to the economy and society when understood and adopted early, resulting in improved health, environmental outcomes, economic growth, and sustained innovation (Balakrishnan et al., 1979;

Martin, 1995). In the era of the Fourth and even Fifth Industrial Revolution (Industry 4.0 and 5.0), researchers are delving deeply into defining what constitutes emerging technologies (Alvarez-Aros & Bernal-Torres, 2021; Rotolo et al., 2015). These technologies may be entirely new innovations or reimagined applications of existing technologies. Their critical role in national competitiveness (Coccia, 2019; Dahlman, 2007) and a firm's innovative growth (Lee & Grewal, 2004) is increasingly evident. We can see this with the digital transformation of the retail industry. Traditional store-based retailers who adopted the Internet as a communication channel and formed e-alliances experienced a positive impact on firm performance. This is supported by a study reviewing 181 companies across multiple countries, demonstrating the broad impact of digital adoption in the retail sector (Vaishnav & Ray, 2023). Technologies related to artificial intelligence, smart devices, information and communication technologies, new materials, robotics, automation, sensors, and mechatronics, among others, are proving to be pivotal elements in the competitiveness of both developed and emerging economies (Alvarez-Aros & Bernal-Torres, 2021).

Both qualitative and quantitative methods are currently used to identify emerging technologies. Traditionally, the most common way is to use qualitative "top-down" processes with panels of experts discussing and voting on critical technologies and themes. Many leading global organisations use this approach, for example, the Organisation for Economic Co-operation and Development (OECD), World Economic Forum (WEF), and MIT Technology Review each produce annual reports on emerging technologies (Agenda & News, 2023; OECD, 2016; Review, 2023). The best-typified one is the Delphi method, a structured methodology designed to systematically elicit and refine the opinions of experts through iterative rounds of questionnaires and controlled feedback (Brown, 1968). Quantitative methods evolve rapidly with text analysis and deep learning techniques. There are two main limitations in current methods of identifying emerging technologies: the types of data used and the "top-down" methodology applied. At an early stage, most data used to explore emerging technologies started with publications, patent information, News articles or a mix of these (Abercrombie et al., 2012; Zhang et al., 2014). Counts of keywords, authors, publications, citations, patents, or even News articles are commonly used as data resources of quantitative methods (Bettencourt et al., 2008; Ho et al., 2014). The development of data analysis tools and methods, including natural language processing (NLP), subject-action-object (SAO) structure, similarity matrix, knowledge networks, full-text analysis, and large language models (LLMs), has the potential to enable the creation of new types of data for detecting emerging technologies (Bonaccorsi et al., 2020; S. Kim et al., 2020). The vast majority (89%) of data sources on emerging technology forecasting are related to patents, pub-

lications and News articles (Viet et al., 2021). The rapid advancement of text-mining techniques enhances data diversity and offers new, unique insights into both explicit and latent knowledge about the nature, structure, and functions of emerging technologies.

This paper presents a dataset of two broad components. The first comprises entity embeddings of 23k technology-adjacent entities (TA23k) with a hierarchical structure, including an identifiable subset of 100 manually verified, highly recognisable emerging technologies (ET100). The second component is a set of technology indices that can be used to filter both mature and emerging technologies from the universe of technology-adjacent entities. Those indices study the detection and diffusion of technologies ranging from old inventions to the latest advancements in various research and industry fields included in Wikipedia. This dataset aims to help researchers, policymakers, and corporations make informed decisions, allocate resources wisely, and foster the development of these technologies for sustainable growth and competitive advantage. By identifying these emerging technologies early, stakeholders can better prepare for the changes and opportunities they present, encouraging proactive adaptation and strategic planning.

This study uses a "bottom-up" approach to identify and explore the underlying structure of technology-adjacent space by leveraging the Wikipedia corpus and NLP techniques. Wikipedia is edited by numerous experts and contains texts from reliable sources (Jemielniak, 2019; Mesgari et al., 2015). For most technologies, Wikipedia provides relevant description text and links relevant articles using hyperlinks. Wikipedia2Vec, a pre-trained language model with entity embeddings, enables us to filter technology-adjacent articles related to emerging technologies based on cosine similarity (Yamada et al., 2020). Entity embeddings provide context-specific representations of real-world entities (Wikipedia article) compared to the more generalised semantic and syntactic information captured by word embeddings. After defining a universe of technology-adjacent entities from the Wikipedia corpus, dimensionality reduction algorithms and clustering algorithms are used to detect and visualise the clusters and hierarchical structure among the technology-adjacent space. The output of this "bottom-up" approach is a three-level hierarchical tree called three meta tech-clusters (TC3), seven theme tech-clusters (TC7) and ET100 from top to bottom level. The ET100 at the bottom level are the 100 emerging technologies that have been manually verified from the larger technology-adjacent universe. In the meantime, we manually name the three meta tehc-clusters and the seven theme tech-clusters based on the typical characters of the cluster of those technologies (See Subsection Hierarchical Structure of Emerging Technologies for more details). Moreover, we collect and create a series of external indices for the final dataset (See Section Data Records for more details). The workflow

of the Cosmos 1.0 dataset is shown in Figure 4.1.

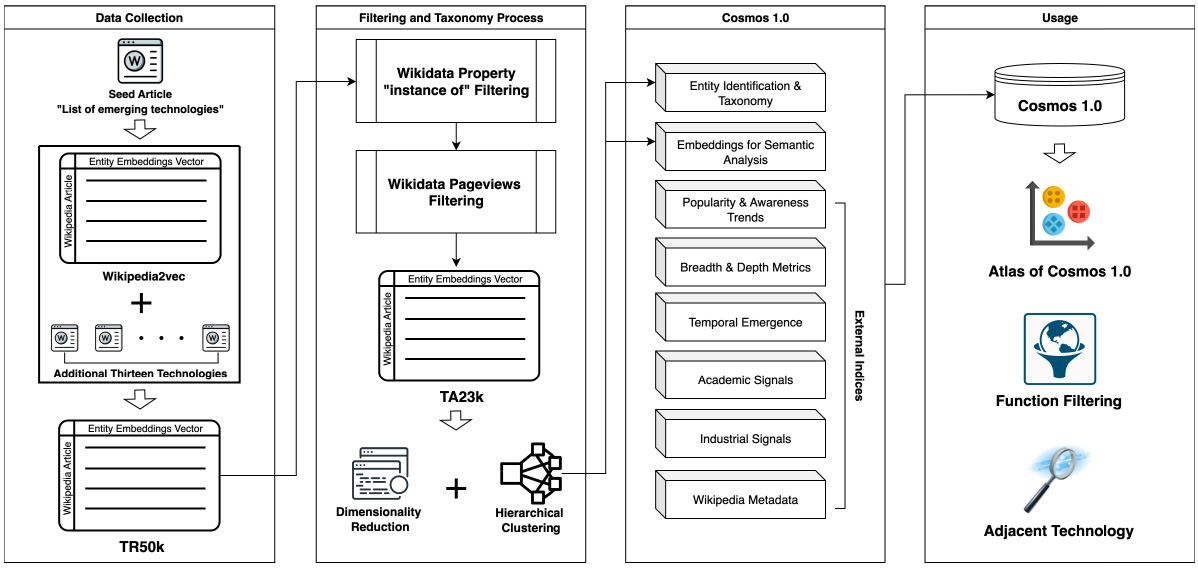


Figure 4.1: The repeatable workflow of creating and mapping the Cosmos 1.0 Dataset. It illustrates the simplified process of data collection, taxonomy, all features and potential usages of the Cosmos 1.0 dataset.

The remainder of this paper is structured into four sections: Methods (Section 4.2), Data Records (Section 4.3), Technical Validation (Section 4.4), and Usage Notes (Section 4.5). Section 4.2 details the data sources, embedding construction using Wikipedia2Vec, filtering procedures, and the development of the hierarchical structure (TC3, TC7, and ET100). Section 4.3 introduces the multidimensional technology indices and describes their conceptual foundations and computational implementation. Section 4.4 presents the data records and feature architecture of Cosmos 1.0, followed by Section 5, which provides technical validation and comparative analysis with third-party datasets and existing emerging-technology lists. This specific structure aligns with Scientific Data's emphasis on transparency, reproducibility, and the creation of high-quality, reusable research datasets.

4.2 Methods

We aim to leverage large language models to automatically identify rising technologies across a broad spectrum of research areas instead of concentrating on just one area or limiting ourselves to technologies that have only recently gained popularity. Current lists of emerging technologies (Gartner, Inc., 2024; McKinsey & Company, 2024; World Economic Forum, 2024) are often limited and lack depth and complex categorisation. We use insights from language models to streamline the exploration of emerging technologies and to

categorise them into clusters with machine-learning methods. We also construct technology indices from reliable third-party sources to quickly and accurately position potential emerging technologies within the technology-adjacent space, validating the approach and strengthening the dataset. These are designed to offer governments, entrepreneurs, and researchers valuable perspectives. Since technologies continuously evolve, we document in detail the processes of collecting data, creating indices, and building Cosmos 1.0. It facilitates future updates to the next generation of the Cosmos dataset.

4.2.1 Data Collection based on Wikipedia2vec

One significant drawback of patents and publications is that the naming of emerging technologies can vary across different publications, depending on the authors' preferences and interpretations (Jaeger et al., 2015). Wikipedia's advantage lies in providing each technology with a consistent and universally recognised name and a detailed and standardised description. Previous efforts have utilised the hyperlink structure of Wikipedia to identify more extensive lists of emerging technologies (Bonaccorsi et al., 2020). However, these approaches overlooked the textual content of Wikipedia and failed to recognise the structure of these technologies or create additional metrics for comparing them.

Wikipedia2Vec (Yamada et al., 2020) is a toolkit that generates embeddings from each article and word within Wikipedia. With this toolkit, Wikipedia articles are trained as entities and thus represent the concepts and linkages within a broad context of 2.6 million other articles. The embeddings for the English language Wikipedia (2018) are used in this research. It is a powerful tool to leverage the vast amount of knowledge encoded in Wikipedia to create high-quality word and entity embeddings. It uses a skip-gram model, a type of neural network architecture that learns to predict the context of a given word based on its embedding. It then applies modifications to the original skip-gram model to incorporate the Wikipedia structure. For example, it uses the categories, links, and redirects in Wikipedia to define the context of a given word or article. It also uses a negative sampling technique to train the model efficiently. Word embeddings represent words in a vector space, while entity embeddings do the same for Wikipedia entities, such as articles for people, places, and concepts. Since the embedding space captures semantic similarities based on Wikipedia's context, co-occurrence, and linkage structure, we retrieve emerging technology-related articles using cosine similarity measurements.

To create the Cosmos 1.0 dataset, we start with the seed article "List of emerging technologies" and collect the 100,000 most similar words and entities based on cosine similarity to the seed. The whole list contains 45,149 words and 54,851 entities. Since entity embed-

dings offer the advantage of capturing entities' specific contexts and relationships more accurately than word embeddings, we only keep the entities as the universe of emerging technology "candidates". Moreover, we incorporated an additional thirteen technologies found in Wikipedia that were not initially included in our set. These technologies are: "Advanced Computer Techniques", "Biopharmaceutical", "Communication", "Computer security software", "Customer relationship management", "Educational technology", "Electronic data interchange", "Enterprise resource planning", "Financial technology", "Genetics", "Microsatellite", "Online service provider", "Research". Those technologies align with other emerging technology frameworks, including World Economic Forum, OECD, and MIT Technology Review.

We refer to the complete entity list as TR50k (54,864 technology-related entities). It includes technologies, theories, organisations, people and other concepts, which are technology-related but not directly close to technologies themselves. These articles represent "noise" for our purposes, and we remove and minimise their impact using properties of Wikidata and pageviews. Firstly, we utilise the Wikidata property "instance of" to refine and clean this data. The "instance of" attribute connects an item with its broader category, denoting it as a particular instance within that category. Since we aim to remove those apparent anomalies and retain as many entities as possible, we manually select technology-adjacent instance labels and keep all Wikipedia articles without "instance of" information. Although this less rigorous screening may preserve a degree of noise, we consider this approach justified as it maximises the retention of potential emerging technology "candidates" within our dataset. There are 29,030 entities left after we filter the data based on critical labels of this attribute. The kept instance labels are available in supplementary information (Section 1 Supplementary Methods). Secondly, we collected pageviews of the TR50k for the last three years (2021 - 2023). We noted many articles with no or very few pageviews as being tangentially related to technologies rather than being technologies themselves. We assume that valid technologies, even specialised ones, attract some degree of general attention and thus are not among the bottom 10% percentile of the whole pageviews distribution. We note that the articles with the least attention, that is, those in the bottom 10% percentile on the distribution tails for the last three years are those with less than 2085, 2039 and 1766 pageviews, respectively. To be more restrictive again, we manually raised the pageviews filter value to 2500 to filter the noise. In other words, we only retain entities that have received more than 2500 pageviews per year over the past three years. Finally, we end up with 23,544 technology-adjacent entities called the TA23k, which consist of the universe of emerging technology "candidates".

Applying the 2,500 pageviews threshold reshapes the log-transformed distribution of Wikipedia pageviews from approximately normal to positively skewed. This indicates that a large portion of ambiguous, non-technical entries have been effectively filtered out. The resulting skewed distribution aligns with the typical long-tailed patterns found in bibliometric and innovation datasets, where a small number of technologies attract widespread attention while the majority remain niche (Clauset et al., 2009). Kostoff et al. (Kostoff et al., 2004) are concerned that disruptive technologies are challenging to detect early because they often emerge in niche domains, outside mainstream research, and may be overlooked by conventional and strict metrics. To validate that the threshold does not exclude meaningful technologies, we compared our filtered list against established foresight sources. Over 90% of technologies identified by the OECD fall above the 2,500 pageviews threshold (See Subsection Comparison ET100 with other emerging technology lists for more details). Together, distributional evidence and benchmarking results provide converging support for the 2,500 pageviews threshold as an empirically grounded method for curating emerging technologies, while acknowledging that additional methods (e.g., expert review) may be required to capture early-stage innovations with limited public visibility.

Since we frequently mention technology-related entities, technology-adjacent entities, and emerging technologies in this paper, we distinguish these terms to avoid ambiguity. Technology-related entities, exemplified by TR50k, constitute the broadest set of technologies identified on Wikipedia through cosine similarity, though this set contains considerable "noise". Within our framework, we define technology-related entities as those positioned furthest from emerging technologies. Following refinement of TR50k using Wikidata and pageviews, we derived a more focused subset, TA23k, representing technology-adjacent entities that are closer to emerging technologies, yet still affected by "residual noise" due to the limitations of our filtering process. By contrast, the 100 emerging technologies (ET100) identified in this paper through a combination of tailored indices and manual screening represent the set of genuine emerging technologies (See Subsection Hierarchical Structure of Emerging Technologies for more details). Next, we apply machine learning algorithms to the TA23k dataset to detect the hierarchical structure of this technology-adjacent space. This provides the initial input of approximately 23k technology-adjacent entities with 100-dimensional embedding vectors. We now have a valuable approach for identifying potential emerging technologies "candidates" within the space and exploring the relationships between different technologies and research fields.

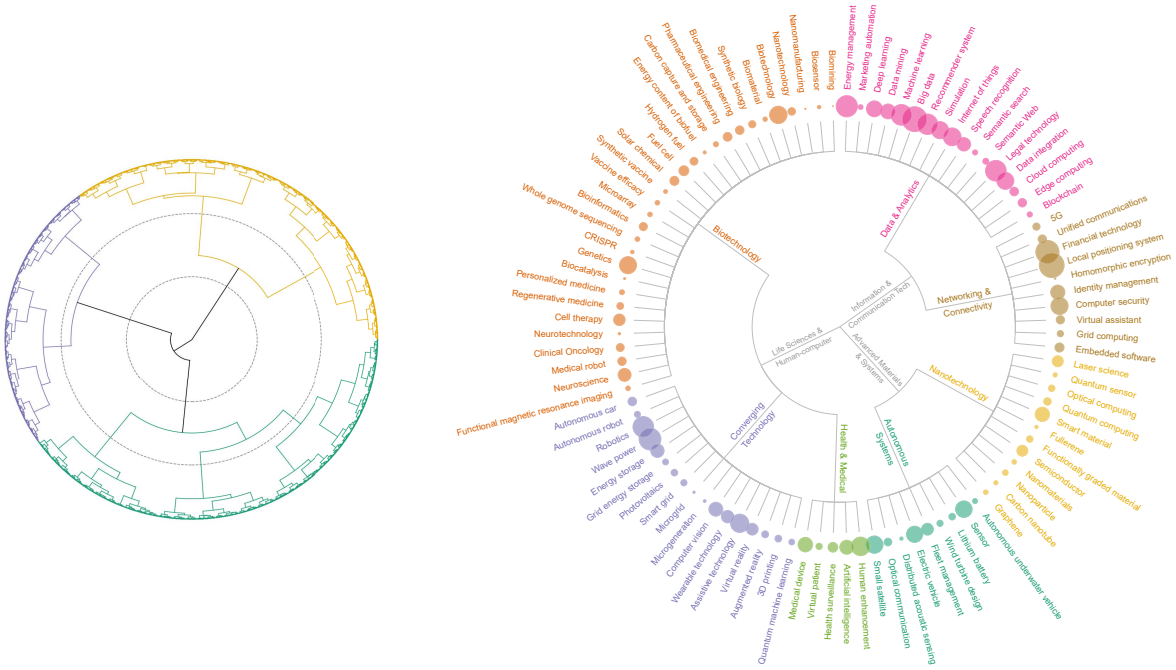
4.2.2 Hierarchical Structure of Emerging Technologies

Since no globally standardised classification system exists to classify technologies, we introduce unsupervised machine learning algorithms to cluster the TA23k based on their embedding vectors. Before clustering, we use the t-distributed Stochastic Neighbour Embedding (t-SNE) algorithm to map the 100-dimensional entity embeddings space to a two-dimensional space for visualisation and further clustering analysis. The t-SNE (Van der Maaten & Hinton, 2008) algorithm excels at preserving local structures within high-dimensional data, making it particularly well-suited for revealing intricate cluster patterns that linear methods, such as PCA, might miss. Its non-linear approach allows it to capture complex relationships between features, offering a more nuanced view of data's inherent structure. This preprocessing step mitigates the curse of dimensionality by simplifying the data, which enables visualising and understanding complex data structures.

The second step is to utilise agglomerative hierarchical clustering (AHC) (Gowda & Krishna, 1978) to reveal the hierarchical structure of the TA23k. It is a bottom-up clustering method where each data point starts as its cluster, and pairs of clusters are merged as one moves up the hierarchy. The process begins with calculating the similarity (or distance) between each pair of points or clusters. It then iteratively combines the closest pairs into larger clusters until all points are merged into a single cluster or a desired number of clusters is reached. This method is known for producing a dendrogram. This tree diagram illustrates the series of merges and the multi-level hierarchy of clusters, enabling detailed analysis of data grouping and structure.

As Figure 4.2(a) shows, each branch represents a cluster, and the height of the branches reflects the dissimilarity between merging clusters. The dendrogram reveals two levels of optimal clustering: one at seven theme tech-clusters (TC7) and another at three meta tech-clusters (TC3). At the TC7 level, the dendrogram displays smaller, more finely grained clusters, indicating greater similarity within these groups. This level is suitable for detailed analyses with significant nuances between data points. A more pronounced gap is observed before merging into three clusters as we move higher up the dendrogram. This considerable increase in dissimilarity suggests a natural consolidation of the data into three broader categories, each representing a more general grouping of the data points. These broader clusters are ideal for high-level insights and understanding overarching patterns within the dataset.

Within the framework of the TC7, a curated compilation of 100 emerging technologies (ET100) was manually reviewed. For each thematic tech-cluster, we first ranked a broad pool of candidate technologies using the **Google_Patent_Counts_2023** feature (See Sec-



(a) The Cosmos 1.0 (n=23,544) **(b)** ET100 (n=100)

Figure 4.2: Radial Tree Dendorgam of Cosmos 1.0 and ET100. (a) The full radial tree dendrogram with two grey dotted circles demonstrates three and seven are optimal numbers of clusters for the TA23k. We use colours to distinguish the three meta tech-clusters (TC3). (b) The curated radial tree dendrogram shows the technologies of ET100 and cluster names of TC7 and TC3 from bottom to top. The circle sizes of ET100 represent the normalised value of a technology index called **Generality_Index**. The theme tech-cluster "Data & Analytics" is frequently mentioned across Wikipedia articles than other theme tech-clusters.

tion Data Records for more details), which captures the number of patents filed in 2023 on Google Patents that explicitly mention the technology. This patent-based signal served as a proxy for recent technological activity and intensity of innovation. Our experts manually assessed the top 1,000 technologies per cluster, using qualitative criteria such as investment potential, public and industry visibility, and relevance to broader sociotechnical trends. From this refined pool, 91 technologies were selected in the seven theme tech-clusters. Moreover, we refer to other prominent technology listings by leading global organisations such as the OECD, WEF, and Gartner, resulting in a final list of 100 emerging technologies. We ensure that each identified technology aligns precisely and unambiguously with one of the seven thematic groups with manual inspection, facilitating thematic cluster-level analyses. Figure 4.2(b) provides a dendrogram that illustrates the hierarchical relationship and integration among ET100, TC7, and TC3. This structured approach enables the exploration of various levels and relationships within the hierarchy, including the

dynamics between parent and child entities.

In addition to the hierarchical structure, the ET100 gathering sequence is also worth investigating. Close-knit technologies are more likely to be grouped in the same cluster. The spectrum of technologies ranges from foundational ones, such as cloud computing and big data, to specialised innovations like CRISPR and quantum computing. Foundational technologies provide the base infrastructure, while data and analytics technologies, including machine learning and data mining, build on this foundation to process vast data volumes. Specialised industrial technologies, such as robotics and 3D printing, transform specific sectors, while cutting-edge technologies like graphene represent the forefront of scientific research. Energy and environmental technologies focus on sustainability, while health and biotechnologies, such as regenerative medicine, are critical for medical advancements. This progression reflects a move from broad, widely adopted technologies to experimental and transformative innovations at the cutting edge of technological progress. Each technology contributes to an interconnected ecosystem of advancements, pushing the boundaries of the digital and physical worlds.

The data-centric approach utilised in developing the Cosmos 1.0 stands out for its robustness and verifiability. It facilitates the analysis of connections between technologies and their relatedness. Unlike traditional methodologies that rely on manual, high-level analyses or one-off in-depth studies, our method offers comprehensive mapping capabilities compared to other emerging technology lists of OECD and the WEF, ensuring complete and precise alignment. It is designed to allow for the future inclusion or exclusion of technologies as the landscape evolves. Due to its hierarchical structure and extensive technology corpus, it also supports more detailed explorations within specific areas. This methodology ensures that both large-scale and niche technologies are included, spanning the breadth of the landscape. The utility and relevance of Cosmos 1.0 are demonstrated through the application of external technology indices, as outlined in the next section.

4.2.3 Technology Indices

We create multiple metrics, including the Technology Awareness Index, Generality Index, Deeptech Index, Age of Tech Index and Technology Proximity Index, to capture each technology's growth, popularity, research depth and emergence in our TA23k, respectively. The "Technology Awareness Index" measures the evolving public and academic interest in emerging technologies by analysing their Wikipedia pageviews trends over time. The "Generality Index" evaluates emerging technologies' broad applicability and foundational importance by analysing their prevalence across diverse contexts within Wikipedia, providing insights

into their potential as general-purpose innovations. The "Deeptech Index" leverages advanced data analytics to quantify technologies' scientific depth and innovation potential, providing a crucial tool for identifying and assessing Deeptech activities within the broader technology landscape. The "Age of Tech Index" utilises historical literature data from the extensive Google Books collection to determine the birth year of technologies, providing a foundational metric for analysing their emergence, societal impact, and evolution over time.

Technology Awareness Index

The use of Wikipedia pageviews to measure the growth of interest in emerging technologies is based on the assumption that increased public and professional curiosity about new technologies translates to more visits to their respective Wikipedia articles. By tracking the pageviews statistics over time, researchers can potentially gauge the rising or waning interest in specific technologies. This method offers a proxy measure of popularity or awareness rather than direct technological adoption or development. It leverages the accessibility and extensive use of Wikipedia as a primary source of information, making it a useful, albeit indirect, tool for analysing trends in technological engagement and public interest.

To calculate the "Technology Awareness Index", we use linear regression on the pageviews data from Wikipedia articles over the past three years. By plotting the yearly pageviews against time and fitting a linear regression model, we derive the slope of this line, which represents the rate of change in pageviews. A positive slope indicates an increase in interest, reflected by more pageviews over time, whereas a negative slope suggests declining interest. This slope, quantified as the index, provides a numerical value that helps gauge the growth or decline in attention toward specific technologies, offering insights into trends in technological engagement.

Generality Index

The aim of the "Generality Index" is to quantify and understand emerging technologies' versatility and foundational nature. It quantifies the breadth of application of a technology by analysing its presence across different contexts within Wikipedia's vast knowledge repository. This index uses computational linguistics and document frequency techniques to measure how often technology is mentioned across various Wikipedia articles.

To construct the index, each technology is represented by a canonical search phrase (e.g., "artificial intelligence", "quantum computing"), which is queried using the Wikipedia API (https://en.wikipedia.org/w/api.php). The API returns a value indicating the number

of unique Wikipedia articles in which the phrase appears (totalhits). This raw count serves as the Generality Score, a proxy for the document frequency (DF) of the term across Wikipedia.

A higher Generality Score indicates that technology is mentioned in broader contexts, suggesting its versatility and importance as a general-purpose technology (Bekar et al., 2018; Bresnahan & Trajtenberg, 1995) that serves as a foundation for innovation in multiple industries and markets. In contrast, a lower score indicates a more specialised technology with limited applications. This index supports the identification of critical generative technologies that serve as essential building blocks across many industries and sectors. Such technologies are pivotal in driving continuous innovation and have a broad impact due to their wide applicability.

The Generality Index helps researchers, policymakers, and industry leaders recognise technologies likely to be more influential and essential for future developments. Essentially, the Generality Index offers a systematic approach to gauge technologies' universality and potential ubiquity, aiding in strategic decision-making on research focus, funding allocation, and policy formulation. Limitations of this measure include coverage bias, temporal lag, and focus on popularity rather than impact. It can also be influenced by language and cultural differences and does not account for the depth or context of technology discussions, suggesting the need for third-party data for a verified assessment.

Deeptech Index

The "Deeptech Index" is a metric designed to evaluate technologies based on their depth of scientific research and potential for disruptive innovation. Deeptech technologies are characterised by their strong foundations in science-based research and development (R&D), often protected by intellectual property. They are distinct from technologies that primarily enhance business or service models. The index aims to distinguish these technologically and scientifically intensive technologies from those less anchored in rigorous R&D.

We use an advanced methodology blending business and research dimensions to compute the Deeptech Index. This approach utilises the Wikipedia2Vec model, representing Wikipedia articles as embedding vectors that can be quantitatively analysed. For this index, specific Wikipedia articles related to "Business" and "Industry" serve as seeds to capture the business orientation of technology. In contrast, articles like "Basic research", "Research and development", and "Research" represent the research dimension. The cosine similarity between the embedding vectors of each technology and these seed vectors is calculated to quantify how closely a technology aligns with business or research themes.

The index calculation involves first determining the average cosine similarity of a technology to the business seeds and the research seeds. The business component is then inverted (1 - "business") to prioritise less business-oriented technologies and more grounded in deep research. The sum of the research-oriented score and the inverted business score is computed for each technology. Finally, these summed scores are ranked on a percentile scale to create the Deeptech Index, offering a relative measure of each technology's position.

This index provides a scalable and robust method for assessing Deeptech activity, overcoming limitations found in traditional datasets like patents or research investment records, which can be incomplete or outdated. By leveraging real-time and widely available Wikipedia data, the Deeptech Index offers a timely and insightful tool for analysing at scale the landscape of technologies foundational to future innovations.

Age of Tech Index

Determining a technology's "age" is a methodological approach aimed at understanding when it becomes significant within society rather than simply tracking its first mention. This approach is encapsulated in the "Age of Tech Index", which uses a data-driven method to establish a technology's birth year based on its presence in literature, specifically leveraging the expansive Google Books database.

Google Books, a large repository of more than 40 million books in more than 400 languages, is an ideal source for this analysis. It includes collections from major research libraries like the University of Oxford Bodleian Libraries, Harvard University Library, and Stanford University Libraries. This makes it reflective of a broad societal understanding of concepts over time. This vast collection tracks the mentions of technologies from as early as 1500 up to 2019, providing a long-term perspective on technological evolution.

To determine the birth year of technology according to the Age Index, the analysis focuses on identifying the point in time when mentions of the technology reach a certain threshold, specifically, 5% of its maximum mentions during 1900-2019. This threshold is considered significant because it indicates a notable level of recognition and integration of technology into societal or academic discourse (Kousha et al., 2011). The data is analysed year by year, starting from 1900, aligning with the period of rapid technological advancement and better documentation.

This methodology is crucial for understanding technologies' emergence, growth, and relationship to other technologies. For example, it allows for examining how technologies like CRISPR and blockchain have rapidly reached broad audiences, contrasting with older

technologies like neuroscience and renewable energy, which have built momentum over more extended periods. Metrics like audience reach on digital platforms provide insights into the rapid adoption of specific applications, but the Age of Tech Index offers a broader, more historical perspective. It analyses mentions across a wide range of literature, reflecting broader societal engagement with and the significance of various technologies.

Technology Proximity Index

The introduction of the "Technology Proximity Index" aims to expedite the process of distinguishing between technologies and concepts or terms related to technology. Technology is defined as consisting of three fundamental aspects-purpose, function, and benefit-representing its timeless essence (Carroll, 2017). Within the universe of Cosmos 1.0, certain companies, research institutions, concepts, and terms are highly associated with technology but do not yet meet this definition. These entities are retained to ensure the breadth and diversity of the universe. Therefore, a classifier has been trained to measure the proximity of each Wikipedia entity in Cosmos 1.0 to the concept of technology as defined.

We utilise ChatGPT with manual screening to assist in selecting entities as targets for the classifier. Specifically, we identified 100 entities from each of the seven theme tech-clusters (TC7) within Cosmos 1.0, resulting in 700 positive targets. To create a balanced input dataset, we subsequently selected 700 negative targets from the entities that were removed during the filtering process (See Subsection Data Collection based on Wikipedia2Vec for more details). An XGBoost classifier is trained on the input data with hyperparameter tuning, resulting in an average accuracy and F1 score of approximately 86% on the test sets.

The classifier assigns a probability score to each entity, representing the likelihood and confidence that it aligns with the "technology." We designate this probability score as **Tech_Proximity_Prob**. By applying a decision threshold of 0.5, these probability scores are converted into discrete class labels, termed **Tech_Proximity_Index**. This process enables us to distinguish entities that closely correspond to the "technology".

4.3 Data Records

The Cosmos 1.0 dataset is openly accessible at Figshare (Xian et al., 2025). The file Cosmos_Dataset.xlsx is a table, with each row corresponding to a technology-adjacent entity with a corresponding Wikipedia page. Each row includes the following variable fields. We categorise all features into eight meaningful categories and explicitly explain their importance, providing simple examples.

4.3.1 Data Structure

Entity Identification & Taxonomy

These features contain identity and location in the hierarchical structure, which enables users to explore the position of a technology in the broader technology-adjacent space.

- **Wiki_Entity**: a specific concept or item with a dedicated Wikipedia page.
- **TC3 meta tech-clusters**: Cluster labels for the three meta tech-clusters cut the dendrogram at a higher level, resulting in more generalised groups of the data points.
- **TC7 theme tech-clusters**: Cluster labels for the seven theme tech-clusters cut the dendrogram at a lower level, yielding finer-detailed clusters that reflect closer similarities among the data points.
- **ET100_Flag**: Flag marks the 100 technologies manually selected.
- **tsne_x**: The x-axis of the 2D map in Figure 4.4, also the first dimension of t-SNE on TA23k.
- **tsne_y**: The y-axis of the 2D map in Figure 4.4, also the second dimension of t-SNE on TA23k.

Embeddings for Semantic Analysis

The feature captures the semantic context of each technology-adjacent entity, which enables clustering, similarity search, transfer learning, or visualisation in latent space.

- **Feature_1-100**: Entity embedding vector representations of Wikipedia article (100 dimensions).

Popularity & Awareness Trends

This category captures public interest by tracking Wikipedia pageviews over time, enabling the identification of technologies that are rapidly gaining or losing visibility.

- **Pageviews_2019-2023**: The annual sum of pageviews of Wikipedia technology-adjacent entity pages for the last five years (2019-2023).
- **Pageviews_3yr_slope**: The regression slope on the annual pageviews counts of technology-adjacent entities for the last three years (2021-2023).

- **3yr_Awareness_Index**: Standardised **Pageviews_3yr_slope** to compare how awareness of the technology-adjacent entity is growing or declining relative to others.
- **Pageviews_5yr_slope**: The regression slope on the annual pageviews counts of technology-adjacent entities for the last five years (2019-2023).
- **5yr_Awareness_Index**: Standardised **Pageviews_5yr_slope** to compare how awareness of the technology-adjacent entity is growing or declining relative to others.

Breadth & Depth Metrics

This category helps distinguish general-purpose technologies, deeptech innovations, and core technological concepts from adjacent or peripheral entities.

- **Generality_Index**: Document frequency of a technology-adjacent entity mentioned across various Wikipedia articles.
- **DeepTech_Index**: A measure of assessing the deeptech activity of each technology-adjacent entity.
- **Tech_Proximity_Prob**: Evaluate the proximity of **Wiki_Entity** as the core technology.
- **Tech_Proximity_Index**: Flag marks **Wiki_Entity** as a core technology if its **Tech_Proximity_Prob** is larger than 50%.

Temporal Emergence

This category helps distinguish recently emerging technologies from more established ones, supporting analyses of technology life cycles and diffusion timelines.

- **Age_of_Tech_Index**: Predictive age of each technology-adjacent entity.
- **First_Pub_Year**: The year of the first publication that mentions the technology-adjacent entity in the title, as recorded in the OpenAlex database.

Academic Signals

This category indicates how actively a technology-adjacent entity is being studied, helping to identify research frontiers and emerging academic focus areas.

- **Publication_Counts_2019-2023**: The annual counts of publications that mention the technology-adjacent entity in the title for the last five years (2019-2023) from OpenAlex.
- **Publication_Counts_3yr_slope**: The regression slope on the annual publication counts of technology-adjacent entities in OpenAlex for the last three years (2021-2023).
- **3yr_Publications_Growth_Index**: Standardised **Publication_Counts_3yr_slope**.
- **Publication_Counts_5yr_slope**: The regression slope on the annual publication counts of technology-adjacent entities in OpenAlex for the last five years (2019-2023).
- **5yr_Publications_Growth_Index**: Standardised **Publication_Counts_5yr_slope**.
- **GS_Author_Counts**: The number of scholars, as collected from Google Scholar, who have shown interest in or are actively working in a specific technology-adjacent entity area.
- **GS_Author_Counts_Scaled**: Scale **GS_Author_Counts** from the smallest to the largest count, making it easier to compare across different technology-adjacent entities.

Industrial Signals

This category helps identify technologies with growing industrial investment and innovation potential, signaling their relevance to markets and applied R&D.

- **TFR**: The sum of the cumulative amount of money companies that mention specific technology-adjacent entity in their description have raised across all their funding rounds.
- **Google_Patent_Counts_2019-2023**: The annual counts of patents that mention the technology-adjacent entity for the last five years (2019-2023) from Google Patents.
- **Google_Patent_Counts_3yr_slope**: The regression slope on the annual Google Patent counts of technology-adjacent entities for the last three years (2021-2023).
- **3yr_Google_Patent_Growth_Index**: Standardised **Google_Patent_Counts_3yr_slope**.
- **Google_Patent_Counts_5yr_slope**: The regression slope on the annual Google Patent counts of technology-adjacent entities for the last five years (2019-2023).
- **5yr_Google_Patent_Growth_Index**: Standardised **Google_Patent_Counts_5yr_slope**.

Wikipedia Metadata

This category supports qualitative analysis, global relevance assessment, and content-based filtering for downstream applications.

- **Wikipedia_Content**: The full text of a Wikipedia page, excluding images and tables.
- **Wikipedia_Summary**: A concise overview of the main points of a Wikipedia page's content.
- **Wikipedia_External_Links**: The counts of links on a Wikipedia page connect to other related Wikipedia articles, external sites, and sources that provide additional information or verify the content discussed.
- **Wikipedia_Num_Ref_Links**: The count of reference links used to substantiate the information presented in the article.
- **Wikipedia_Coordinates**: The geographical coordinates of locations mentioned in articles, linking to interactive maps for easy visualisation and navigation.
- **Wikipedia_Num_Language**: Each technology's number of Wikipedia language editions in 2024.

4.3.2 Multidimensional Framework of the Cosmos 1.0

Identifying emerging technologies requires multidimensional assessments rather than a single evaluation. We introduce eight categories to evaluate all technology-adjacent entities, providing a holistic framework for assessing emerging technologies within the technology-adjacent space by integrating structural, semantic, temporal, academic, industrial, and public interest dimensions.

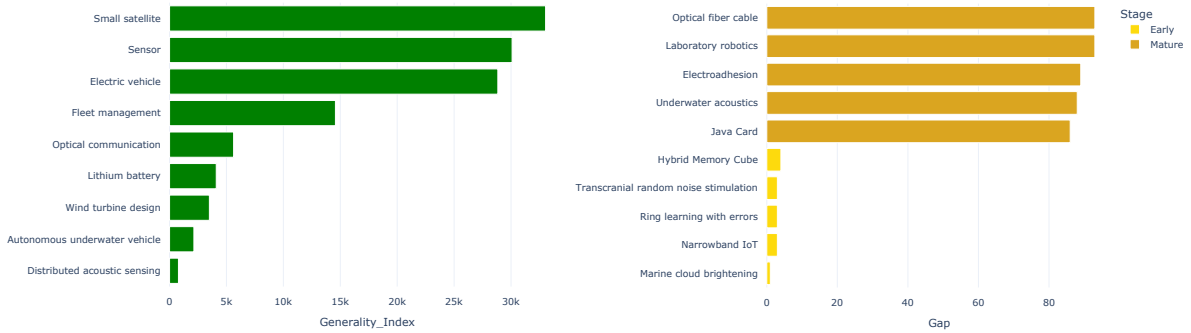
Organising technology-adjacent entities into a hierarchical structure (TC3, TC7, ET100) provides a foundational taxonomy for understanding the emerging technology landscape (in Figure 4.2(b)). Such a taxonomy helps users navigate complex domains and identify clusters of interested technologies. Katy Börner (Börner, 2010) highlights the importance of structured knowledge systems and hierarchical maps for facilitating foresight and exploration across scientific domains. This multilevel design enables both broad overviews and granular analysis of technology domains on different scales. It also allows users to quickly focus on specific domains of interest, making the dataset highly adaptable to diverse needs.

Moreover, semantic embeddings derived from Wikipedia2Vec capture rich contextual relationships among technology-adjacent entities that traditional keyword-based methods often miss (Mikolov et al., 2013; Yamada et al., 2020). By applying cosine similarity to rank technology-adjacent entities relative to a curated set of seeds, one can identify overlooked but semantically proximate technologies within the same cluster. It enhances the discovery of specific emerging areas within broad fields, such as "Renewable Energy Technologies" (See subsection Intra-Cluster Similarity for more details).

Prior studies (Mestyán et al., 2013; Roll et al., 2016) have demonstrated the predictive and interpretive power of Wikipedia pageviews, which can serve as a meaningful indicator of emerging attention in our dataset. Fortunato et al. (Fortunato et al., 2018) emphasise the importance of tracking temporal dynamics in scholarly and public engagement to map the evolution of scientific and technological knowledge. Accordingly, these features quantify both the volume and growth of public attention toward each technology-adjacent entity, capturing not only how widely known a technology-adjacent entity is but also how rapidly its visibility is changing. By standardising growth rates, they are helpful for comparative assessment of emerging momentum relative to other technology-adjacent entities (See subsection Multi-Indices Filtering Strategy for more details).

Features in the Breadth & Depth Metrics category are largely independent and exhibit low intercorrelation. Each one offers a distinct dimension and perspective on the technology universe, revealing its generality, scientific depth, and technical proximity, and is critical to assessing the potential impact and long-term development of a technology-adjacent entity. Emerging technologies often exhibit both specialisation and diffusion: some originate in narrow domains and evolve into general-purpose platforms (e.g., "Small satellite" and "Sensor"), while others gain traction through deep scientific advancement (A. Arora et al., 2018; Chiarello et al., 2018; Gourevitch et al., 2021; Jovanovic & Rousseau, 2005). In Figure 4.3(a), technologies with high generality indices, such as "Small satellite" and "Electric vehicle", demonstrate widespread applicability across diverse domains including aerospace, environmental monitoring, and consumer electronics. In contrast, "Distributed acoustic sensing" and "Autonomous underwater vehicle" are more specialised use cases and narrower application scopes. Traditional indicators, such as publication or patent counts, often overlook these subtleties. By integrating generality, deeptech intensity, and technical closeness, these indices offer a multidimensional view of technology-adjacent entities, enabling more accurate identification of technologies and more precise differentiation between fleeting trends and fundamental breakthroughs.

The Temporal Emergence category is crucial for identifying the stage of technologies in



(a) ET100 in "Autonomous Systems" **(b)** Stage of Technologies

Figure 4.3: Example usage of features of Cosmos 1.0 (a) Technologies of ET100 in the theme tech-cluster "Autonomous Systems" are ranked by feature **Generality_Index**. (b) Early-stage and mature-stage technologies are filtered by the difference between features **Age_of_Tech_Index** and **First_Pub_Year**.

the innovation life cycle. Tracking temporal indicators of innovation is a well-established approach in technology forecasting and bibliometrics (Callon et al., 1983; A. L. Porter & Cunningham, 2004; Small & Upham, 2009). The **First_Pub_Year** feature captures the earliest appearance of a technology-adjacent entity in scientific literature. The **Age_of_Tech_Index** estimates the onset of societal or academic recognition by identifying the first year in which the technology-adjacent entity reached 5% of its historical peak mentions. The combination of these two features allows for distinguishing between technology-adjacent entities that have emerged recently and those that have existed longer but only recently gained traction. For example, a technology with an early publication date but a recent age index may indicate delayed adoption or a resurgence of interest. In contrast, a close alignment between the two features suggests steady and early uptake, thereby helping to classify technologies as early-stage or mature with greater precision. We focus on technologies where the **Age_of_Tech_Index** exceeds the **First_Pub_Year**, indicating a delay between initial publication and broader recognition. To ensure relevance, we retain only those with an **Age_of_Tech_Index** after 2010 and a **First_Pub_Year** after 1920. The size of this gap ranks the technology-adjacent entities: a larger gap suggests a mature technology-adjacent entity with delayed adoption, while a smaller gap signals a recently emerging technology-adjacent entity. From the top and bottom 60 entities, five technologies are manually selected as examples for each group shown in Figure 4.3(b).

Features in the Academic Signals quantify the intensity and growth of scholarly activity, capturing how actively it is currently being studied and how rapidly that attention is growing. Increasing publication counts and growth slopes indicate an increasing interest in research (S. K. Arora et al., 2013; Cozzens et al., 2010), while standardised indices allow fair

comparison among domains with different baseline activity levels. The **GS_Author_Counts** feature adds another dimension by measuring the size of the engaged research community. A larger community often indicates that the technology has attracted broad, interdisciplinary attention, increasing its potential for diffusion (A. Porter & Rafols, 2009). It also suggests a stronger foundation for future innovation, collaboration, and funding, making such technologies more likely to transition from emerging research areas to impactful applications.

The Industrial Signals category helps identify technologies with commercial attraction by capturing signs of capital investment and patent data. Capital investment and patent data serve as practical indicators for identifying technologies with strong commercial potential and strategic value (Caviggioli, 2016; Ernst, 2003; Ghosh & Nanda, 2010). For example, a high **TFR** indicates that startups referencing a technology have successfully attracted significant investment, reflecting market confidence and anticipated demand. Meanwhile, steep patent growth slopes represent intensifying innovative activity and protection efforts, which often precede product development and commercialisation. When these signals align, they help screen potential technologies that are strategically positioned for industry adoption and policy support. Therefore, this category reflects the market closeness, investment appeal and technological maturity.

The Wikipedia Metadata category provides rich textual and structural signals that enable qualitative evaluation, content filtering, and assessment of global relevance for technologies. Wikipedia content and link structures reflect collective intelligence and serve as proxies for public understanding, topical coverage, and notability (Holloway et al., 2007; Yasseri et al., 2012). Features such as full article content and summaries support semantic analysis and contextual enrichment, while external and reference links serve as indicators of information quality and interconnectedness, similar to how citation networks are used in scientometrics. Compared to traditional bibliographic databases, these Wikipedia-based features offer open access and real-time insights across a broader spectrum of knowledge domains. They are particularly useful for downstream tasks such as filtering ambiguous entries and identifying technologies with international presence.

4.4 Technical Validation

4.4.1 Comparison indices with third-party data sources

Correlation analysis is used to validate the indices in our dataset. Employing three different correlation tests, including Pearson, Spearman, and Kendall Tau, provides a robust analysis by measuring relationships from various perspectives, accommodating different data characteristics like non-normal distributions and outliers, and confirming consistency across tests for more reliable results. This approach offers a statistical method to measure and establish the strength and direction of relationships between the dataset's indices and external, reputable data sources such as publication counts from OpenAlex, total funds raised from Crunchbase and patent records from Google Patents. It provides empirical evidence of validity, showing that these indices reflect real-world trends and activities surrounding technologies. By using correlation analysis in this context, we quantitatively assess the reliability of these indices as accurate indicators of technological significance and impact, thereby supporting the robustness of the dataset's construction.

There are also limitations to this approach. Correlation does not imply causation; it merely indicates patterns of association between two variables without establishing a direct cause-and-effect relationship. Additionally, the validity of the results depends heavily on the quality and relevance of the third-party data used. Since the third-party data are unavailable for all 23,544 technologies, the validation process may not comprehensively cover the entire dataset, potentially leading to biases or gaps in the validation efforts. However, all third-party data used had significant overlap in entities with the whole technology-adjacent space (TA23k), with all having more than two thousand data points in common.

Table 4.1 summarises the correlations between the four constructed indices and the third-party data. To justify the Awareness Index, the growth in public interest in technologies based on Wikipedia pageviews, we evaluate the growth in academic interest in technologies based on publication counts from OpenAlex. We use the same method and compare these two trends through correlation analysis. For the trend of the last three years, both Pearson and non-parametric tests (Spearman and Kendall Tau) indicate a statistically significant positive correlation, supporting the reliability of the Awareness Index as an indicator of growing interest in these technologies over a shorter timeframe. It statistically confirms that increases in Wikipedia pageviews are associated with increases in scholarly publications. The significance of these correlations in the shorter term suggests that the Awareness Index effectively reflects real-time shifts and trends, making it a reliable tool for gauging immediate academic and public interest in new technologies. However, for the 5-

Table 4.1: Technology Indices and their Correlations with Third-Party Data Sources. The statistical tests validate the accuracy, utility, and significance of the internal Cosmos dataset by demonstrating that the internal structure of the Cosmos technology model aligns with research metrics, patent counts, and venture capital investments.

Technology Indices (Constructed)	Third-Party Data Sources	Sample Size	Pearson	Spearman	Kendall Tau
Awareness Index (3yr)	Publication Counts Trend (3yr)	19,080	0.0146**	0.1045***	0.0725***
Awareness Index (5yr)	Publication Counts Trend (5yr)	19,078	0.0071	0.1379***	0.0956***
Generality Index	Patent Counts 2023	12,938	0.2043***	0.2666***	0.1807***
Generality Index	Total Funds Raised	4,101	0.2047***	0.0256	0.0170
Deeptech Index	Wikipedia Reference Links	22,109	0.1305***	0.1794***	0.1212***
Deeptech Index	Scholar Counts	2,272	0.0439**	0.0712***	0.0475***
Age of Tech Index	First Publication Year	15,614	0.2897**	0.4152***	0.2942***

** Significant at the 0.05 level; *** Significant at the 0.01 level.

Third Party Data Sources: Publication Counts (Microsoft Academic / OpenAlex); Patent Counts (Google Patents); Total Funds Raised (Crunchbase); Reference Links (Wikipedia); Scholar Counts (Google Scholar / League of Scholars); First Publication Year (Microsoft Academic / OpenAlex).

year periods, only the non-parametric tests remain significantly positive, while the Pearson correlation does not show significance. This divergence may suggest that over longer periods, the relationship between public interest and scholarly output becomes less linear or direct, possibly due to the maturation of the technology or shifts in research focus. These findings highlight the importance of considering both short-term and long-term trends in technology awareness and their relation to academic activities. Granger causality could further explore the dynamic relationship between public interest and scholarly activity, helping to determine if there's a time-lagged effect where public interest in Wikipedia predicts subsequent changes in research activities. The limited data points and the requirement for data stationarity restrict the effectiveness of our analysis in this scenario.

The correlation analysis results for the Generality Index, which assesses a technology-adjacent entity's breadth of application across various contexts within Wikipedia, reveal significant insights about its broader relevance and applicability. It can serve as an indicator of a general-purpose technology (GPT). General-purpose technologies are defined by their wide-ranging impact across many sectors, fundamentally altering industries and economies. By measuring the frequency and diversity of mentions across various contexts, the Generality Index captures an essential characteristic of GPTs: their pervasive applicability. The significant positive Pearson correlation between the Generality Index and patent counts for 2023, with a coefficient of 0.2043, indicates that technologies with a higher generality score tend to have more patent filings. This suggests that technologies mentioned

frequently across different Wikipedia articles are more likely to be innovative and patentable. Similarly, the positive Pearson correlation with total funds raised (coefficient of 0.2047) underscores that more general technologies attract more investment, reflecting their potential commercial viability. However, the lack of significant positive correlations in the Spearman and Kendall Tau tests with total funds raised suggests that this relationship may not be strictly monotonic, indicating variations in how generality impacts funding across different types of technologies. The observed variation in patenting levels across industries can be meaningfully correlated with the specific characteristics of the technologies and the R&D process, particularly in relation to the nature of the technological regime (Arundel & Kabla, 1998; Breschi et al., 2000). Overall, these results support the validity of the Generality Index as a measure of a technology's broad relevance and potential impact in both academic and commercial areas.

The significant positive correlations across all tests for the Deeptech Index provide substantial support for its effectiveness in capturing the scientific and innovative depth of technologies. The correlation derived from Wikipedia2Vec cosine similarity measures with real-world data, such as reference links and scholar counts, indicates that the Deeptech index represents relevant aspects of advanced technologies as they are recognised and valued in practical and academic domains. This suggests that quantifying technological alignment with deep research and business aspects via Wikipedia2Vec embeddings reliably reflects a technology's depth and impact in academia. It validates using Wikipedia-based data and the Wikipedia2Vec model as robust tools for assessing technological significance, bridging the gap between theoretical conceptualisation and real-world relevance and application. These findings underscore the Deeptech Index as a valuable tool for determining technologies based on their scientific rigour and potential impact, reflecting its utility in pinpointing emerging areas likely to influence future technological landscapes.

The Age Index, identifying the emergence year of technology-adjacent entities based on historical mentions, shows significant positive correlations with the **First_Pub_Year** from the OpenAlex database. High correlation coefficients across all three statistical tests indicate a strong alignment between the Age Index and the earliest recorded scholarly mention of technology-adjacent entities. It demonstrates the Age Index's effectiveness in approximating the time when technologies first gained noticeable recognition and began to impact scholarly discourse. The positive results reinforce the reliability of using historical documentation frequencies to estimate the birth year of technologies, highlighting the index's value in historical analysis of technological development.

4.4.2 Comparison ET100 with other emerging technology lists

We reviewed a number of other emerging technology lists and chose to compare our emerging technology list with the robust and extensive OECD dataset (OECD, 2016). The list in the OECD report is derived from technology foresight exercises conducted by or for national governments of several OECD countries (Canada, Finland, Germany, United Kingdom) and the Russian Federation, as well as an exercise by the European Commission. These exercises assess promising technologies over a future 10 to 20-year horizon, pooling insights from various experts and analytical techniques. The OECD limits its focus on transformative technologies that significantly impact global socio-economic conditions and address critical challenges over the next two decades.

The ET100 list provides a more detailed and specialised overview of emerging technologies than the OECD's broader categorisations. For instance, while the OECD lists 40 technologies in total under four general fields like Digital Technologies, Biotechnologies, Advanced Materials, and Energy plus Environment, the ET100 dives deeper into specific applications and innovations within these areas—such as edge computing in digital technologies, personalised medicine in biotechnologies, and microgeneration in energy technologies. This granularity makes the ET100 more practical for stakeholders needing precise, actionable information on current technological trends, whereas the OECD's approach is better suited for broad strategic planning and policy guidance. The ET100's emphasis on cutting-edge and specific applications and a structured hierarchical system of meta tech-clusters and theme tech-clusters offers a clear, up-to-date roadmap for navigating the rapidly evolving tech landscape.

The ET100 overlaps significantly with the OECD list. The ET100 contains 31 (77.5%) technologies that exactly match, and 37 (92.5%) of the technologies appearing in the OECD list. The overlap primarily features in core areas like digital technologies and biotechnologies, where both identify significant trends such as AI, IoT, and synthetic biology. The ET100 and TA23k may not include some technologies due to the 2018 limitation of Wikipedia2Vec; our lists may not capture new concepts, terms, or updated information that has emerged since 2018. However, the ET100 employs a data-driven approach, using techniques (Wikipedia2Vec) and other third-party data combined with hierarchical clustering to analyse and update its list of emerging technologies. This allows for detailed, nuanced insights into specific technologies and their interconnections. In contrast, the OECD uses a more traditional method involving expert panels and foresight exercises, focusing on broader technological areas and long-term policy implications. Our approach enables more frequent updates and might also lead to a focus shift towards newer or rapidly evolving technologies, poten-

tially omitting stable ones that still appear in the OECD's less frequently updated list. This technical and selective approach helps the ET100 maintain a more dynamic and detailed roadmap tailored to current market and innovation cycles, complementing the OECD's strategic overview.

4.5 Usage Notes

The Cosmos 1.0 dataset comprises 23,544 technology-adjacent entities with a hierarchical structure, and 100 emerging technologies have been manually verified. We enhance the dataset with unique technology indices, such as Age of Tech, Deeptech, Generality, and Tech Awareness, and enrich it with metadata from Wikipedia and data from third-party sources, including Crunchbase, Google Books, and Google Scholar. In this section, we demonstrate how to utilise the Cosmos 1.0 dataset through several examples.

4.5.1 Atlas of Comos 1.0

We illustrate the 2D mapping of Cosmos 1.0 based on the features **tsne_x** and **tsne_y** from the "Embeddings for Semantic Analysis" category. Figure 4.4 visualises TA23k colored by seven theme tech-clusters while the stars represent the ET100. The three circular areas are the groups that are easily noticeable. The two circles at the edge are separated from the main body, forming two offshore islands (labelled "Traditional Knowledge Systems" and "Patent Law Concepts"). The third circle contains multiple technologies from ET100, notably ("Smart grid", "Microgrid", "Microgeneration", "Photovoltaics", "Grid energy storage", "Energy storage", "Wave power"), and we define these seven ET100 technologies as "Renewable Energy Technologies". We gather technologies through coordinates and notice that the two islands are technology-related terms rather than actual technologies. For example, "Traditional Knowledge Systems" primarily consists of philosophical concepts, religious doctrines, classical texts, epistemological frameworks, and metaphysical ideas from Indian, Chinese, and Buddhist traditions. "Patent Law Concepts" encompasses legal doctrines, administrative processes, classifications, and tools related to the patent system. The third circle, located at the centre of the map, includes the domains of renewable energy generation, storage, grid integration, and power system management, affirming its technological nature. This spatial layout demonstrates the utility of Cosmos 1.0 in distinguishing between technological and technology-related entities. This visual and structural differentiation is crucial for identifying emerging technologies, as it allows researchers and poli-

cymakers to focus on regions of dense innovation, observe thematic overlaps, and detect underexplored areas with high growth potential.

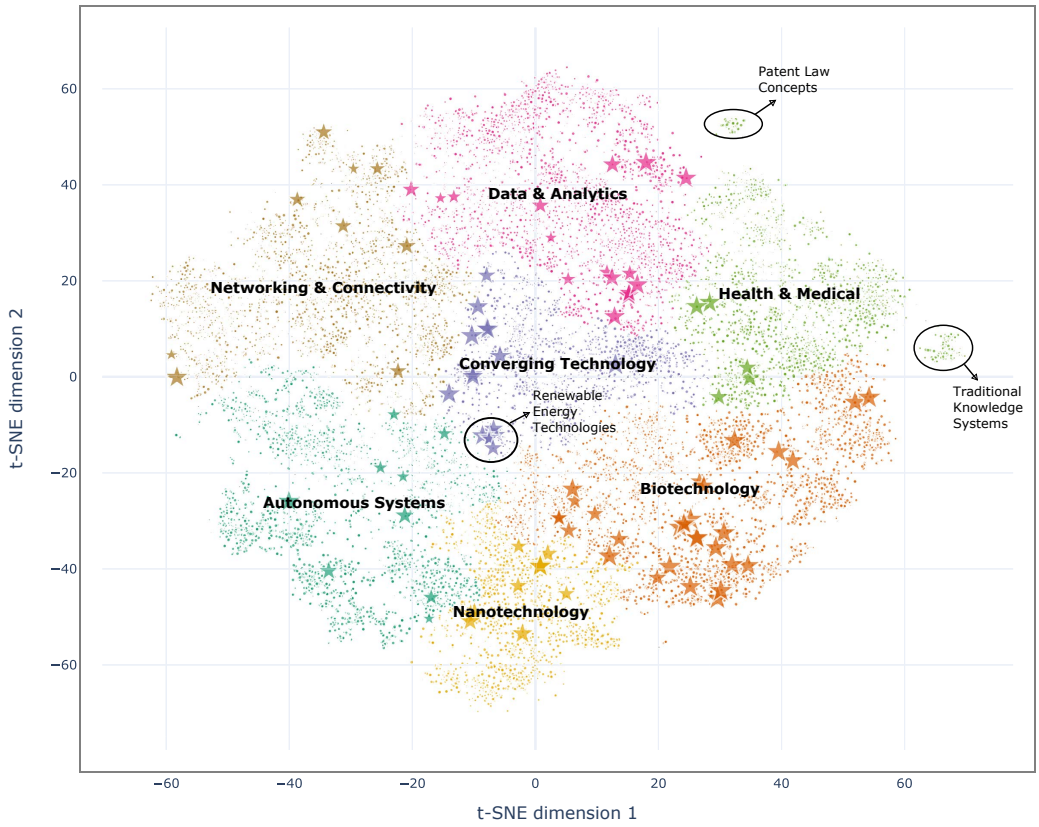


Figure 4.4: Map of Cosmos 1.0 (n=23,544) with highlighted ET100. Utilising the dimensionality reduction and hierarchical clustering algorithms to visualise the 7 theme tech-clusters of the TA23k in the 2D map. The dots represent 23,544 technologies in Cosmos 1.0, and the stars represent the ET100, a manual selection of widely recognised 100 technologies. This map facilitates the observation of technological distributions. The interactive version of the map at here.

4.5.2 Intra-Cluster Similarity

To further understand a group of closely related emerging technologies, we utilise intra-cluster cosine similarity to identify and rank technology-adjacent entities most relevant to a given seed set. We define "Renewable Energy Technologies" in the subsection Atlas of Comos 1.0, as the seven ET100 technologies within this circle ("Smart grid", "Microgrid", "Microgeneration", "Photovoltaics", "Grid energy storage", "Energy storage", "Wave power"). Since those technologies are included in the same theme tech-cluster "Converg-

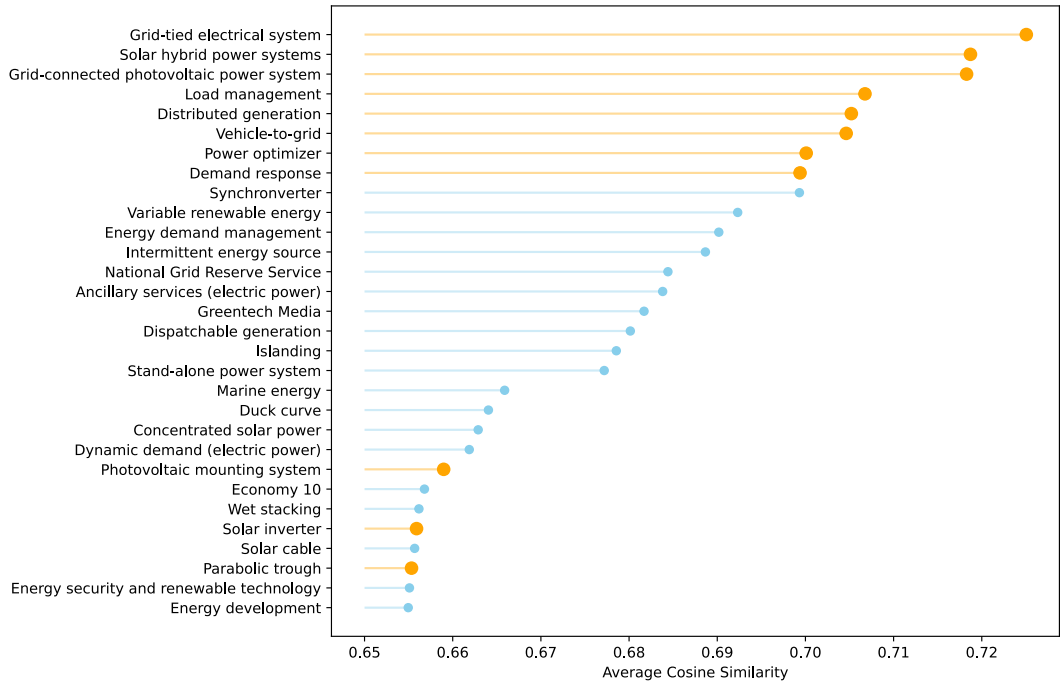


Figure 4.5: Top 30 technologies that are closest to the ET100 group of seven technologies "Renewable Energy Technologies" based on average cosine similarity. The higher-ranking entities demonstrate strong relevance. The entities highlighted in orange represent the technologies we observed for different potential subdomains of "Renewable Energy Technologies" in the space of solar and photovoltaic technologies.

ing Technology", we limit the scope to this theme and calculate the average cosine similarity between each candidate technology and the seven ET100 technologies within "Renewable Energy Technologies". A higher average cosine similarity indicates a stronger semantic relationship to "Renewable Energy Technologies". Based on this metric, we rank all candidate technologies in descending order and select the top 30 most relevant entities, shown in Figure 4.5. The retrieved list demonstrates high quality and strong relevance. Most entities (highlighted in orange) fall within a subdomain of renewable energy, which includes solar and photovoltaic technologies (e.g., "Power optimiser", "Solar inverter", "Photovoltaic mounting system", "Parabolic trough"). Energy experts can detect more granular subdomains on the basis of the list. These ranked entities represent a cohesive technological framework encompassing generation, storage, distribution, and demand-side management. This approach helps contextualise and deepen our understanding of "Renewable Energy Technologies" and its broader ecosystem by revealing closely related and complementary technologies within the same semantic space.

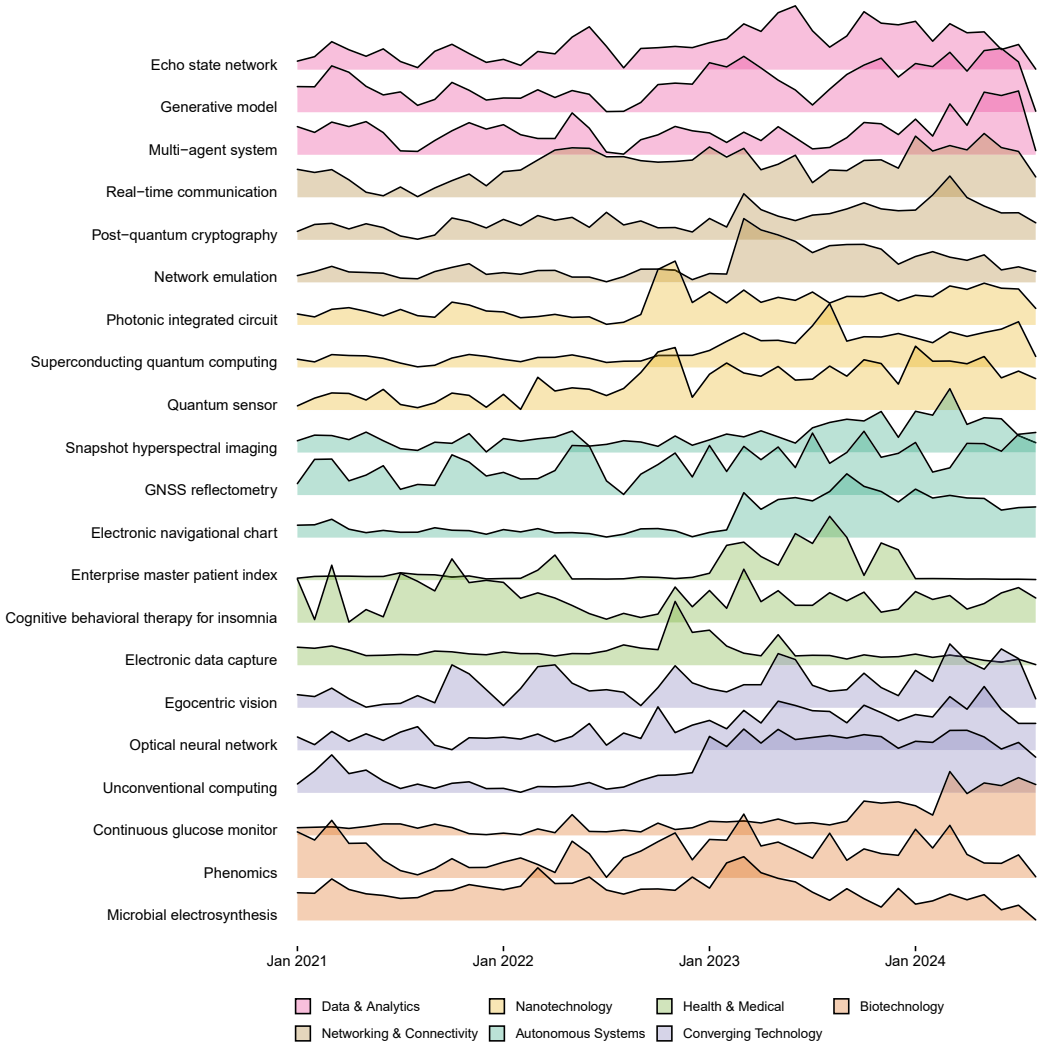


Figure 4.6: Monthly Pageviews Trends of Technology-Adjacent Entities of TC7. Three fast-moving frontier technologies are selected as examples for each of the seven theme tech-clusters (TC7) identified. This custom list of technologies is filtered using various technology indices that meet the different criteria. The criteria include all positive indices values, such as **Awareness_Index**, **Publications_Growth_Index** and **Google_Patent_Growth_Index** (See the description of indices in Section Data Records). Moreover, it only keeps the technologies that appeared after the year 1950 (**Age_of_Tech_Index** > 1950).

4.5.3 Multi-Indices Filtering Strategy

Academics and researchers can leverage this dataset to track technological advancements and trends within their fields of interest using multi-index filtering. Figure 4.6 illustrates the monthly pageviews counts movement of an example of custom lists of technology-adjacent entities filtered by multiple technology indices. We select technology-adjacent

entities from each of the TC7 theme tech-clusters that consistently exhibit strong signals across multiple features, retaining only those with positive values in **3yr_Awareness_Index**, **3yr_Publications_Growth_Index**, and **3yr_Google_Patent_Growth_Index**. These indicators respectively capture public interest, academic momentum, and industrial innovation. Additionally, we limit the selection to technologies that emerged after 1950, based on the **Age_of_Tech_Index**, to ensure relevance to recent developments. Finally, we manually review and highlight three fast-moving frontier technologies as examples of rapidly evolving technologies for each theme tech-cluster.

The multi-indices filtering strategy facilitates a deeper understanding of how specific technologies are selected and their broader implications within the technology landscape. It increases the understanding of measurements of emerging technologies. Most studies focus on the patent record (Eilers et al., 2019; Sun et al., 2024) or research literature (Sick & Bröring, 2022), such as robotics and internet-of-things (Kose & Sakata, 2019; Sun et al., 2024). The Cosmos 1.0 provides a rich data source for identifying technologies on a broad scale through the eight categories.

Overall, the Cosmos 1.0 dataset provides valuable insights for researchers, policymakers, and corporations by capturing emerging technologies through complementary signals. For researchers, it enables targeted discovery of fast-evolving research frontiers. Policymakers can leverage the strategy to design policies based on technological socio-economic shifts. Corporations can use it to identify commercially promising technologies early, optimise R&D investments, and inform strategic planning in competitive landscapes. This multidimensional framework enhances the relevance and ability of technology detection across sectors.

4.6 Data Availability

The Cosmos 1.0 dataset generated and analysed in this study is publicly available on Figshare at https://doi.org/10.6084/m9.figshare.28268561 or on Github repository at https://github.com/xian-gong-elaine/Cosmos_1.0. The file includes all the features we discussed in the paper.

4.7 Code availability

The code for the analyses and the experiments in this paper is available at https://github.com/xian-gong-elaine/Cosmos_1.0.

Author Declaration

The following chapter contains content from the following publication.

Gong, Xian, McFarland, C., McCarthy, P. X., Griffith, C., & Rizoiu, M.-A. (2023). Informing innovation management: Linking leading r&d firms and emerging technologies. *Proceedings of the International Society for Professional Innovation Management (ISPIM)*, 1–16

Author Contributions: Xian Gong led the data collection, preprocessing, and analysis, developed the methodology integrating entity embeddings with firm–technology mapping, produced the main visualisations and results, and drafted the initial manuscript. C. McFarland contributed to the design of the case study and drafted and revised the manuscript. P.X. McCarthy provided conceptual framing and theoretical grounding for linking emerging technologies to global R&D firms. C. Griffith assisted with robustness checks, and critical review of the manuscript. M.-A. Rizoiu supervised the research, guided methodological development and validation.

Production Note:
Signature removed prior to publication.

Xian Gong

Production Note:
Signature removed prior to publication.

Claire McFarland

Production Note:
Signature removed prior to publication.

Paul X. McCarthy

Production Note:
Signature removed prior to publication.

Colin Griffith

Production Note:
Signature removed prior to publication.

Marian-Andrei Rizoiu

CHAPTER 5

INFORMING INNOVATION MANAGEMENT: LINKING LEADING R&D FIRMS AND EMERGING TECHNOLOGIES

Understanding the relationship between emerging technology and research and development has long been of interest to companies, policy makers and researchers. In this paper new sources of data and tools are combined with a novel technique to construct a model linking a defined set of emerging technologies with the global leading R&D spending companies. The result is a new map of this landscape. This map reveals the proximity of technologies and companies in the knowledge embedded in their corresponding Wikipedia profiles, enabling analysis of the closest associations between the companies and emerging technologies. A significant positive correlation for a related set of patent data validates the approach. Finally, a set of Circular Economy Emerging Technologies are matched to their closest leading R&D spending company, prompting future research ideas in broader or narrower application of the model to specific technology themes, company competitor landscapes and national interest concerns.

5.1 Introduction

Understanding how emerging technologies are being used by companies, including their capabilities and specialisations, is of increasing interest for a wide range of stakeholders including investors, customers, suppliers, employees, as well as researchers and policymakers.

However, traditional data sources, taxonomies, and methodologies have limitations when it comes to understanding the dynamic and nascent nature of emerging technologies, and the individual capabilities and strengths of the leading firms investing in research and development. For example, patent analysis - a key source for many existing studies - works well for some industries such as pharmaceuticals but less well for others such as software development.

The foundation of this research is a methodology to create a dynamic and structured list of the most significant emerging technologies using machine learning trained on the knowledge contained in Wikipedia. This method leverages the output from the tens of thousands of scholars and experts who contribute to articles about technology on Wikipedia, the world's largest and most read reference work in history and assessed as highly reliable for science and technology information (Bruckman, 2022).

The resulting list of 100 emerging technologies (the ET100) has been identified based on analysis of over 50,000 Wikipedia articles most related to the concept of emerging technologies and grouped into hierarchical clusters (that can be interrogated at different levels of detail). The significance of the ET100 list has been demonstrated through positive correlations with the level of attention paid to the Wikipedia articles and current investment in these technologies. Additionally, the ET100 maps to other lists of emerging technologies published by organisation such as the OECD, WEF and Gartner.

A similar machine learning methodology is then used to map the relationship between the ET100 and the activities of the businesses investing the largest sums in R&D worldwide. This approach uses the European Commission's most recent list of the world's top spending R&D companies – the 2022 EU Industrial R&D Scoreboard (Nindl et al., 2023), and the knowledge embedded in their corresponding Wikipedia profiles, to find the closest association between the companies and emerging technologies.

Recent research has demonstrated the potential of deriving detailed technical information and other latent knowledge from word embedding language models (Tshitoyan et al., 2019). In this paper we illustrate the potential to use the latent knowledge in a novel word embedding language model to create new maps of:

- Leading companies by R&D investment and their relationship to one another.
- The global emerging technology landscape.
- The relationship between these leading R&D investing companies and an established set of emerging technologies (the ET100).

This paper is structured into six remaining sections following the Introduction. Section 2 defines the core problem; Section 3 reviews the current understanding from both firm and policy perspectives; Section 4 highlights the current limitations and research gaps; Section 5 details the methodology underlying the ET100/R&D Links model, including embedding construction, alignment, and clustering; Section 6 presents the empirical findings, visualisation results, proximity analyses, and patent-based validation; Section 7 discusses limitations, implications for innovation management and policy, and potential avenues for future research. This structure focuses on analytical coherence while highlighting the model's practical relevance for both managers and policymakers.

5.2 What is the problem?

The relationship between business R&D and emerging technologies is difficult to determine at scale and on a dynamic basis.

This is the case for two main reasons, firstly because business R&D is often tightly directed towards the development of new products and services, and secondly the way in which information about the nature of the business R&D is communicated by firms, and collected by national reporting organisations.

The evidence of business R&D focus on the development of products and services can in some cases be gleaned from the record of patents applied for, from company annual reports and media releases. There are some examples of companies being public about their investment in advancing understanding in emerging technologies[1] but much of the detail of business R&D is hidden.

The academic literature is a rich source of business R&D case studies, however it is lacking in linking a generalised group of firms (such as is found in the EU Industrial R&D Investment Scoreboard) with a robust list of emerging technologies.

While national governments are interested in levels of business R&D expenditure (termed BERD in relation to national accounts) because of its contribution to economic growth, in this context there is generally little information collected about the nature of the R&D investment beyond broad categories based either on traditional industry classifications or custom expert derived taxonomies. Business R&D expenditure is measured at a firm level,

[1]For example, a number of big tech companies, including IBM, Microsoft, Google and Intel are making significant investments in quantum computing R&D via VentureBeat https://venturebeat.com/data-infrastructure/quantum-progress-how-ibm-microsoftgoogle-and-intel-compare/ accessed Apr 2023.

often reported at an industry or general research field level[2] and compared at a global level by organisations such as the OECD.

By contrast emerging technologies are often examined at an individual technology level or as a category by researchers and policy makers alike. Existing approaches to understanding emerging technologies as a group rely mostly on established legacy taxonomies, such as Standard Industrial Classification codes - long held to be inadequate in examining high tech environments (Kile & Phillips, 2009); or on expert developed categories that are not dynamic and therefore are hard to maintain. Neither resulting group of emerging technologies provides insights into the complex relationships or associations with other technologies let alone with any business.

There is a gap therefore in looking at the relationship between those firms spending the most on R&D globally and the emerging technology landscape. Since both R&D and emerging technologies play an important role in innovation, better understanding each of these factors and the relationship between them is of key interest to innovation managers.

5.3 Current understanding, what do we know about the problem?

Looking in more detail at the current situation in relation to business R&D and emerging technologies from the perspectives of both commercial interests and policy makers provides important background to this research.

5.3.1 Importance of emerging tech to companies

As well as being at the heart of economic growth, technological change is having a profound influence on reshaping the worldwide economy as is evidenced by almost any study of recent history. Looking at publicly listed companies, over the last 30 years the largest companies in the world have changed from predominantly car and fossil fuel manufacturers to digital tech giants.

However, from a company level perspective, technology is so pervasive that it can be hard to determine when a new technology is worth paying attention to. Indeed, the academic definition of emerging technologies as novel, relatively fast growing, with the potential for significant impact on society and the economy while being also somewhat un-

[2] Examples of Fields of Research collected include 'Information and Computing Sciences', 'Engineering', and 'Agricultural, Veterinary and Food Sciences'.

certain and ambiguous (Rotolo et al., 2015) presents challenges for companies in terms of balancing the risk and reward of deciding to make an investment.

Yet, estimates of the size of the 'New Technologies' market in 2023 exceed $US 1.3 trillion[3]. IDC Global ICT Spending forecasts for 2020 – 2023 predict that over the next 5-10 years new technologies such as robotics, AI and AR/VR will grow to 25% of ICT spending, particularly as firms generate cost savings from implementation of cloud and automation services. These estimates are important inputs for company decision making as they seek to deliver value to shareholders and customers, invest in employees, partner with suppliers, and support the communities in which they operate[4].

An element of this decision making is whether a company will adopt emerging technology, assemble emerging technology into new products and services or develop emerging technology. Elements of business R&D exist within each of these approaches.

The extent to which there are links between leading R&D spending companies and individual emerging technologies is the key focus of this paper.

5.3.2 Importance of emerging tech to policy makers

As recognised formally by the United States government in their announcement of the Office of the Special Envoy for Critical and Emerging Technology, "technology is increasingly central to geopolitical competition and the future of national security, economic prosperity, and democracy"[5]. While much of the commentary is inevitably focused on issues of national security and democracy, the emphasis on economic prosperity is of key interest to businesses. The US government has identified a list of critical technologies which broadly encompasses biotechnology, advanced computing, artificial intelligence, and quantum information technology[6].

Likewise, the European Union has recently released a draft report on critical technologies for security and defence[7] and announced a new 'Observatory of Critical Technologies'

[3] IDC – Global ICT Spending Forecast 2020 – 2023

[4] Statement on the purpose of a corporation via Business Roundtable https://system.businessroundtable.org/app/uploads/sites/5/2023/02/WSJ_BRT_POC_Ad.pdf accessed Apr 2023

[5] Department Press Briefing - January 3, 2023 via United States Department of State https://www.state.gov/briefings/department-press-briefing-january-3-2023/ accessed Apr 2023

[6] Critical and Emerging Technologies List Update via Whitehouse.gov https://www.whitehouse.gov/wp-content/uploads/2022/02/02-2022-Critical-andEmerging-Technologies-List-Update.pdf. Also US Critical and Emerging Technology Strategy via Vivekananda International Foundation (vifindia.org) https://www.vifindia.org/article/2023/january/30/us-critical-and-emerging-technologystrategy accessed Apr 2023

[7] Critical technologies for security and defence: state of play and future challenges (2022/2079 (INI)) https://www.europarl.europa.eu/doceo/document/ITRE-PR738598_EN.pdf via European Parliament Committee of Industry, Research and Energy accessed Apr 2023.

with a focus on value and supply chains particularly for space, defence and related civil sectors[8].

These two come together in the US-EU Trade and Technology Council created to coordinate key global trade, economic and technology issues with specific focus on a number of emerging technologies[9].

The 2022 EU Industrial R&D Investment Scoreboard, reflects the EU's focus on understanding how Europe compares with both the US and China in terms of overall investment in business R&D; where there are strengths and how the EU wants to excel with 'green and digital transformations'. The New European Innovation Agenda, released in July 2022, has a goal of promoting 'firm creation and growth in emerging technologies to trigger spill overs between sectors'.

Signalling by governments of areas of focus inevitably concentrates investment by businesses as it provides the certainty conducive to a strong investment climate.

5.3.3 Importance of R&D to firms

It is clear from the 2022 EU R&D Scoreboard that a significant amount of funding is directed by firms towards R&D. In 2021 the 2,500 companies investing the most in R&D in the world account for approximately 86% of the world's business funded R&D, a total of 1093.9 billion Euro[10].

The relationship of business R&D expenditure and business performance is widely studied by academics and business consultants alike, with evidence that R&D expenditure increases future profitability (Branch, 1974; Eberhart et al., 2004; Sougiannis, 1994), and has a positive impact on market value (Armstrong et al., 2006; Sougiannis, 1994) and stock returns (Chan et al., 2001; J. Kim, 2022; Lev & Sougiannis, 1996).

The relationship of all this business R&D activity with emerging technologies is the key question of interest in this paper.

[8]Note, the EU approach is a new endeavour due to a new Action Plan on synergies between civil, defence and space industries. https://ec.europa.eu/commission/presscorner/detail/pt/QANDA_21_652 Q&A: Synergies between civil, defence and space industries via Europa.eu accessed Apr 2023.

[9]U.S.-EU Trade and Technology Council https://www.trade.gov/useuttc via International Trade Administration accessed Apr 2023.

[10]The 2022 EU Industrial R&D Investment Scoreboard | IRI (europa.eu) https://iri.jrc.ec.europa.eu/scoreboard/2022-eu-industrial-rdinvestment-scoreboard via Europa.eu accessed Apr 2023.

5.3.4 Importance of business R&D to policy makers

R&D investment has long been found to be a key driver of economic growth. Business R&D makes up a significant portion of the Gross Expenditure on R&D for many countries, leading to policy instruments to stimulate business R&D, including both favourable tax treatment and direct subsidies for specific industries. The practice of comparing the levels of R&D investment (public and private) at a national level[11] has become an indicator of commitment and performance (Schot & Steinmueller, 2018). The notion of R&D spillovers, where investment by one entity (for example government) has a beneficial impact to other entities (for example businesses) (Griliches, 1991; Grossman & Helpman, 1993), is a key consideration of government investment in R&D. Additionally, the commercialisation of government R&D by companies has become an area of increasing focus[12].

5.4 Limits of current understanding, thesis and research questions

The macro level intersection of these things is interesting for policy makers, academics, and future-focused companies. Current available data is isolated. For example, media releases / annual report updates from firms about their R&D projects provides a biased (company initiated) perspective. The value of R&D expenditure captured by OECD governments provides a macro view, and while this provides some insight at the industry level, it gives little indication of expenditure direction at the technology level.

Since both R&D at leading firms and the relationship of these firms to emerging technologies play an important role in innovation, better understanding each of these factors and the relationship between them is of key interest to innovation managers.

This paper sets out the thesis that word embedding language models can provide ways to map the proximity of emerging technologies to leading R&D investing companies (and vice versa) and that these maps can be cross validated through established third party data such as patent counts. In doing so it seeks to address the following questions:

[11] The OECD collects and publishes this data in multiple forms on a regular basis. OECD publications include the Main Indicators and Science and Technology (MSTI), the OECD Science Technology and Industry Scoreboard and numerous datasets. https://www.oecd.org/sti/inno/researchanddevelopmentstatisticsrds.htm via Research and Development Statistics (RDS) – OECD accessed Apr 2023

[12] For example, the United States government has a series of initiatives related to commercialising federally funded R&D, including https://www.nist.gov/tpo/lab-marketviaLab-to-Market (L2M) | NIST. In Australia, the NSW Government also has a strong focus, https://www.dpc.nsw.gov.au/publications/categories/turning-ideas-into-jobsaccelerating-research-and-development-in-nsw/ Turning ideas into jobs: Accelerating research and development in NSW via Premier & Cabinet accessed Apr 2023.

- Which emerging technologies are closest to the top 500 firms by R&D expenditure?
- For each of the 100 identified leading emerging technologies, which company from the top 500 R&D spending companies is closest?
- What does this combined dataset look like from a two-dimensional perspective? What are the noticeable clusters? What outliers appear? What further questions does this raise?

5.5 Methodology

This innovation-in-practice paper draws on previous work in emerging technology categorisation using machine learning clustering techniques and links a robust set of defined leading emerging technologies to the firms spending the most on R&D measured at an international level. The list of emerging technologies (the ET100) has been constructed by mapping thousands of individual technologies and their relationships to each other using machine learning analysis of the similarity between the way different technologies are described. This approach produces a novel multi-level technology taxonomy that reveals the complex relationships between all the technologies, providing insight into the most adjacent technologies and those that are most connected within different technology ecosystems. It also serves as a foundation that can be used to interrogate other sources of data.

The natural language processing tool used in this research has been trained on the full corpus of English language Wikipedia consisting of 4.4 million articles and 1.9 billion terms created by over 40 million editors. Known as Wikipedia2vec (Yamada et al., 2020), this toolkit represents defined entity vectors in the same high-dimensional space. It learns high-quality word embedding from Wikipedia, placing semantically similar words and entities close to one another in the vector space, and incorporating additional information from Wikipedia, such as article titles, links and categories. The entity vectors are particularly powerful because they are trained on a large, multilingual corpus of Wikipedia articles, which covers a wide range of topics and domains. This means that the entity vectors will likely capture a wide variety of semantic relationships between entities, including those that span different languages and cultures.

The research work underpinning this paper defines as entity vectors all the subcategories and pages contained within the top level category of emerging technologies in Wikipedia; and all the Wikipedia pages associated with the businesses listed in the 2022 EU Industrial R&D Investment Scoreboard. An algorithm was designed to reveal the em-

beddings or 'similarity' of the defined entities also referred to here as links or associations. The resulting model is referred to from here as the 'ET100/R&D Links' model.

Since technologies and organisations are different types of entities, the concept of orthogonal matrices in machine translation is introduced to rotate the company space to align with the technology space. These matrices are particularly useful because they preserve the distances and angles between vectors, which can help to maintain the semantic relationships between firms and technologies during the alignment process. Dimensionality reduction and clustering analysis creates visualisations that can effectively convey the relationships among technologies and companies. Moreover, it improves interpretability and identifies relationships due to grouping similar entities together into clusters.

The resulting 'ET100/R&D Links' model can be visualised as a map of the 100 identified emerging technologies and the leading 700 or so companies by R&D expenditure. Also revealed is the closest emerging technology to each of these companies, the closest company to each emerging technology and the relationships between individual companies and individual technologies, and between groups of companies and groups of technologies.

The notion of 'similarity' and 'distance' refer to the mathematical translation of the positioning of the emerging technologies and the companies in the 'ET100/R&D Links' model. Each emerging technology and company is represented by a unique vector, enabling the distance between each to be measured.

5.6 Findings

The novel approach to examining the relationships between emerging technologies and leading R&D companies resulting in the 'ET100/R&D Links' model, visualised as a map, enables investigation of a number of exploratory research questions and raises others.

5.6.1 A map of the emerging technologies and leading R&D companies reveals clusters and white space

Taking a broad view initially, Figure 5.1 shows an optimised and cross-validated map of the companies and the emerging technologies. There are clearly several clusters of companies and technologies. There is also a section of the map that is less dense – where there is more white space between companies. Note the size of the Company bubbles reflects the magnitude of their R&D investment. Larger equals more.

Nine natural groups or clusters of companies and technologies have been identified within this map using hierarchical clustering, as shown in Figure 5.2. GPT3 was used to suggest labels to describe each group of companies and the authors have settled on the current list taking this input and the most distinctive industry in the cluster into account.

Two company clusters have high concentrations of close emerging technologies – the Technology and Financial Services cluster (17 technologies) and the Pharmaceuticals and Biotech cluster (20 technologies). This is unsurprising and reflects the maturity and widespread application of both these fields.

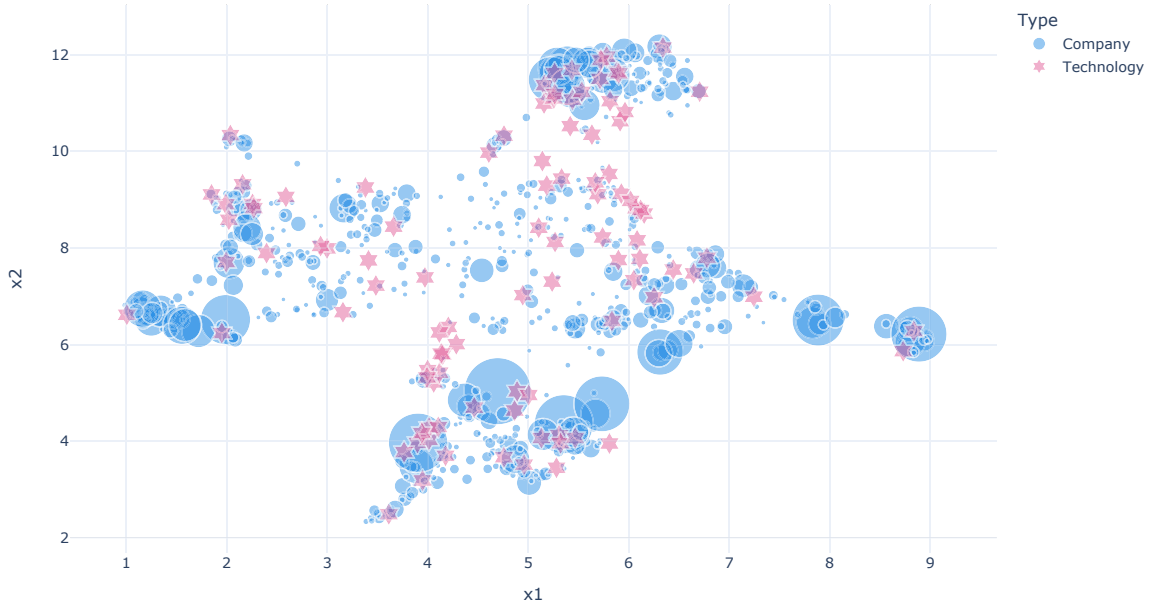


Figure 5.1: Visualisation of emerging technologies and leading R&D spending companies map showing distances and relative sizes of R&D expenditure.

By contrast, the company cluster for Telecommunications has the least concentration of close emerging technologies, with just three emerging technologies close to the companies within this clusters. Additionally while the closest adjacent cluster is Electronics and Manufacturing, the Telecommunications cluster is some distance away from several of the emerging technologies often associated with the telecommunications industry, such as Local Positioning System and Internet-of-Things.

Returning to the area of relative white space within the map, where there is less density of both companies and technologies, there are two noticeable insights. Firstly, the technologies within this location tend to be more specific in terms of both definition and ap-

plication than some of the technologies appearing in the more closely linked clusters. For example, Small Satellite and Distributed Acoustic Sensing have very particular applications at this point in time.

The second insight is the companies within this white space are generally spending less on R&D comparatively, with a large proportion below 500m Euro. There are no 'mega' spending companies within this part of the map. Note that the more tightly defined clusters of companies and technologies tend to contain a number of 'mega' spenders and a number of closely linked technologies. Could this be evidence of the development of R&D ecosystems within these clusters?

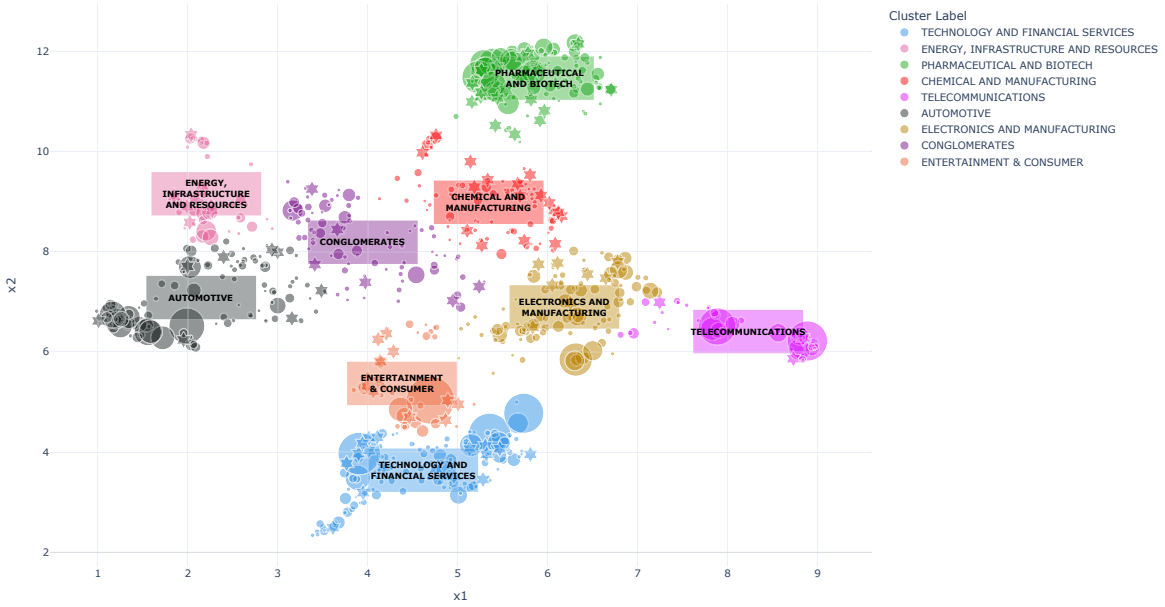


Figure 5.2: Hierarchical clustering reveals nine natural groups of leading R&D spending companies and emerging technologies.

5.6.2 Zoomed in perspective one: Closest emerging technologies to the leading R&D spending companies

Beyond the map visualisation it is possible to zoom in to the 'ET100/R&D Links' model to examine two perspectives. Firstly, the closest technologies to the companies appearing in the 2022 EU Industrial R&D Investment Scoreboard. While revealing the closest technologies, this also provides a comparative perspective on the relative position of the set of closest technologies to different companies. For some companies there are a number

of technologies clustered together, for others there is a single closest technology and then some distance to the next closest technology, for others there is a more even spacing between close technologies. Figure 5.3 sets out one subset of companies with a variety of top five closest technology positions to illustrate this.

Looking in detail at Figure 5.3, it is clear that there is some distance between 5G, the closest technology to telecommunications company Nokia, and the next closest technologies which are Unified Communications, Internet of Things, Local Positioning System and Cloud Computing, with these latter four technologies clumped together.

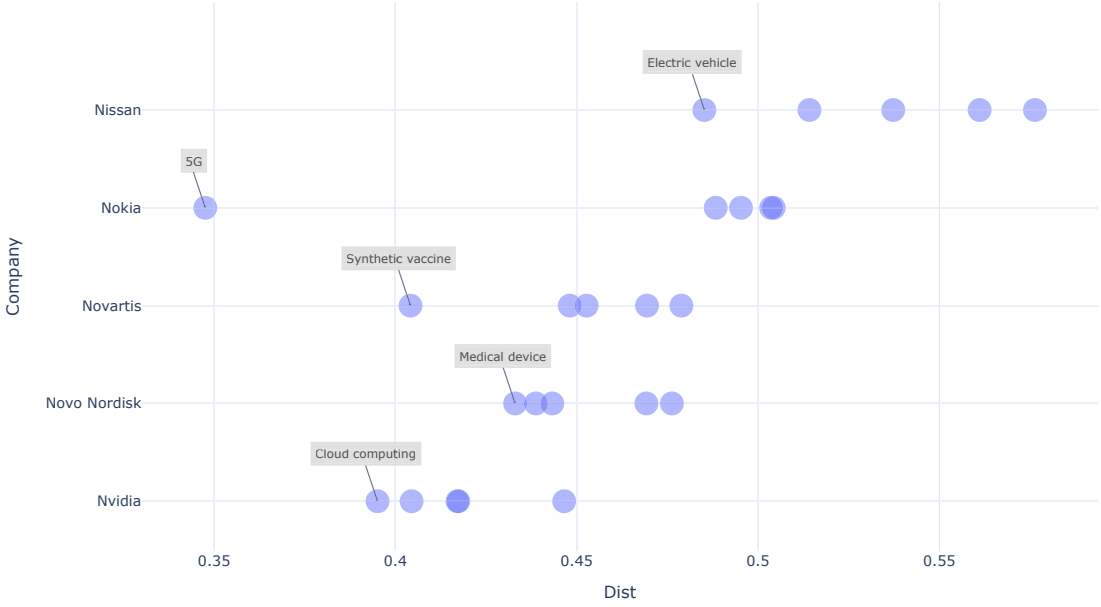


Figure 5.3: A subset of leading R&D spending companies with their closest emerging technologies.

By contrast, the closest technology to automotive company Nissan is Electric Vehicle, although this is only about as close as Unified Communications is to Nokia. The other close technologies to Nissan are Fleet Management, Fuel Cell, Autonomous car and Energy content of biofuel. These four technologies are relatively evenly spaced and all at a further distance from Nissan than the position of the top five close technologies of Nokia.

Novartis and Novo Nordisk are both pharmaceutical companies and while their closest technologies are different, with Synthetic vaccine for Novartis and Medical device for Novo Nordisk, they do share some similarities beyond this. For both companies, their second and third closest technologies are Biotechnology followed by Cell Therapy, both of which

are relatively closer to Novo Nordisk. Beyond this for Novartis, Vaccine Efficiency followed by Neurotechnology round out the top five closest technologies, while these positions are held by Regenerative Medicine and Clinical Oncology for Novo Nordisk.

5.6.3 Zoomed in perspective two: Closest companies to emerging technologies defined in ET100

Looking at the alternative perspective of which companies in the EU Industrial R&D Investment Scoreboard are closest to the emerging technologies within the ET100 also shows some strong variance in positioning in a one dimensional view. Figure 5.4 shows a subset of technologies to illustrate this.

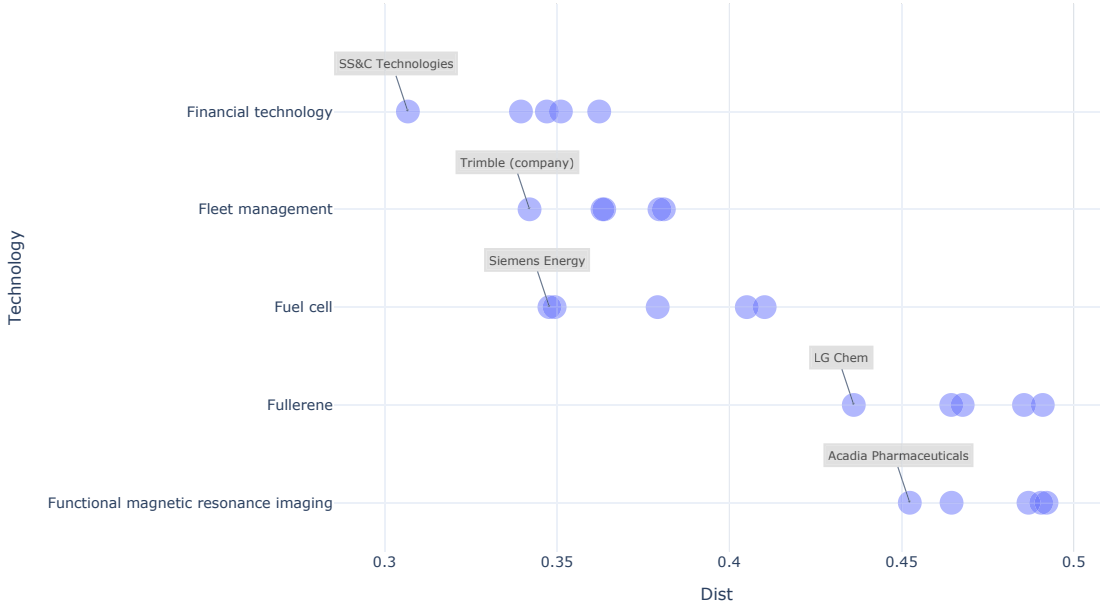


Figure 5.4: A subset of emerging technologies with their closest leading R&D spending companies.

In this group of technologies Fullerene and Functional magnetic resonance imaging have no companies relatively close to them suggesting they are still nascent; whereas the more mature financial technology (fintech), fleet management and fuel cell technologies have established companies close to each of them.

Across the whole set of 100 emerging technologies, some companies appear to have more of a dominant lead in their association with technologies when compared to nearest

Table 5.1: Identified Circular Economy Emerging Technologies matched to their closest leading R&D spending company within the 'ET100/R&D' Links model.

Circular Economy Emerging Technology	**Closest Company from the 2022 EU Industrial R&D Investment Scoreboard**
Biocatalysis	Novozymes
Biomaterial	Mitsui Chemicals
Carbon capture and storage	Siemens Energy
Energy management	Huaneng Power
Grid energy storage	Goldwind
Microgrid	Siemens Energy
Precision agriculture	Hexagon AB

Sources: Source: ET100/R&D Links model.

rivals. Illumina for example in Whole Genome Sequencing is a long way ahead of its nearest rival, and for 5G a number of global telcos and OEMs led by China Mobile, Nokia and CommScope are in relatively close proximity. This raises the question about whether the closeness of association corresponds to relative market share or other measures of advantage, which is perhaps an opportunity for future research.

5.6.4 Zoomed in perspective three: Circular Economy Technologies and leading R&D spending companies

Looking specifically at Circular Economy Technologies (CETs), the 'ET100/R&D Links' model can be used to investigate the closest emerging technologies to the top companies identified as the leading inventors in Circular Economy Technologies (CETs) set out in the 2022 EU Industrial R&D Investment Scoreboard[13].

For example, European multinational BASF is identified in the 2022 EU Industrial R&D Investment Scoreboard as being a top European company inventing in CETs in multiple industries (Construction, Chemicals and Plastics, Fertilisers, Metals, and Batteries & Fuel Cells). The five closest emerging technologies to BASF in the 'ET100/R&D Links' model are Smart Material, Biotechnology, Synthetic Vaccine, Semiconductor, and Carbon capture and storage.

Conversely, the 'ET100/R&D Links' model can also be interrogated from the emerging technology perspective. Taking a handful of emerging technologies with application to the

[13]Table 5.3 Top five Scoreboard companies inventing in CETs per industry (2010-2019) from the 2022 EU Industrial R&D Investment Scoreboard | IRI (europa.eu) https://iri.jrc.ec.europa.eu/scoreboard/2022-eu-industrial-rd-investment-scoreboard via Europa.eu sets out companies from both the Global and European perspective. A small number of examples are used here to illustrate the strength and applicability of the 'ET100/R&D Links' model.

Circular Economy and determining which company from the 2022 EU Industrial R&D Investment Scoreboard is closest as set out in Table 5.1 provides an insight into the CET landscape.

5.6.5 Patent analysis validation

Finally, since this is exploratory research, an additional question of interest is whether the relationship between this set of leading R&D spending companies and emerging technologies, as constructed in the 'ET100/R&D Links' model, is consistent with existing research or third-party data? To determine this an analysis of global patent data as it relates to the emerging technologies and the companies in the 'ET100/R&D Links' model was undertaken.

Looking at the counts of global patents for all emerging technologies in the ET100 for a set of top companies from the EU Industrial R&D Investment Scoreboard there is a significant positive correlation for almost half (n=48) with the similarity measures within the 'ET100/R&D Links' model.

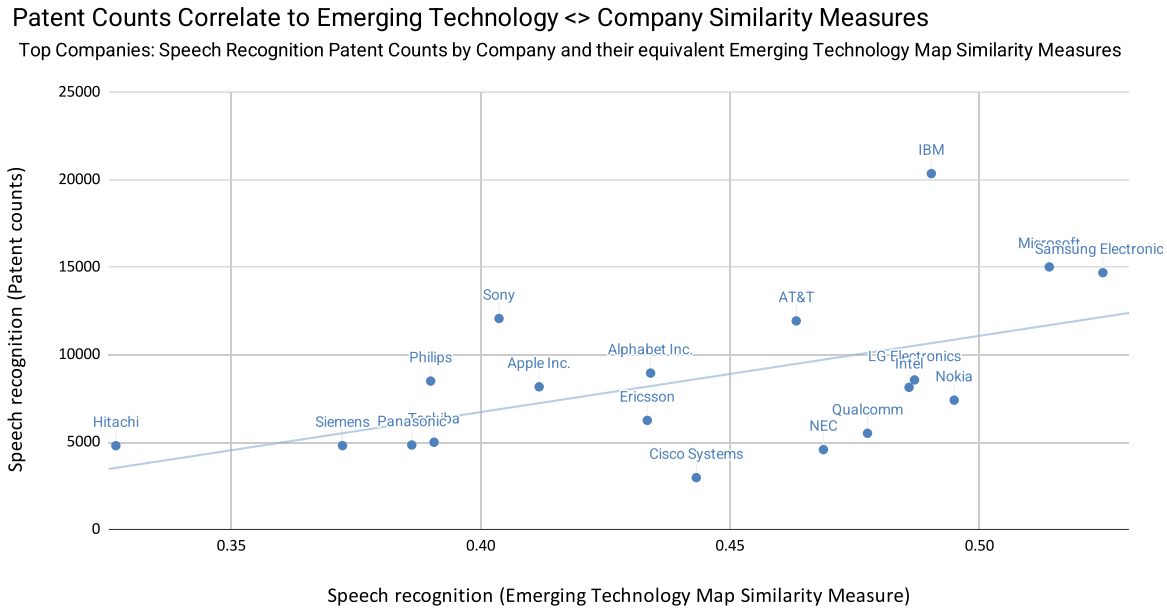


Figure 5.5: Example of significant positive correlation between patent counts and 'ET100/R&D Links' model for Speech Recognition technology.

The medical technologies within the ET100, where there is a lot of patenting activity (Clinical Oncology, Vaccine efficacy, Biotechnology) are among the most strongly corre-

lated. Many of the technologies that are not significantly correlated are either very new and may not have much patent momentum or are very specific technologies, which also may not have much patenting activity. Recall also that patents are highly applicable in some industries (eg pharmaceuticals) and much less so in others (software). This raises another opportunity for future research to better understand the significance of this pattern of correlation.

An example of the patent analysis undertaken is shown in Figure 5.5, which sets out the correlation of patent counts to 'ET100/R&D Links' model similarity measures for Speech Recognition.

5.7 Limitations, significance and future

In taking this approach to investigating the linkages between the top R&D spending companies and a defined list of emerging technologies some qualifications are required, and this is to do with the characterisation of the linkages for some companies with some of the emerging technologies. Due to the nature of the methodology, the exact nature of the association between a technology and a company may require further interpretation. It could show an existing investment in a technology, or an adjacency that could reveal a potential to leverage a technology in a novel manner, or it could show other associations related to some specific use or interest in a technology.

It is not certain that the linkages observable in the 'ET100/R&D Links' model are due to investments by the companies in question into the emerging technology in question. For example, a company such as food multinational Mondelez appears to be associated with vaccine efficacy due to a policy to vaccinate all its employees rather than involvement in the development of such technology in the vaccine space. Additionally, some linkages reflect multiple or ambiguous word meanings. For example, the term microsatellite has association with both genomic analysis within the biotechnology field, and the space industry.

This variation is a natural outcome from a natural language processing model such as the 'ET100/R&D Links' model underpinning this paper and is the reason why expert interpretation is so important in research. In this case to interpret the nature of the linkages between specific companies and technologies based on specific context. To ask the question "why has the embedded knowledge in both these sets revealed a specific association?".

For commercial innovation managers this research provides an insight into the competitive landscape amongst leading R&D firms. This provides rich information for the development of strategic planning and potentially for strategic partnership opportunities or

M&A activity. The correlation between the 'ET100/R&D Links' model and the patenting activity of companies and specific technologies (per the Speech Recognition example) reveals that in part, the model reflects technology strength and / or capability. Therefore, the information presented in the 'ET100/R&D Links' model provides insight into company-specific areas of potential investment in emerging technology development.

For policy-makers, insight into the proximity of particular emerging technologies to particular companies may provide another layer of detail as to where R&D investment is flowing, knowledge that may be valuable in a variety of policy contexts.

For innovation researchers this paper and accompanying presentation may be thought-provoking for new avenues of research. For other researchers this may provide ideas for research partnerships.

Beyond the approach examined in this paper, the methodology underpinning the 'ET-100/R&D Links' model can be applied more broadly or more narrowly to investigate more extensive lists of technologies and companies. Some examples of potential deep dives into specific technology themes include circular economy technologies, pandemic resilience and response related technologies, quantum computing related technologies. Or the methodology could be applied to examine the competitor landscape by defining a set of competitive companies with a view to developing understanding of the technology landscape within which they operate. This could be insightful for both established companies and new entrants. The methodology could also be applied to custom lists of technologies and companies to examine specific areas of national interest, for example the critical technology situation or the major supply chain environment.

The overall advantage of the methodology is the ability to identify linkages and therefore associations at scale between emerging technologies and leading R&D spending companies. The complexity of the web of linkages is significant and in aggregate accurate. It enables adjacencies to be discovered and explored, at a company level, at an emerging technology level, and at a national/regional level. This research opens up many additional questions, some of which have been set out here. The authors are keen to explore further opportunities arising from the ISPIM 2023 conference where this work is being presented for the first time.

CHAPTER 6

CONCLUSION

This chapter synthesises the key contributions of the thesis, reflecting on how each study advances the overarching goal of discovering latent knowledge about individuals, enterprises, and technologies. By integrating techniques from machine learning, natural language processing, large language models, and network science, the research demonstrates how heterogeneous data sources can be transformed into meaningful insights for understanding complex innovation ecosystems. In addition to summarising the key findings and contributions, this chapter also discusses the broader implications of the research, highlighting how the results collectively inform our understanding of the interconnections among different domains. Finally, the chapter outlines the limitations of the study and identifies potential directions for future research, emphasising opportunities to further integrate these domains and extend the developed computational approaches to new datasets, contexts, and applications.

6.1 Key Findings and Contributions

In Chapter 2, we examine how founders' personality traits, inferred from large-scale linguistic analysis of social media texts, impact the success trajectories of startups. Using natural language processing and machine learning, the study generates personality profiles for thousands of startup founders, correlating them with firm-level performance indicators, including funding raised, revenue growth, and exit outcomes. The analysis reveals significant associations between specific personality traits and measures of entrepreneurial

success. For instance, higher scores on openness and conscientiousness are associated with better fundraising outcomes, while traits such as emotional stability are correlated with a higher likelihood of long-term survival. The study also identifies interaction effects, where the specific combinations of personality traits of the founder team play a pivotal role in shaping performance, particularly in high-uncertainty environments. Furthermore, industry-specific patterns emerge, with certain personality profiles more strongly aligned with success in technology-intensive sectors compared to service-oriented domains. By integrating individual-level psychological characteristics with organisational and market performance data, the research demonstrates that latent personal attributes can be systematically measured and linked to entrepreneurial outcomes on a large scale. These findings contribute to bridging the gap between personal behavioural attributes and macro-level organisational performance. In doing so, the study not only advances academic understanding of the founder–startup performance relationship but also offers practical insights at the individual level for investors and accelerators seeking to identify and support high-potential start-up founders.

In Chapter 3, we apply harmonic centrality to the Australian web domain network, integrating structural link analysis with firm-level non-structural attributes to reveal latent patterns in enterprises' online presence. Quantitative analysis shows that harmonic centrality effectively differentiates enterprises with strong, diverse online connections from those with weaker or more isolated positions. Higher harmonic centrality scores are significantly associated with greater sectoral diversity, broader geographic engagement, and stronger economic performance in some industries. The results highlight that while large, well-known companies often occupy high-centrality positions, certain smaller enterprises achieve similar network prominence through niche sectoral connectivity or regional bridging roles. Qualitative findings complement these insights by identifying clusters of enterprises that hold strategic positions in the national domain space. These include sector-based hubs, such as technology and education, as well as geographically anchored clusters in major cities and regional centres. The study also uncovers patterns of interlinking between enterprises, revealing pathways of information flow, potential for collaboration, and competitive positioning. Enterprises with high harmonic centrality often act as intermediaries, connecting otherwise disparate network segments and thereby playing a critical role in shaping the digital economic ecosystem. By combining web network metrics with organisational and economic data, the research demonstrates that structural positioning in the online domain space can serve as a proxy for hidden competitive advantages, innovation potential, and influence within the broader economy. These findings illustrate how

network science and link analysis can yield actionable insights at the company level for policymakers, industry leaders, and researchers.

In Chapter 4, we construct a Cosmos 1.0 dataset, which can be read as a large-scale, multidimensional map of the emerging technology frontier, constructed using Wikipedia-derived entity embeddings and temporal dynamics to capture the relationships among 23,000 technology entities. The work identifies a curated set of 100 emerging technologies (ET100) through filtering, noise reduction, and manual validation, enabling a more precise focus on early-stage, high-potential developments. By embedding these entities into a semantic vector space and applying clustering, the study reveals distinct thematic groupings and structural relationships, including both tightly connected technology clusters and more isolated technologies with niche applications. Temporal features reveal varying maturity trajectories, indicating that some technologies are rapidly gaining attention, while others emerge more gradually. The study also demonstrates the adaptability of the Cosmos framework by linking emerging technologies to related concepts, industries, and application domains, thereby enabling richer contextual understanding. Furthermore, the framework enables cross-domain analysis, illustrating how technologies from different fields converge over time, potentially signalling opportunities for interdisciplinary innovation. By making the Cosmos platform interactive, the work supports dynamic exploration of technology landscapes, providing policymakers, researchers, and industry stakeholders a tool to monitor technological change, identify innovation hotspots, and anticipate shifts in the frontier. These findings uncover latent knowledge about technologies using advanced computational methods, illustrating how large-scale knowledge graph embeddings can reveal hidden structures, interconnections, and trends in global technology development.

In Chapter 5, we develop and validate a novel methodology that uses Wikipedia-derived entity embeddings to systematically link 100 emerging technologies with the world's top R&D spending companies. By combining semantic similarity measures with embedding space transformation, the study identifies both expected and previously hidden associations between firms and technologies, offering a dynamic and scalable map of the innovation landscape. The findings include identifying nine distinct clusters and areas of "white space" with lower company and technology density. Highly concentrated clusters, such as Technology & Financial Services and Pharmaceuticals & Biotech, contrast with sparse ones like Telecommunications, where some expected technologies are relatively distant. Zoomed-in analyses reveal variation in the proximity of companies to their closest technologies, with some firms tightly linked to a few technologies while others are connected to more evenly spaced sets. Conversely, some technologies, such as Fullerene and Func-

tional magnetic resonance imaging, are distant from all companies, indicating they are in their nascent stages. In contrast, others, such as fintech, fleet management, and fuel cells, are closely associated with established firms. Focusing on Circular Economy Technologies highlights specific firm-technology pairings, such as Siemens Energy's partnership with carbon capture and storage, and Novozymes' collaboration with biocatalysis. Patent analysis validates nearly half of the firm-technology associations, with strong correlations observed in patent-intensive fields such as medical technologies, while weaker correlations are noted for very new or niche technologies. These findings underscore the model's potential for mapping technological engagement, identifying innovation opportunities, and guiding further investigation into the dynamics of firm-technology relationships. This research not only provides actionable intelligence for innovation management and policy but also demonstrates the value of open, continuously updated data sources for tracking technological change and corporate R&D strategies over time.

6.2 Interconnections and Implications

The four studies presented in this thesis reveal a broader conceptual framework linking individuals, enterprises, and technologies within innovation ecosystems. While each chapter focuses on a distinct level of analysis, a common theme emerges: latent signals embedded in digital traces, such as language and networks, can be computationally extracted to reveal hidden structures that shape economic and technological outcomes.

Chapter 2 begins at the individual level, demonstrating that founders' personality traits can be inferred from linguistic patterns and systematically associated with startup success. Importantly, the results also suggest that personality patterns between founders and employees are distinctive and combinations of personality traits within founder teams may shape organisational trajectories. These findings have practical implications for venture capital firms, startup accelerators, and entrepreneurial support programs, which could incorporate personality-based assessments derived from founders' public communications to better evaluate team dynamics and leadership potential. Such insights may also assist founders themselves in building more complementary teams, improving decision-making diversity and resilience during early-stage venture development. This theoretical framework is currently being operationalised within the Japanese Black Belt Global Venture Studio initiative, where personality-informed venture design and team formation strategies are being explored in applied startup development contexts.

Chapter 3 shifts the analytical lens from individuals to enterprises, examining how or-

ganisations are positioned within the digital economy through the structure of the Australian web domain network. The analysis shows that harmonic centrality captures meaningful differences in firms' structural connectivity, revealing patterns of multiple aspects, such as sectoral influence and regional clustering. Considering Chapter 2, there is a potential pathway linking founder characteristics to organisational outcomes. Founder personalities may plausibly influence how firms build networks, establish partnerships, and position themselves within broader digital ecosystems. Although this thesis does not empirically test such causal relationships, the findings point to a promising line of inquiry connecting individual-level attributes with organisational network structures.

Chapters 4 and 5 extend the analysis to the technological domain. Chapter 4 constructs the Cosmos 1.0 dataset, mapping the global emerging technology landscape through embedding based representations of over 23,000 technology-adjacent entities. By providing a transparent and reproducible dataset, Cosmos 1.0 enables researchers and policymakers to systematically monitor emerging technology trends, conduct comparative innovation studies, and build downstream applications such as technology foresight tools and national innovation policy analyses. The resulting ET100 reveals clusters of emerging technologies and areas of convergence across scientific and industrial domains. Chapter 5 then connects these technologies to the world's leading R&D firms, identifying innovation clustering and areas of technological "white space". This mapping of the firm-technology framework provides a practical approach for potential users to identify strategic technology–industry alignments and monitor how corporate R&D activity evolves in relation to frontier technologies. More broadly, the approach enables scalable monitoring of global innovation ecosystems by revealing latent relationships between firms and technological domains that may not be visible through traditional classification systems or industry taxonomies.

Overall, the studies suggest a multi-layered perspective on innovation systems. Individuals influence a startup's success their personality and the composition of the founder team. Enterprises shape technological trajectories and economic performance through R&D investment decisions and network relationships in the digital ecosystem. Emerging technologies then reshape organisational opportunities and competitive landscapes. By uncovering latent signals across these three levels, the thesis contributes not only individual empirical findings but also a broader conceptual pathway for understanding how the latent attributes of individuals, enterprises, and technologies interact to shape innovation systems.

6.3 Limitations of the Study

In this thesis, several limitations should be acknowledged.

First, the research relies extensively on digital trace data, including social media text, web domain networks, Wikipedia-derived embeddings, and Crunchbase records. Although these sources enable large-scale analysis and capture rich behavioural signals, they may also reflect biases inherent in digital platforms. Certain regions, industries, or users may be underrepresented due to limited online presence or uneven digital participation. As a result, the inferred relationships may not fully capture the offline dynamics of innovation.

Second, the computational methods in this work rely on probabilistic inference and representation learning techniques. Personality inference from linguistic data, for example, captures behavioural signals expressed through social media texts but cannot fully represent the complexity of human psychological traits. Similarly, embedding-based representations of technologies and firms capture semantic proximity rather than direct causal relationships. While robustness checks and validation analyses were performed, these models should be interpreted as analytical approximations rather than definitive measures of underlying phenomena.

Third, the empirical analyses across chapters are largely correlational. They identify statistically significant associations between founder personality traits and startup success, between network centrality and organisational characteristics, and between firms and emerging technologies. However, these relationships do not establish causal mechanisms. Future research could address this limitation by incorporating longitudinal designs, experimental methods, or quasi-natural experiments to better isolate causal effects.

Fourth, although we advance a conceptual framework linking individuals, enterprises, and technologies, the studies remain empirically separate due to the thesis-by-publication structure. Each chapter draws on distinct datasets and methodological approaches, which limits the ability to test cross-level hypotheses directly. For instance, the potential relationship between founder personality traits and organisational network centrality, or between entrepreneurial characteristics and technological exploration, remains an open question for future work. Integrating these levels into a unified multi-layer dataset or embedding space would represent a significant step forward in understanding the dynamics of innovation systems.

Finally, the breadth of the research, spanning psychology, entrepreneurship, network science, and technology forecasting, poses challenges for communicating findings across disciplinary boundaries. While we demonstrate the potential of computational methods to

generate new insights, translating these insights into practical frameworks for policymakers, investors, researchers and industry practitioners requires further refinement.

These limitations represent opportunities to deepen the research and strengthen the empirical integration of the framework in future work.

6.4 Future Research

Beyond the text-derived personality profiles presented in Chapter 2, future research could incorporate founders' social and professional networks, investment histories, and prior entrepreneurial experiences, thereby enabling a deeper understanding of how personality interacts with social capital and past performance. Machine learning models could be applied to link these comprehensive profiles with startup performance metrics such as funding success, revenue growth, and survival rates over time. Additionally, extending the analysis beyond startups to include public and private firms of varying sizes and across diverse sectors would reveal whether relationships between personality and performance are consistent or sector-specific. Longitudinal studies could track changes in personality expression and leadership behaviour across different business stages, identifying patterns linked to resilience, innovation adoption, and crisis management. Advances in large language models could further enhance personality inference from unstructured text, while explainable AI techniques could aid in uncovering the mechanisms underlying observed correlations.

For Chapter 3, future research could extend the application of harmonic centrality in the Australian web domain network by incorporating longitudinal web-crawl data to examine how enterprises' online network positions evolve and how these changes relate to innovation, market expansion, and resilience (Newman, 2018; Pósfai & Barabási, 2016). Integrating additional datasets, such as social media activity, could reveal deeper connections between online structural positioning and offline economic impact (Kitchin, 2014). Advances in large language models could be employed to semantically enrich link data, allowing for the classification of inter-domain connections by type (e.g., partnership, supply chain, informational), thereby enabling more nuanced interpretation of enterprise networks (Bommasani et al., 2021). Cross-country comparisons using similar methodologies could uncover how national digital ecosystems differ in their structural properties and innovation capacity (Castells, 2011). Finally, scenario modelling could be developed to simulate how targeted policy interventions, such as enhancing connectivity for strategically positioned but underconnected enterprises, might improve the robustness and inclusive-

ness of the national digital economy.

For Chapter 4, future work on Cosmos 1.0 could focus on integrating multimodal data sources, such as pictures and videos, to enrich the semantic and temporal mapping of emerging technologies. Incorporating dynamic graph embedding methods could enable continuous tracking of evolving technology relationships, offering predictive foresight into potential convergence areas and disruptive innovations (W. Hamilton et al., 2017; Zhong et al., 2023). Advances in large language models could enhance entity disambiguation and contextual linking across heterogeneous datasets, thereby improving the accuracy and interpretability of technology identification. Additionally, it is possible to extend the current technology mapping by introducing firms and individuals into the same embedding space, enabling simultaneous exploration of the interconnections among emerging technologies, enterprises, and key actors. By aligning heterogeneous entities in a unified semantic and temporal vector space, this approach allows a comprehensive analysis of how specific companies and individuals interact with, influence, and adapt to technological change.

For Chapter 5, future work could extend the ET100/R&D Links model by integrating individuals, enterprises and technologies into a unified embedding space to capture a more comprehensive view of individual-firm-technology relationships. Incorporating temporal modelling techniques could enable the tracking of evolving associations among individuals, firms, and emerging technologies, offering predictive insights into technology adoption trajectories and competitive positioning (Milojević, 2015). Advances in large language models (e.g., OpenAI GPT-5, Anthropic Claude) could also be leveraged to improve entity disambiguation, semantic similarity measurement, and contextual mapping between heterogeneous datasets (Bommasani et al., 2021). Furthermore, combining the model with network science approaches could help identify critical nodes and clusters in global innovation ecosystems, revealing potential innovation hubs and "white space" opportunities (Newman, 2018). Linking these individual-firm-technology networks to market performance and societal impact metrics would further align the thesis's goal.

From an integrated perspective, several promising directions for connecting the above elements emerge. For example, investigating whether founder personality profiles predict organisational network centrality or engagement with specific emerging technology clusters. Another direction would examine whether organisations occupying central network positions are more likely to participate in emerging technology ecosystems or move across technological domains over time. Finally, a further research direction involves representing individuals, enterprises, and technologies within a unified embedding space, enabling the systematic exploration of latent relationships across these three levels of the innovation

system. Such a unified representation could support new forms of computational analysis that trace how founder characteristics influence organisational positioning and technological exploration, thereby revealing deeper interconnections within innovation ecosystems.

APPENDIX A - THE IMPACT OF FOUNDER PERSONALITIES ON STARTUP SUCCESS

A.1 Industry- and market-based drivers of startups success

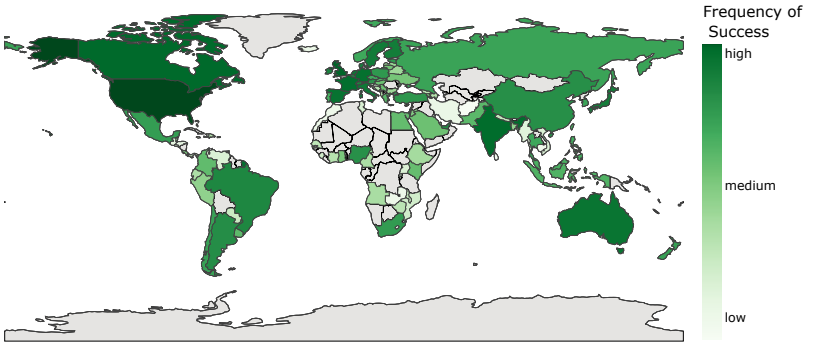
The first lens we looked through was at the firm-level. Much of the previous literature on startups has been focused on firm-level or external factors and their influence on success. Startup success has been shown to relate to how much capital the startup has raised, how old it is and what industry it is in, among other things.

We, therefore, firstly examine a range of firm-level determinants of startup success, including location (Extended Data Figure A.1A), industry (Extended Data Figure A.1B) and age of the startup (Extended Data Figure A.1C) to explore to what extent these factors are associated with success. We find that startup success is influenced strongly by its location (firms from Japan, Scandinavia, USA, France, and Germany are more likely to be successful than those from Turkey, Argentina, Mexico or other countries); industry (firms in Payment Systems and Privacy & Security are most successful) and a company's age.

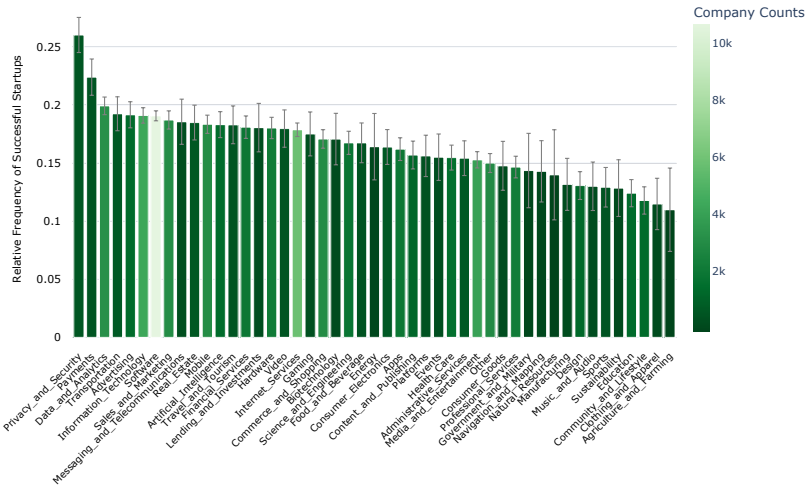
These findings are used in the multifactor analysis shown in more detail below. The predictive capability of the models that also include personality trait information is compared to those that use only firm-level information.

A.2 Entrepreneurs versus Employees

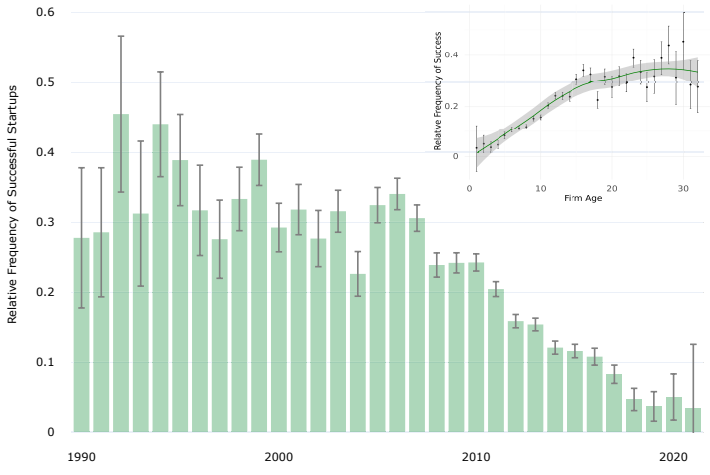
In addition to our main analysis, we compare the personality traits of entrepreneurs with a reference population of employees. We developed an Entrepreneurial Occupational Index (EOI; see Extended Data Figure A.2) based on LinkedIn data that looks at the percentage of people currently employed in that role worldwide and who also hold or have previously held the position of founder or co-founder. We found EOI values for each of the 624 Occupations we have personality profiles for. We ranked the occupations from most entrepreneurial (public speaker 21.11%, chief technology officer 20.75%, and creative director



A



B



C

Extended Data Figure A.1: **Firm-Level Factors of Startup Success.** (**A**), On a country level, chances for success are highest in the US, Japan, West Europe, and Scandinavian countries. (**B**), Firms from the payment and software industries have high chances of success. (**C**), Chances of success are positively related to a firm's maturity, with firms that are seven years or older having higher chances of success.

19.33%) to least entrepreneurial (cashier 0.02%, palaeontologist 0.00%, furniture removalist 0.00%, aged carer 0.00%, bacteriologist 0.00%).

We then created a list of low EOI occupations (n=112), each of which had less than 0.5% of individuals who also held the titles founder or co-founders in their LinkedIn Profile. People in these roles may still be founders and co-founders, but it is unlikely that they are. Any individual in even the most entrepreneurial of these 112 occupations (internal auditor) is still five times less likely also to be a founder or co-founder than the global average (2.5%) across all 624 occupations. From our previous study, we randomly selected a sample of employees (n=6k) for whom we have inferred personality data and who are unlikely to be entrepreneurs as they are drawn from the 112 low EOI occupations. These employees are Twitter users selected as the first twenty-five search-ranked Twitter users with their occupation specified as part of their biotext. The occupations were selected from standard job titles (O*NET and ANZSCO) and represent a wide cross-section of jobs at differing skill levels and in different industries. De-identified data on employees and their occupations can be requested from the authors.

Using the two samples together (entrepreneurs and employees who are unlikely to be founders), we trained and tested a machine-learning random forest classifier to distinguish and classify entrepreneurs from employees and vice-versa using inferred personality vectors alone. As a result, we could correctly predict entrepreneurs with 77% accuracy and employees with 88% accuracy (Figure 2.1A in the main text). Thus, based on personality information alone, we correctly predict all unseen new samples with 82.5% accuracy (See Extended Data Figure A.3 for details on modelling and prediction accuracy).

We explored in greater detail which personality features are most prominent among successful entrepreneurs. We found that the subdomain or facet of Adventurousness within the Big Five Domain of Openness was significant and had the largest effect size. This is consistent with previous research that found higher values in the personality trait Openness significantly predict VC financing even after accounting for observable founder and firm characteristics and openness to be the key Big Five Domain that distinguishes entrepreneurs from non-entrepreneurs. The facet of Modesty within the Big Five Domain of Agreeableness and Activity Level within the Big Five Domain of Extraversion was the subsequent most considerable effect (Figure 2.1B in the main text). All thirty dimensions of the Big Five facet were found to be significantly different in their distribution, with ten features having large effect sizes. (See Extended Data Figure A.4 for more details on Cohen's D analysis with a complete list of features and their effect sizes, and Extended Data Figure A.5 for Big5 personality facets of employees and entrepreneurs visualised as a heatmap and

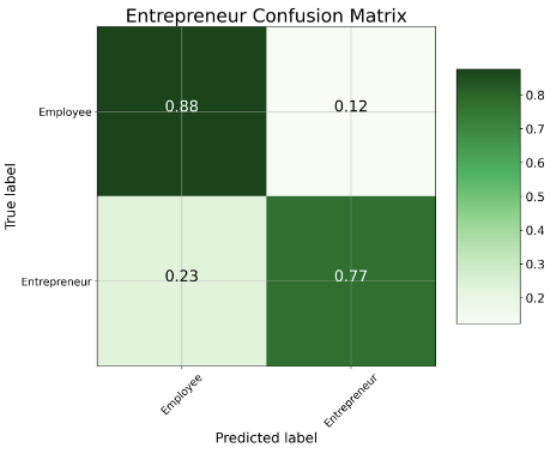
Entrepeneurial Occupations Index
Frequency of people by Occupation
Occupations (n=624)

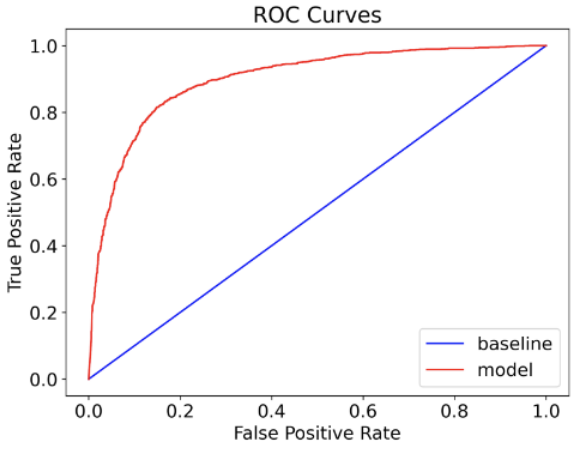
Occupation		Global Count (LinkedIn)	Global with founder or cofounder	Global Index
PUBLIC SPEAKER	Listeners	18,000	3,800	21.11%
CHIEF TECHNOLOGY OFFICER	Leaders	106,000	22,000	20.75%
CREATIVE DIRECTOR	Listeners	388,000	75,000	19.33%
EXECUTIVE PRODUCER	Listeners	83,000	14,000	16.87%
ARTISTIC DIRECTOR	Rebels	45,000	7,300	16.22%
COLUMNIST	Listeners	27,000	3,900	14.44%
ECONOMIC HISTORIAN	Leaders	21	3	14.29%
CHIEF EXECUTIVE	Leaders	1,080,000	149,000	13.80%
AUTHOR	Rebels	273,000	37,000	13.55%
CHOCOLATE MAKER	Listeners	488	66	13.52%
COMMUNITY ARTIST	Rebels	3,300	439	13.30%
MOTIVATIONAL SPEAKER	Leaders	24,000	3,100	12.92%
EXECUTIVE DIRECTOR	Leaders	924,000	108,000	11.69%
CHIEF OPERATING OFFICER	Accomplishers	209,000	23,000	11.00%
PRODUCT DESIGNER	Listeners	101,000	11,000	10.89%
FILM PRODUCER	Rebels	22,000	2,300	10.45%
MEDIA PRODUCER	Listeners	28,000	2,900	10.36%
ORGANISATIONAL PSYCHOLOGIST	Leaders	890	92	10.34%
MARKETING CONSULTANT	Accomplishers	254,000	26,000	10.24%
PLAYWRIGHT	Rebels	6,800	688	10.12%
BOOK EDITOR	Rebels	6,900	671	9.72%
FILM DIRECTOR	Rebels	38,000	3,500	9.21%
MAGAZINE EDITOR	Listeners	10,000	898	8.98%
SPORT PSYCHOLOGIST	Leaders	1,100	96	8.73%
MANAGING DIRECTOR	Leaders	1,990,000	170,000	8.54%
VLOGGER	Observers	4,100	350	8.54%
ART DIRECTOR	Listeners	201,000	17,000	8.46%
CRITIC	Rebels	8,400	689	8.20%
INSTRUMENTALIST	Observers	2,500	205	8.20%

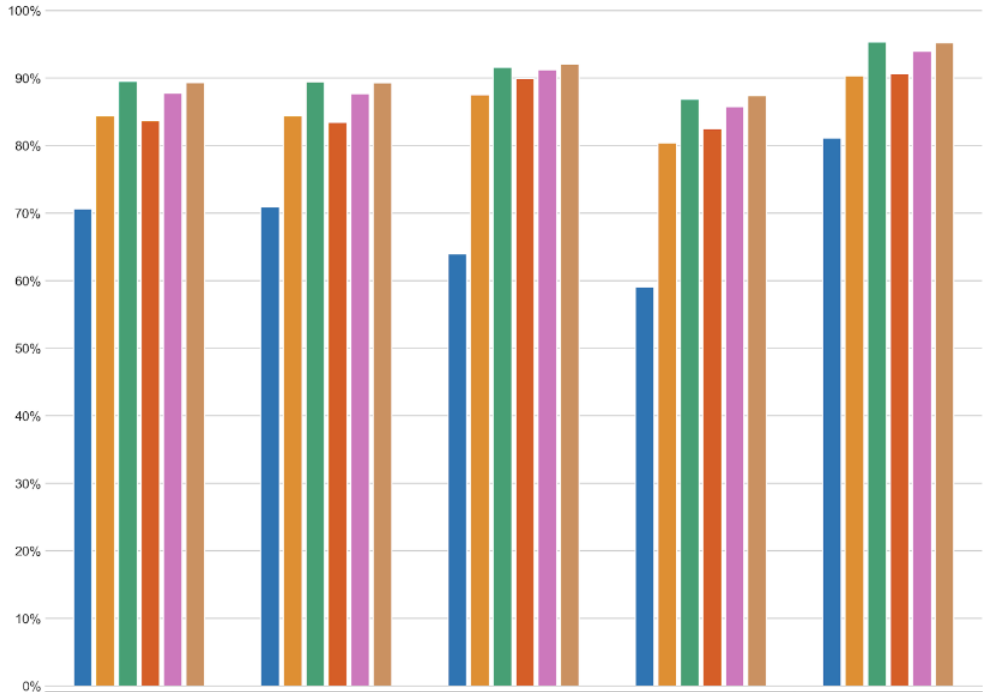
Low EOI Roles
Less than 0.5% Global EOI index & > 100 count in Sample
(Note all Occupation Global EOI = 2.5%)
Occupations (n=112)

Occupation		Global Count (LinkedIn)	Global Index	In Sample Count (LinkedIn)
INTERNAL AUDITOR	Leaders	117,000	0.50%	945
UPHOLSTERER	Fighters	4,700	0.49%	239
PETROLEUM ENGINEER	Fighters	18,000	0.49%	249
PROGRAMMER ANALYST	Fighters	122,000	0.49%	3,100
ANALYST PROGRAMMER	Fighters	122,000	0.49%	3200
SALESPERSON	Fighters	371,000	0.49%	9,800
DETECTIVE	Listeners	51,000	0.48%	1,100
MERCHANDISER	Umpires	313,000	0.48%	4,500
CONTRACT ADMINISTRATOR	Accomplishers	38,000	0.48%	5,400
PARK RANGER	Listeners	8,800	0.48%	363
NAIL TECHNICIAN	Observers	17,000	0.48%	436
RETAIL STORE MANAGER	Accomplishers	51,000	0.47%	2,600
REFERENCE LIBRARIAN	Listeners	8,200	0.46%	137
BOOKKEEPER	Accomplishers	261,000	0.46%	13,000
CONVEYANCER	Accomplishers	7,400	0.46%	2,200
BARISTA	Accomplishers	286,000	0.45%	10,000
CHILD CARE WORKER	Observers	4,800	0.44%	539
COURIER	Accomplishers	67,000	0.44%	2,500
VETERINARY ASSISTANT	Observers	31,000	0.44%	244
DENTAL TECHNICIAN	Fighters	19,000	0.43%	737
CABIN CREW	Umpires	55,000	0.42%	1,500
FLIGHT NURSE	Experts	4,500	0.42%	187
ZOOLOGIST	Rebels	1,200	0.42%	184
BUILDING SURVEYOR	Fighters	13,000	0.42%	1,300
NURSE MANAGER	Leaders	94,000	0.41%	6,700
PURCHASING OFFICER	Accomplishers	25,000	0.40%	3,500
AMBULANCE PARAMEDIC	Leaders	1,000	0.40%	352
ELECTRICAL CONTRACTOR	Accomplishers	17,000	0.40%	1,700

Extended Data Figure A.2: **Entrepreneurial Occupations Index.** To identify a group of individuals who are unlikely to be founders, we rank occupations by the share of founders based on LinkedIn data; occupations with a share of founders smaller than 0.5% are assumed to represent individuals unlikely to be entrepreneurs, that is, they are considered as representing 'employee' personalities.







	Accuracy	F1 Score	Specificity	PPV	AUROC
Naive Bayes	71%	71%	64%	59%	81%
Elastic Net	84%	84%	88%	80%	90%
SVM	89%	89%	92%	87%	95%
Random Forest	84%	83%	90%	82%	91%
Gradient Boosting	88%	88%	91%	86%	94%
Stacked Ensemble	89%	89%	92%	87%	95%

Extended Data Figure A.3: **Entrepreneurial Prediction Performance.** The ROC score of the train set is 0.94, while the score of the test set is 0.9. The confusion matrix also demonstrates high accuracy. On the test set, the model could 88% correctly predict the entrepreneurs and 77% correctly predict the employees.

dendrogram).

Adventurousness in the Big Five framework is defined as the preference for variety, novelty and starting new things - which are consistent with the role of a startup founder whose role, especially in the early life of the company, is to explore things that do not scale easily and is about developing and testing new products, services and business models with the market.

A.3 Clustering of founder personalities

In this part of the Supplementary Information, we provide additional statistics and analyses on the empirical clustering of founder personalities. Extended Data Figure A.6 shows the clustering tendency of the founder personality data. Extended Data Figure A.7 shows the dendrogram of the hierarchical clustering algorithm that was used to identify the groups in the dataset. Extended Data Figure A.8 presents statistics on the optimal number of clusters in the data and the frequency count of founders in each cluster. Extended Data Figure A.9 lists the personality attributes of the occupation tribes corresponding most closely to the founder personality clusters. Extended Data Figure A.11 investigates the robustness of the founder personality clusters against resampling.

To better understand the unique personality characteristics of each of the six different clusters of founders and co-founders we:

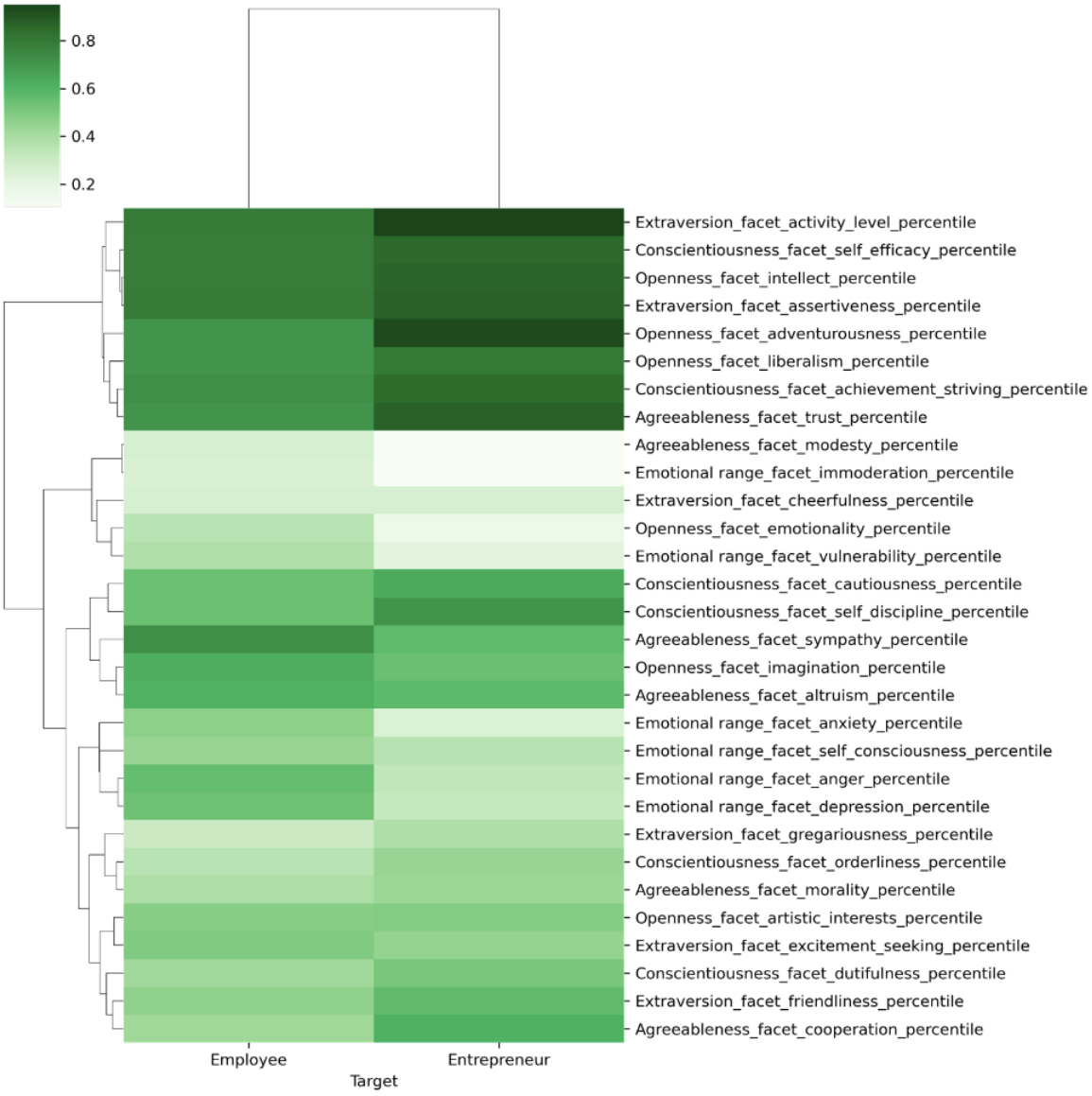
1. **Analysed the personality footprints of each cluster.** We examined the distinctive personality traits of each group. We identified which clusters were home to the maximums in each of the 30 personality facets. We created a heat map revealing the complete personality footprint of each of the six types (Figure 2.1D of the main text).

2. **Matched the occupation closest to the centre of each cluster** using the personality-occupation matrix from our previous research. For each founder and co-founder, we found the closest corresponding occupation tribe for each based on personality similarity.

3. **Identified which of the eight occupation-tribes from previous research each founder or co-founder belonged to.** Leveraging previous research, we then looked at the distribution of tribe membership of each founder within each cluster. Then we tallied the founders within each cluster by tribe to reveal the level of coherence or the extent to which most founders within each group belonged to one occupation tribe. This

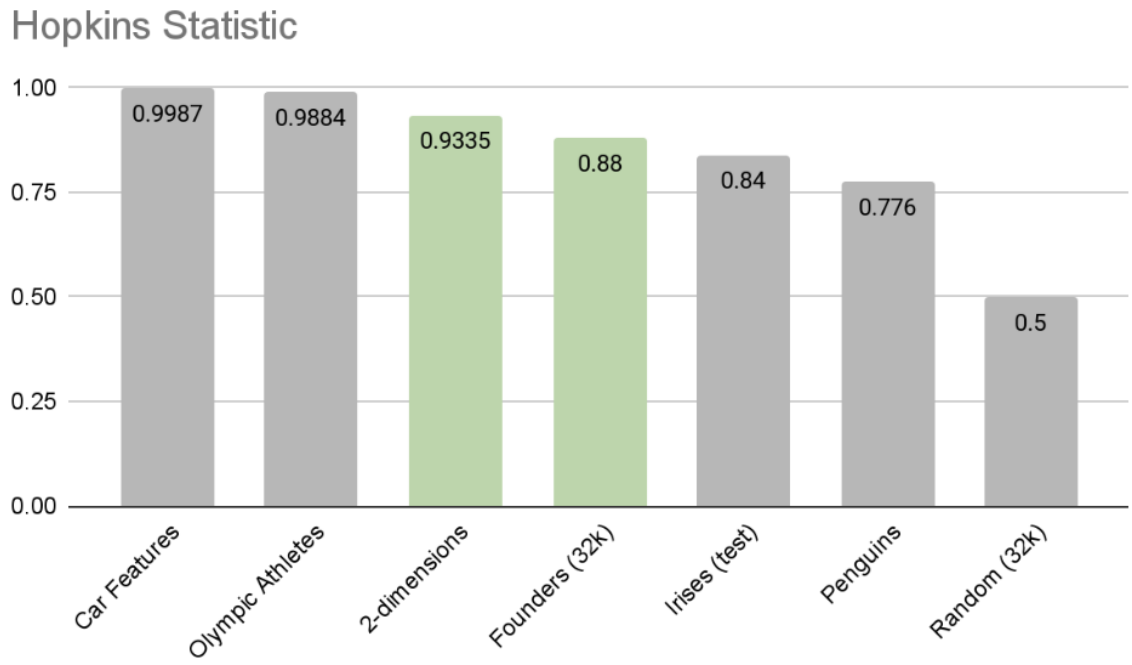
Big 5 Personality Facets	Cohen's D	Effect Size	p Value
Openness (adventurousness)	0.92	Large	0.00
Agreeableness (modesty)	-0.79	Large	0.00
Extraversion (activity_level)	0.78	Large	0.00
Emotional range (anxiety)	-0.77	Large	0.00
Emotional range (immoderation)	-0.73	Large	0.00
Agreeableness (trust)	0.72	Large	0.00
Emotional range (anger)	-0.68	Large	0.00
Emotional range (depression)	-0.68	Large	0.00
Agreeableness (cooperation)	0.67	Large	0.00
Openness (emotionality)	-0.67	Large	0.00
Emotional range (vulnerability)	-0.63	Medium	0.00
Conscientiousness (achievement_striving)	0.59	Medium	0.00
Conscientiousness (self_discipline)	0.55	Medium	0.00
Conscientiousness (cautiousness)	0.50	Medium	0.00
Openness (intellect)	0.45	Medium	0.00
Conscientiousness (self_efficacy)	0.43	Medium	0.00
Extraversion (assertiveness)	0.43	Medium	0.00
Openness (liberalism)	0.43	Medium	0.00
Agreeableness (sympathy)	-0.43	Medium	0.00
Conscientiousness (dutifulness)	0.41	Medium	0.00
Conscientiousness (orderliness)	0.31	Small	0.00
Extraversion (friendliness)	0.24	Small	0.00
Extraversion (excitement_seeking)	-0.21	Small	0.00
Emotional range (self_consciousness)	-0.16	Trivial	0.00
Openness (imagination)	-0.15	Trivial	0.00
Extraversion (gregariousness)	0.15	Trivial	0.00
Agreeableness (morality)	0.12	Trivial	0.00
Extraversion (cheerfulness)	-0.08	Trivial	0.00
Openness (artistic_interests)	0.04	Trivial	0.00
Agreeableness (altruism)	-0.04	Trivial	0.01

Extended Data Figure A.4: **Founder Facet Footprint.** Features that distinguish successful founders (n=4.4k) from successful employees (n=6k). Note all 30 Big 5 facets are significant at $p < 0.05$; however, Artistic Interests and Altruism are not significant at $p < 0.01$.



Extended Data Figure A.5: **Entrepreneurial Personality Facets.** The heatmap visualises the median values of 30 facets of two samples, which intuitively and comprehensively compare all personality traits and patterns. From the heatmap, it could be observed that the difference in median values of the two groups on adventurousness (Openness), activity level (Extraversion) and modesty (Agreeableness) are relatively large.

mapping to occupation tribes forms the basis for labelling the founder personality clusters #0 to #5 as Accomplishers, Engineers, Leaders, Developers, Operators, and Fighters.

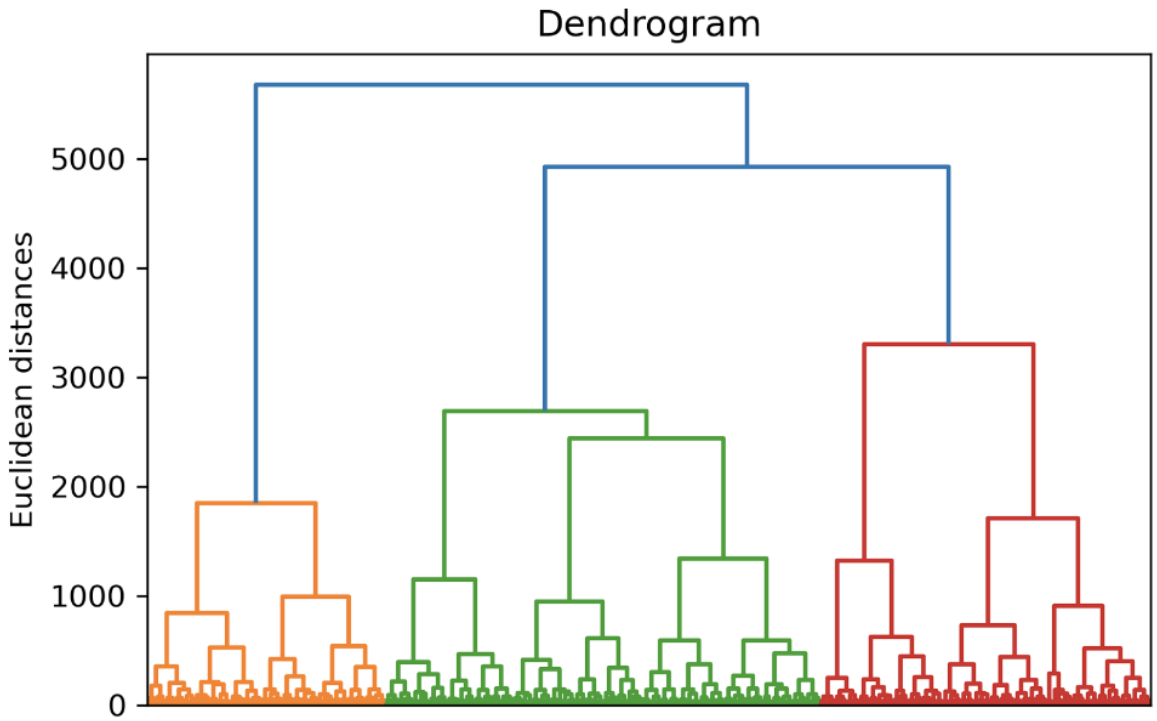


Extended Data Figure A.6: **Clustering Tendency Startup Founders by Personality Features.** Hopkins index (above) measures of the Founders inferred personality traits data, and 2-dimensions data (dimensionality reduction of Founders data) compares favourably to other famous test data sets (Irises; Penguins, Olympic Athletes) that are known to have explicit classes in the data — different breeds of Penguins, various species of Iris flowers and Olympic athletes who have qualified for different categories of events.

A.4 Roles within startups

Information about personality traits not only helps to distinguish between individuals who tend to be founders of startup companies and employees, but it also correlates with the job role that founders will take in the startup companies they establish. For example, Extended Data Figure A.12 shows the distribution of the six founder personality clusters by eight typical job roles in startup companies.

Two personality types are most clearly related to particular job roles. Accomplishers (#0) cluster in the roles of Chief Executive Officer, Chief Financial Officer, Chief Operating



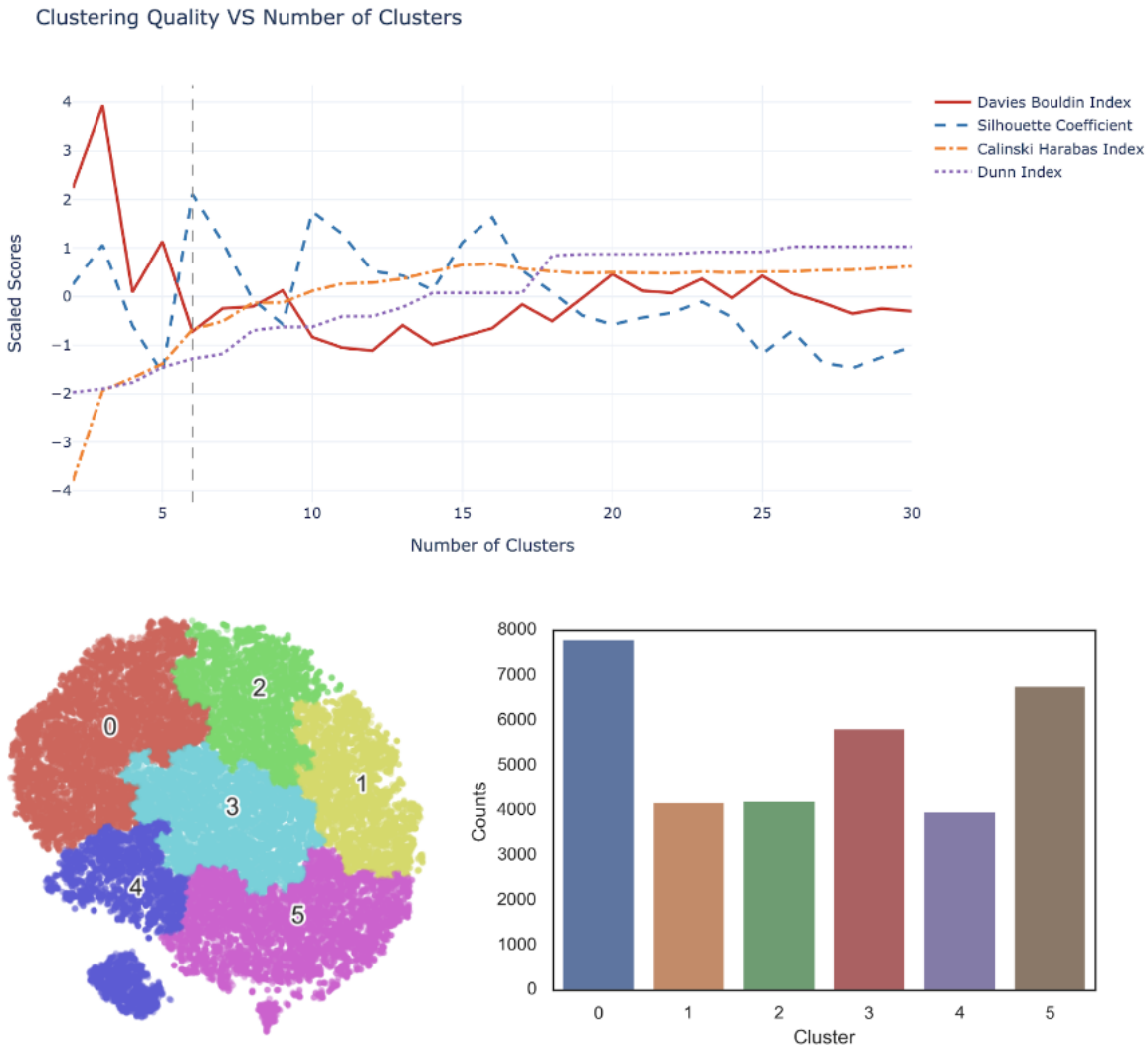
Extended Data Figure A.7: **Hierarchical Clustering of Startup Founders by Personality Features.** Hierarchical clustering was used to model potential clusters of startup founders by their personality features.

Officer, and Chief Marketing Officer. In contrast, Fighters (#5) are most prevalent in the roles of Chief Creative Officer, Chief Product Officer and Chief Technology Officer.

A.5 Data pre-processing for startup success prediction

To use the founder data from Crunchbase described above (32k profiles) for the success prediction, we needed to conduct several preparatory steps, which led to a reduction of the final number of observations as outlined in Extended Data Figure A.13.

The aim is to create a company-founder panel from the Crunchbase data based on exact founder names and company URLs as identifiers. Starting with 32,727 profiles corresponding to 23,292 companies, we removed organisations without names, reducing the data set to 27,181 founders and 23,290 companies. As a next step, we kept only those founders in the data set, founding the 23,290 companies in the data (via the 'founders' column), yielding 25,341 founders and 21,351 companies. Merging these founders with the companies led to



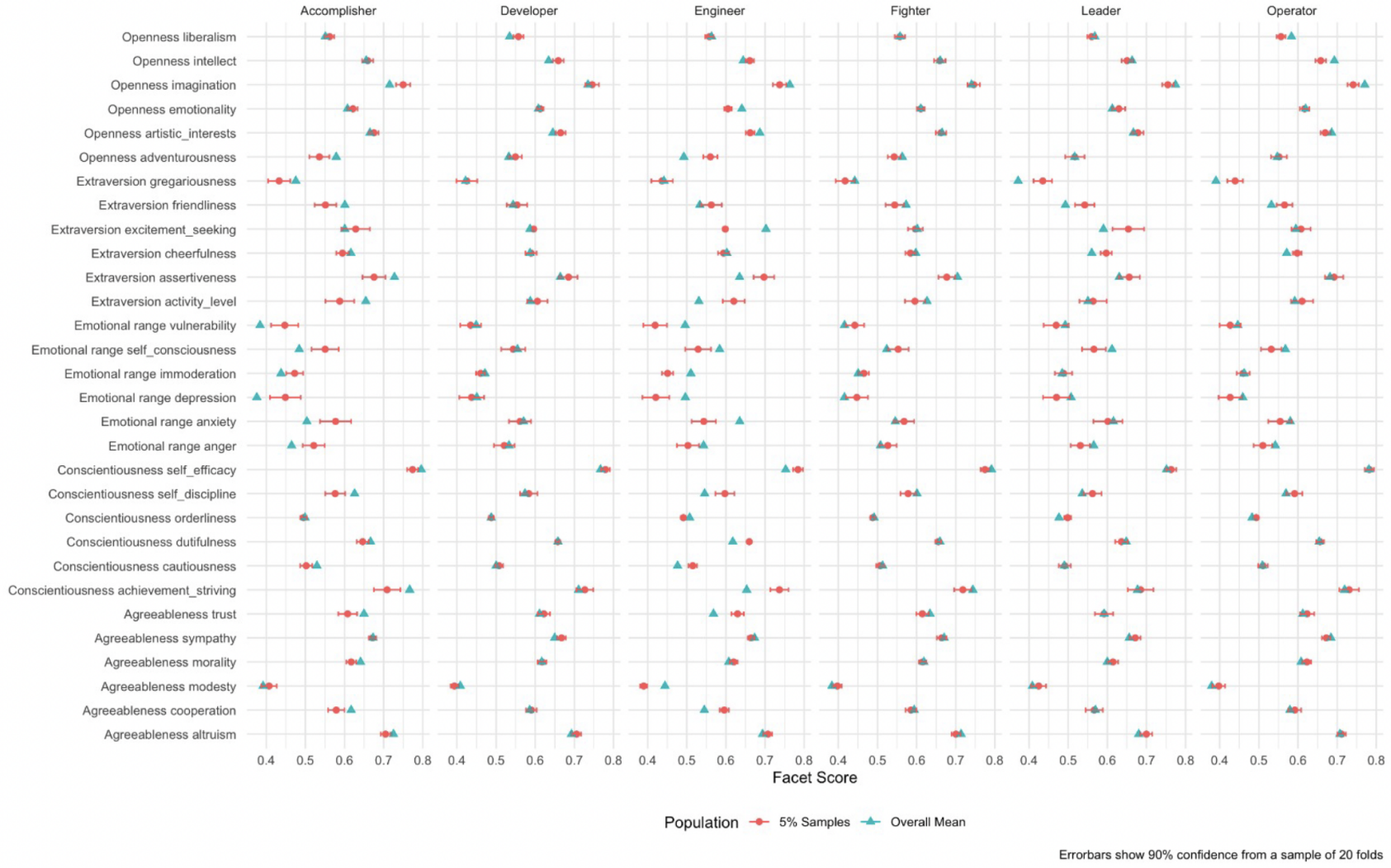
Extended Data Figure A.8: **Optimum Number of Clusters of Types of Founders.** Four different clustering quality indices were used to determine that there are optimally six clusters of startup founders. We labelled each of these #0, #1, #2, #3, #4 and #5 and produced the dendrogram, bar charts, and T-SNE plot above to demonstrate the hierarchy, adjacencies and distribution of all six clusters.

The Eight "Tribes"of Occupations	
Tribe	**Key personality attributes.**
Leaders	Adventurous, persistent, dispassionate, assertive, self-controlled, calm under pressure, philosophical, excitement-seeking & confident.
Listeners	Empathetic and compassionate - can understand others feelings, needs and suffering.
Umpires	Bold, altruistic, respectful of authority, outgoing & uncompromising. Pay attention to the fine details.
Rebels	Authority-challenging, imaginative and appreciative of art.
Experts	Curious, reserved, independent, trusting of others, self-assured, organised, carefree, deliberate, driven & energetic.
Observers	Compassionate, modest, consistent, cheerful, sociable, emotionally aware, laid-back, content
Accomplishers	Super organised & outgoing. Confident, Down-to-earth, Content, Accommodating, Mild-tempered & Self-assured.
Fighters	Spontaneous and impulsive. Tough, sceptical, and uncompromising.

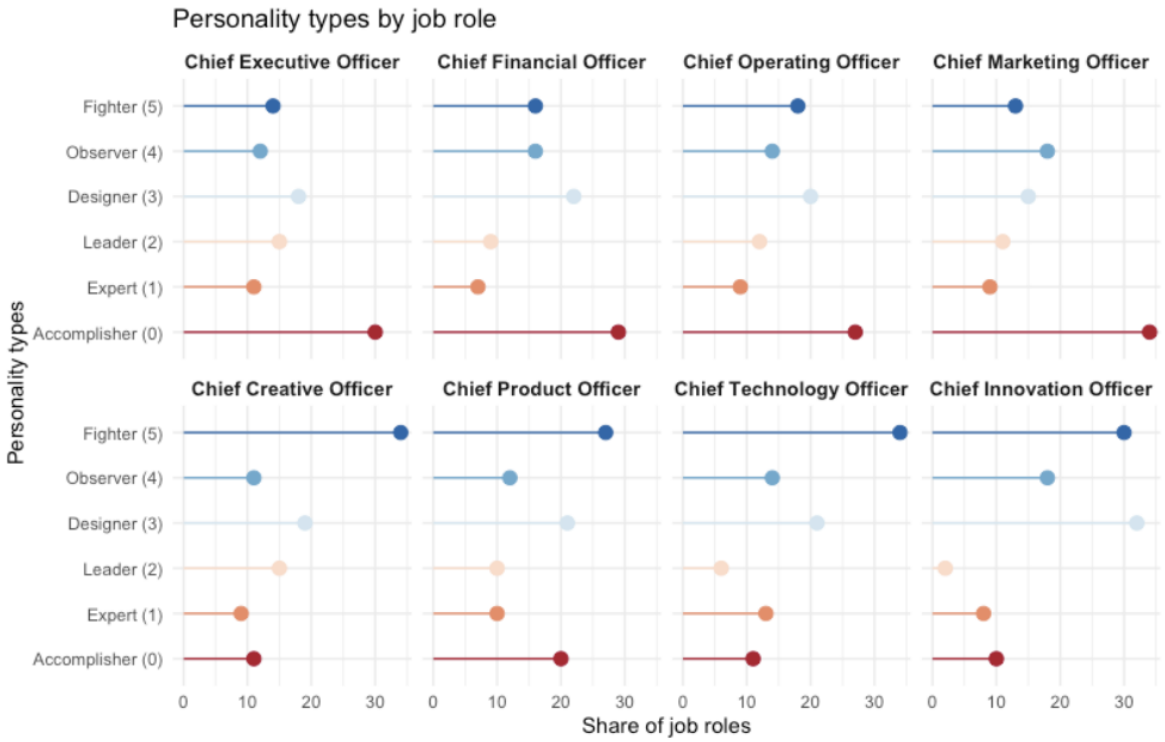
Extended Data Figure A.9: **Eight 'occupation tribes'.** Tribes identified by McCarthy et al., 2022 to empirically describe the fit between occupations and personality types.

Founder Type	**Distinctive Personality Attributes**	**Closest Occupation Fit** From occupation-personality maps
Fighters	Spontaneous and impulsive, tough, skeptical, and uncompromising.	Software Developer Computer Engineer
Operators	Highest in conscientiousness in the facet of orderliness and high agreeableness in the facet of humility for founders in this cluster.	Bicycle Mechanic, Mechanic and Service Manager.
Accomplishers	Organised & outgoing. confident, down-to-earth, content, accommodating, mild-tempered & self-assured.	Chief Information Officer Export Manager
Leaders	Adventurous, persistent, dispassionate, assertive, self-controlled, calm under pressure, philosophical, excitement-seeking & confident.	Executive Director Medical Director
Engineers	Highest in openness in the facets of imagination and intellect.	Materials Engineer and Chemical Engineer.
Developers	"Middle child" cluster — no facets are maximums or minimums but it shares characteristics similar to fighters but higher in extraversion.	Application Developer and related technology roles such as Business Systems Analyst and Product Manager.

Extended Data Figure A.10: **Details on the typology of founder personalities.** The Ensemble theory of Startup Success shows founders come in six types: Fighters, Operators, Accomplishers, Leaders, Engineers and Developers (FOALED), with each founder type having its own distinct personality facet footprint and closest occupation fit.



Extended Data Figure A.11: **Robustness of personality facet clustering against resampling.** The figure compares the overall mean value per facet in each cluster (green triangles) with the distribution of a 20-fold resampling of smaller samples from the founder data (red dot-whiskers) - in many cases the overall mean and the majority of the resampled means overlap; there are some facets with no overlap, but deviations tend to not be bigger than 0.1 points on the facet score scale.

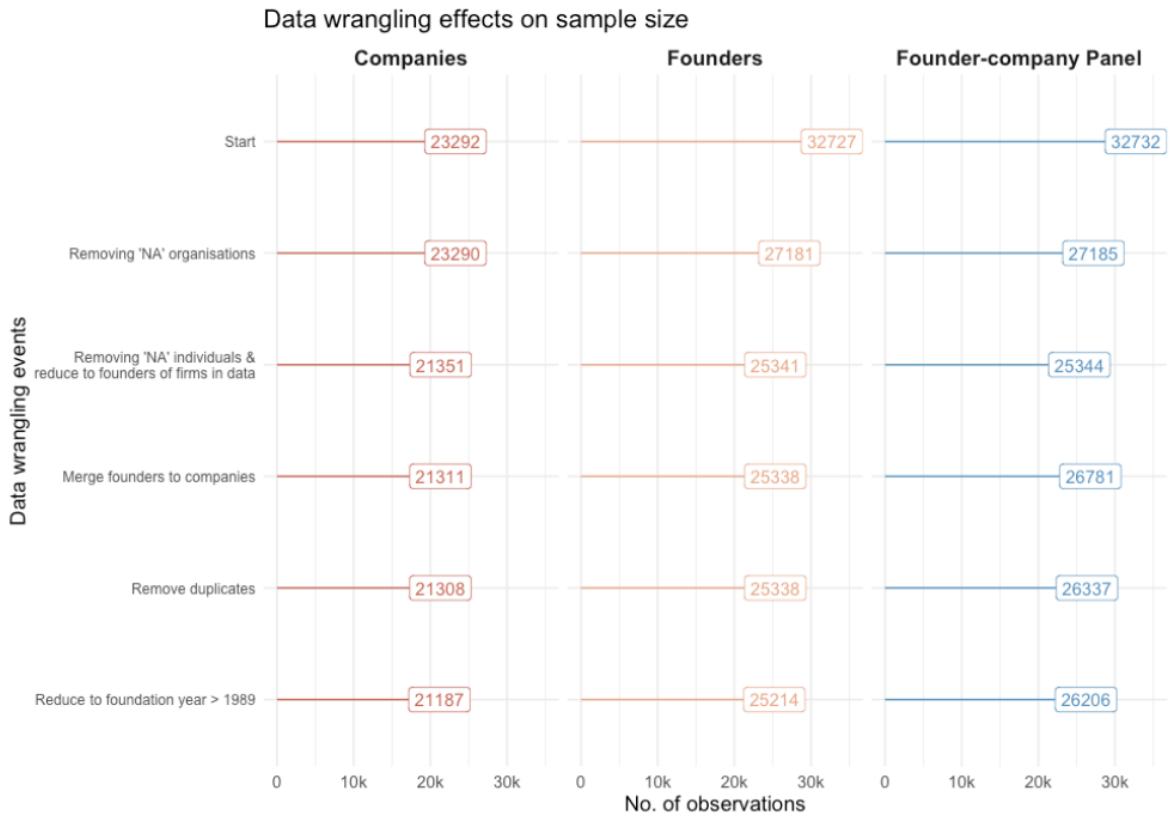


Extended Data Figure A.12: **Occupations within Startups.** While Accomplishers are often CEOs, CFOs or COOs, Fighters tend to be CTOs, CPOs, CCOs.

a further reduction of the data set to 25,338 founders and 21,311 companies. The merging also resulted in some duplicates because of the identical names of some founders. These duplicates were removed by keeping only those company-founder combinations where the company of each potentially duplicated founder was mentioned either as their primary organisation or in their biography. This step did not affect the number of founders but reduced the number of companies by three, which could not unambiguously be assigned to any individual. As the last step, we removed companies that were founded before 1990, leading to a final data set of 25,214 founders involved in the foundation of 21,187 companies.

In reducing the data set to those companies that were founded from 1990 (see Extended Data Figure A.14) onwards, we aimed at limiting the potential bias that could arise from having companies in the dataset that cannot be considered as startup companies because of their age. Therefore, this additional restriction removes less than 0.6% of the companies in our dataset.

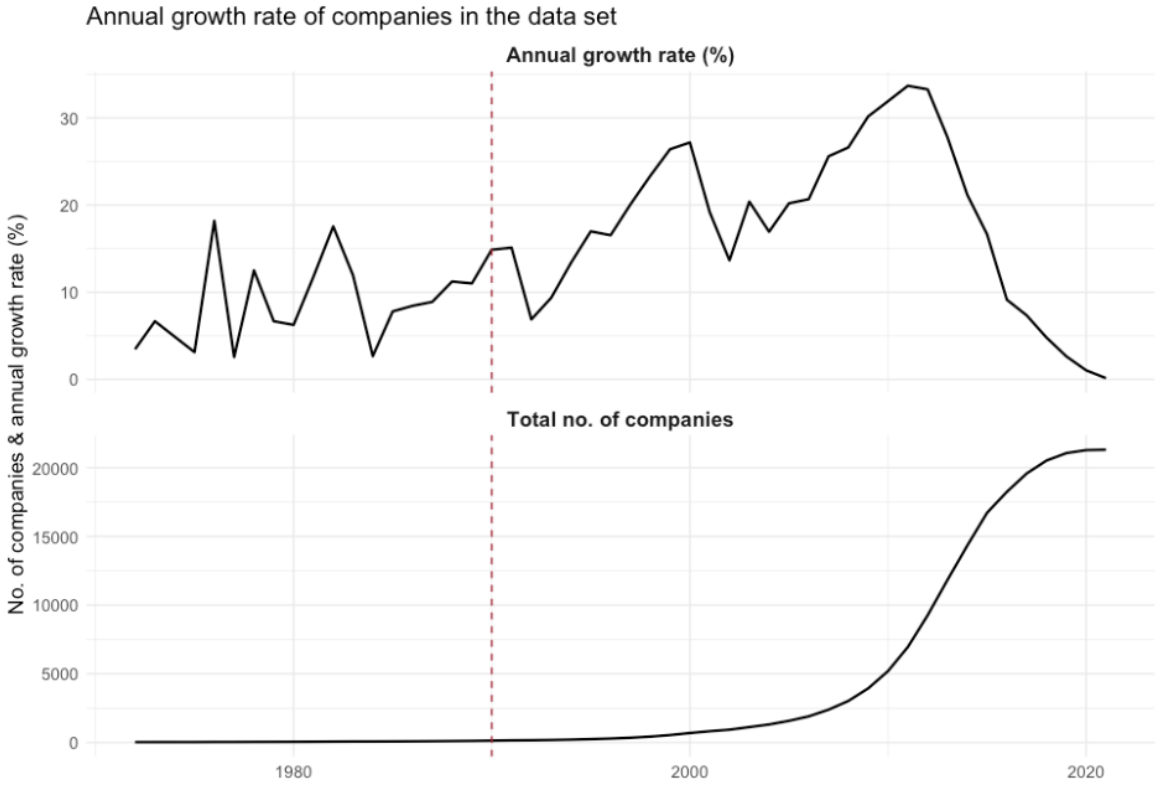
In total, 3,442 of 21,187 companies (16%) in the data set have been successful accord-



Extended Data Figure A.13: **Data Wrangling for Multifactor Analysis.** Effects of the data cleaning on the sample size. Five preparatory steps reduce the data set to 25,214 founders with inferred personality traits who have been involved in founding 21,187 startup companies.

ing to the criterion used by (Bonaventura et al., 2020). On average, successful companies needed 6.38 years to become successful (see Extended Data Figure A.15).

The final data set is a panel with 26,202 observations, i.e. combinations of 25,214 founders involved in founding 21,187 companies. For each data point, we observe a total of 104 variables. Of those, 15 variables relate to the organisations and cover aspects such as a company identifier, description, industry categories, location, and foundation year. In addition, there are six variables related to success in the data: success date & type, success indicator variable, years to success since founded, and an indicator if success occurred within the first seven years after foundation. Eight variables refer to the founders: name, Twitter, biography, primary organisation, primary job role, gender, and social media; 75 variables present different characteristics of the inferred personalities: personality type, individual facets, Big Five traits, etc.



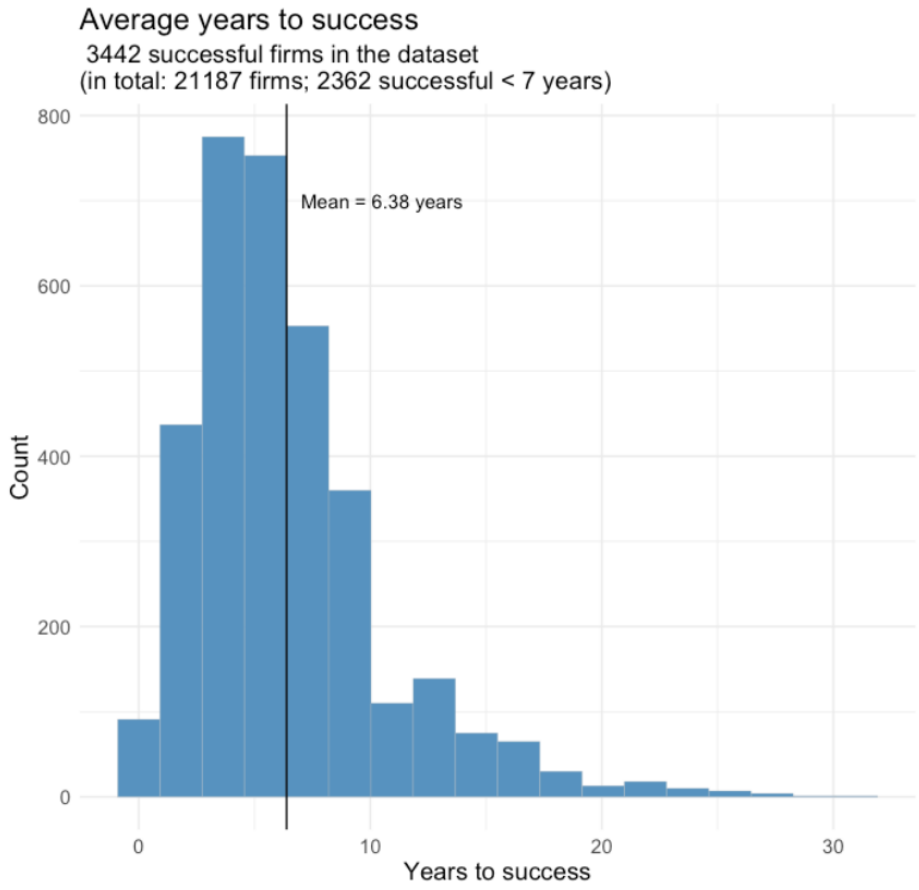
Extended Data Figure A.14: **Foundation Year.** Number of companies in the data set by foundation year. Following the approach taken by Bonaventura et al. (2020), we restrict the dataset to those companies founded from 1990 onwards.

A.6 Startup success prediction

Besides the startup success prediction model presented in the main text, we provide additional analyses on the correlational patterns linking founder personalities to success here. Extended Data Figures A.A.16 and A.17 show that certain personality types tend to be associated more with successful startup companies than others.

In Extended Data Figure A.18, we examine the gender differences between successful and unsuccessful founders. Extended Data Figure A.19 compares the predictive performance of models that include a varying amount of information, and Extended Data Figure A.20 shows the coefficient estimates of the final model displayed in the main text. Extended Data Figure A.22 provides some evidence on mechanisms that might relate certain personality facets to startup success.

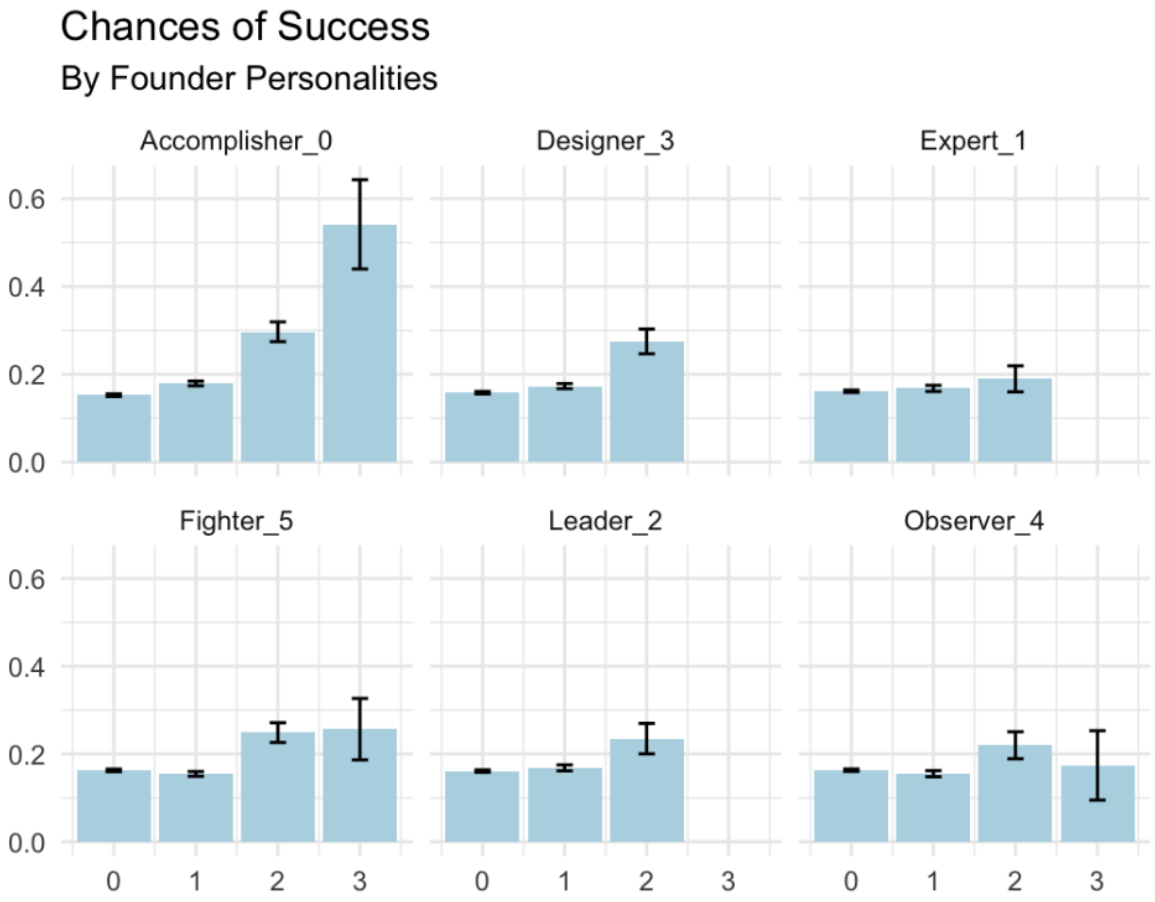
Turning to gender differences in founder personalities, we compare the distribution



Extended Data Figure A.15: **Years to success.** Histogram of the variable 'average years to success' based on 3,442 successful firms in the data set. The mean time to success is 6.38 years.

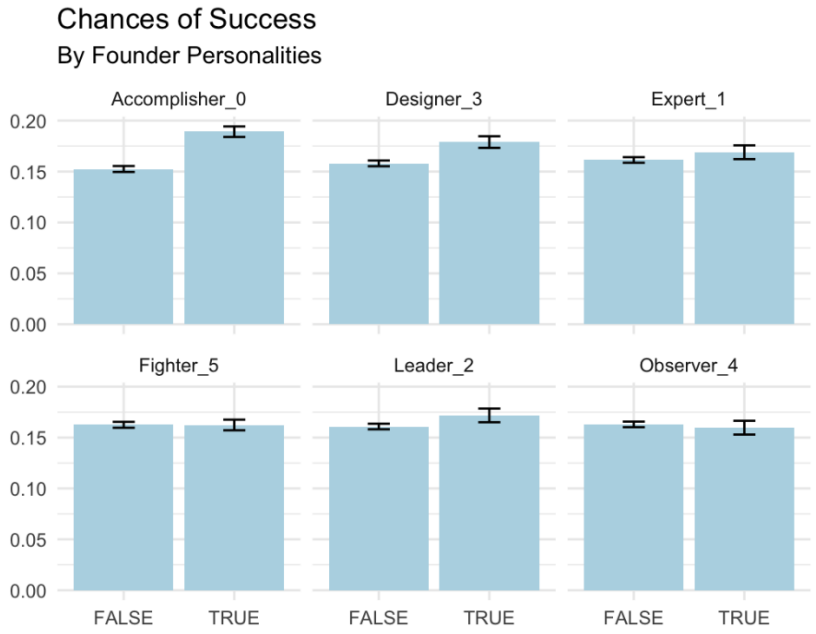
of personality facets between successful and unsuccessful founders stratified by gender. While the multifactor modelling shows having male founders increases a startup's chance of success, this is likely primarily due to the significant gender bias in venture funding. In our data, female solo-founders received, on average, 20% less total funds than their male counterparts, while female co-founded teams received less than half the funding on average as all-male teams.

The amount of venture funds and angel investors specifically targeting female founders has proliferated since 2006, such as Women's Venture Capital Fund (Founded in Portland, Oregon in 2011); Female Founders Fund (founded in NY in 2014) and Halogen Ventures (founded in 2016 in LA). This will likely address some of these inequities over the current cohort of startups.

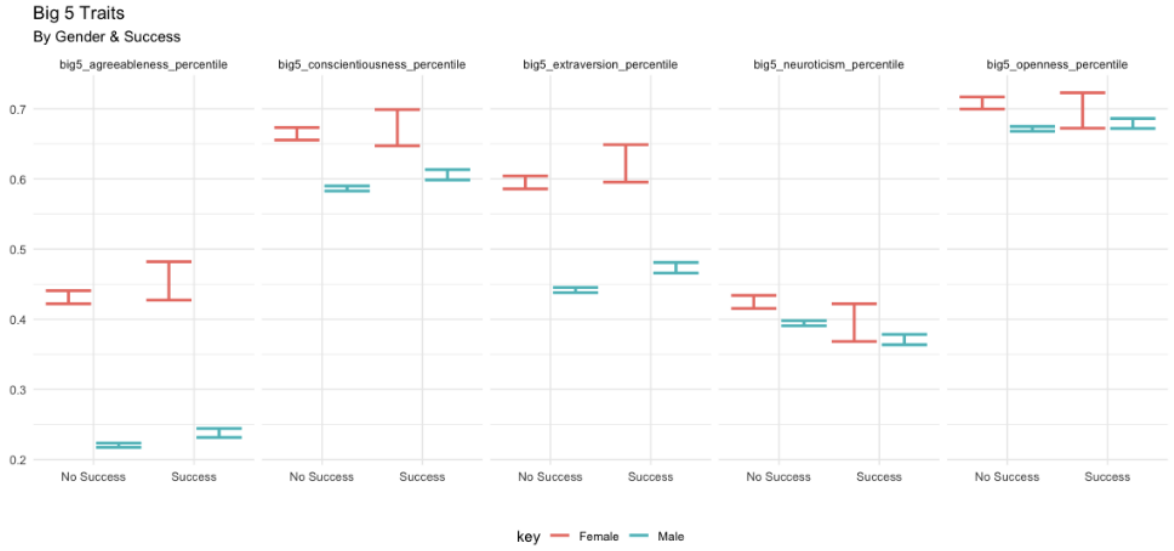


Extended Data Figure A.16: **Founder Team and Success.** Startups with multiple founders of the same personality types have higher chances of success.

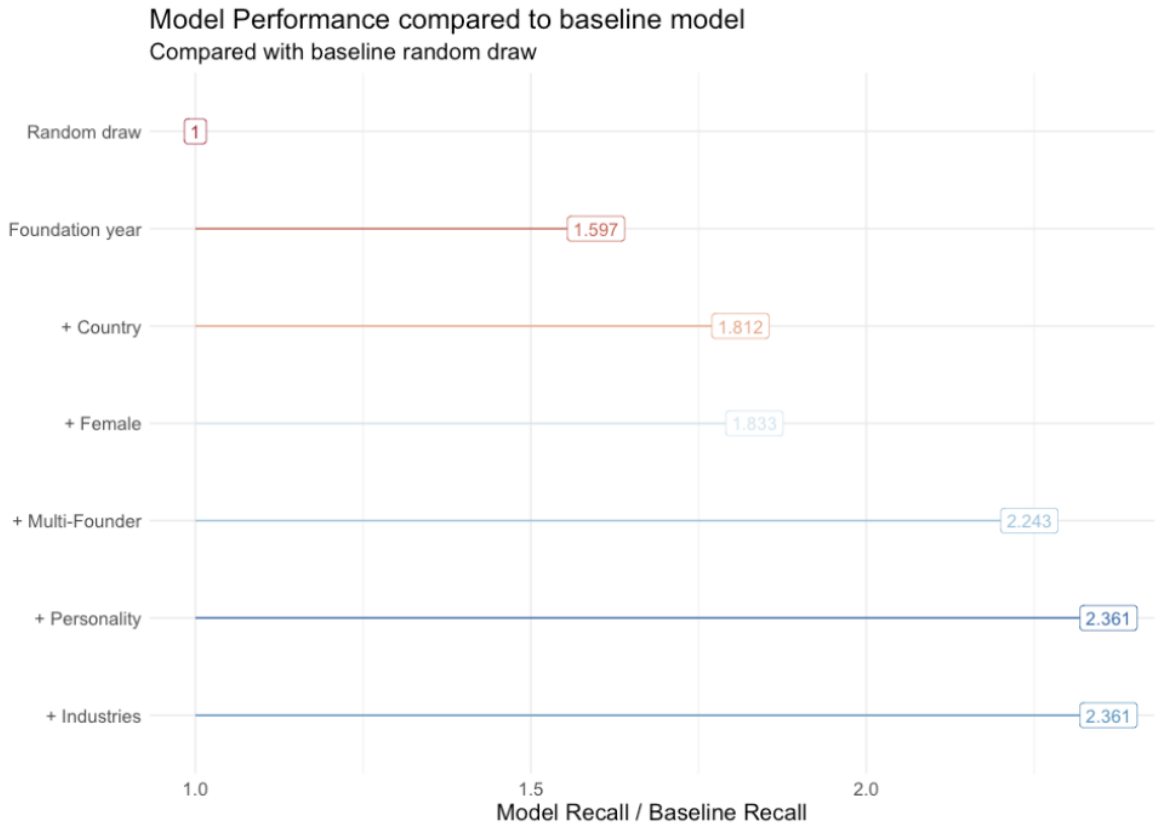
On the one hand, the correlations between external and internal factors and success, on the other hand, are visible when comparing different machine learning models that predict startup success. For example, Extended Data Figure A.19 shows the predictive performance of six logistic regression models compared to a baseline random draw model. According to the recall Machine Learning Performance metric, the best-performing models (5) and (6) are more than 130% better than random draw. These models include several explanatory variables: foundation year, country, female indicator variable, the number of founders, as well as personality types (model 5) plus industries (model 6).



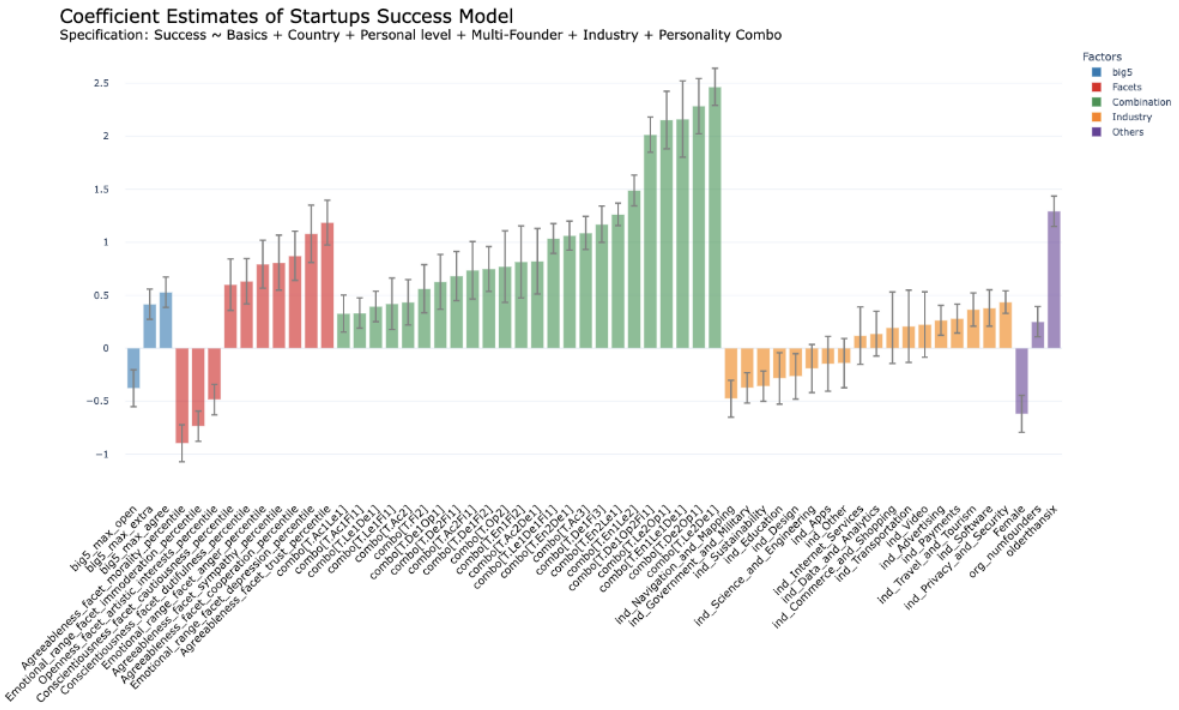
Extended Data Figure A.17: **Founder types and success.** Startups with specific founder personalities have higher chances of success - most significant for personalities of the 'Accomplisher' type.



Extended Data Figure A.18: **Gender bias in startups.** We noted Successful female founders are more similar to the successful male founder.



Extended Data Figure A.19: **Recall Performance.** Comparison of prediction performance according to Recall metric of different startup success prediction models (GLMs): (0) Null model of random draw, (1) simple model including only the foundation year as a predictive variable, (2) model 1 plus country, (3) model 2 plus female indicator variable, (4) model 3 plus count of a number of founders, (5) model 4 plus personality traits, (6) model 5 plus industries.



Extended Data Figure A.20: **Ensemble Model.** Multifactor modelling reveals the relative significance of different Firm-level, Founder-level, and Founder-Team level features on startup success and illustrates how personality-diverse larger teams have one of the most significant impacts on chances of success.

Features associated with Startup Success					
Type of feature	**Instance**	**coef**	**std err**	**p-value**	**odds ratio**
Team personality combinations	Leaders x 2, Developer x 1	2.56	1.17	0.03	12.90
Team personality combinations	Developers x 2, Operator x 1	2.14	0.77	0.01	8.46
Team personality combinations	Leaders x 1, Developer x 1, Engineer x 1	2.13	0.68	0.00	8.44
Startup Age	older than six years old	1.28	0.08	0.00	3.59
Team personality combinations	Accomplishers x 3	1.01	0.48	0.03	2.75
Personality of founders	Maximum Agreeableness (Big5)	0.47	0.12	0.00	1.60
Team personality combinations	Developer x 1, Operator x 1	0.46	0.22	0.03	1.59
Team personality combinations	Accomplishers x 2	0.46	0.14	0.00	1.59
Industry	Privacy and Security	0.46	0.09	0.00	1.58
Industry	Software	0.36	0.05	0.00	1.44
Industry	Payments	0.35	0.12	0.00	1.41
Team size	Number of founders	0.26	0.02	0.00	1.30
Industry	Transportation	0.22	0.09	0.02	1.25
Industry	Advertising	0.22	0.09	0.02	1.25
Industry	Data and Analytics	0.21	0.07	0.00	1.24
Industry	Commerce and Shopping	0.14	0.06	0.01	1.15
Industry	Information Technology	0.14	0.05	0.01	1.15
Industry	Internet Services	0.12	0.05	0.01	1.13
Industry	Other	-0.17	0.06	0.00	0.84
Industry	Apps	-0.18	0.07	0.02	0.84
Team personality combinations	Developer x 1	-0.18	0.08	0.02	0.84
Industry	Community and Lifestyle	-0.18	0.09	0.04	0.84
Team personality combinations	Fighter x 1	-0.24	0.09	0.01	0.79
Industry	Education	-0.24	0.09	0.00	0.78
Team personality combinations	Operator x 1	-0.27	0.09	0.00	0.76
Industry	Design	-0.31	0.10	0.00	0.73
Industry	Navigation and Mapping	-0.51	0.18	0.01	0.60

Note: only features with $p < 0.05$ tabled.

Extended Data Figure A.21: **Coefficient estimates and odds ratios in the Ensemble Model.** The table shows the coefficient estimates, standard errors, p-values, and odds ratios of several features in the Ensemble Model.

Big 5 (Facet) Higher/Lower	How people with this trait may be more successful as founders	Examples of successful startup founders who exhibit this trait
Openness (adventurousness) Higher	Founders with this trait may be more willing to take risks and explore new opportunities, which could lead to innovation and growth for the company.	Elon Musk, the founder of SpaceX and Tesla, is known for his willingness to take risks and explore new opportunities in industries such as space exploration and electric cars.
Agreeableness (modesty) Lower	Founders with this trait may be more confident and assertive in promoting their ideas and vision for the company, which could help them attract investors, customers, and employees.	Steve Jobs, the co-founder of Apple, was known for his confidence and assertiveness in promoting his vision for the company.
Extraversion (activity_level) Higher	Founders with this trait may have high levels of energy and drive, which could help them to accomplish more in less time. They may also be more likely to take on multiple projects and responsibilities, which could lead to growth and innovation for the company.	Richard Branson, the founder of Virgin Group, is known for his high energy and drive, which has helped him to build a successful business empire.
Emotional range (anxiety) Lower	Founders with this trait may be more resilient in the face of challenges and setbacks, which are common in the early stages of building a company. They may also be better able to manage stress and make clear-headed decisions under pressure.	Jeff Bezos, the founder of Amazon, is known for his ability to remain calm under pressure and make clear-headed decisions even in challenging situations.
Emotional range (immoderation) Lower	Founders with this trait may be more disciplined and able to resist temptation, which could help them to make better decisions for the long-term success of the company. They may also be better able to manage their emotions and avoid impulsive actions that could harm the business.	Mark Zuckerberg, the co-founder of Facebook, is known for his discipline and focus, which has helped him to build one of the most successful tech companies in the world.
Agreeableness (trust) Higher	Founders with this trait may be more likely to build strong relationships with employees, customers, and investors based on mutual trust and respect. This could help to create a positive work environment and foster loyalty and commitment.	Larry Page and Sergey Brin, the co-founders of Google, are known for their ability to build strong relationships with employees and foster a positive work environment based on mutual trust and respect.

Extended Data Figure A.22: Potential mechanisms on how personality might affect startup success.

APPENDIX B - COSMOS 1.0: A MULTIDIMENSIONAL MAP OF THE EMERGING TECHNOLOGY FRONTIER

B.1 Supplementary Methods

Instance label: academic discipline, academic major, access control, advanced driver-assistance systems, advertising, aircraft, aircraft model, algorithm, allotrope of silicon, animation technique, application, applications of artificial intelligence, applied science, art genre, artificial intelligence model, artificial satellite, audio effect, automated rapid transit, automation, autonomous car, battery chemistry type, Berkeley Open Infrastructure for Network Computing projects, biological process, branch of biology, branch of chemistry, branch of computer science, branch of physics, branch of science, business, buzzword, carbon neutrality, cellular network, certification, chatbot, chemical element, cinematic technique, class of fictional entities, climate change mitigation, cloud computing, communication protocol, communication technology, computer graphics term, computer network protocol, computer program, computer simulation, computing, computing platform, concept, copy protection, design, digital rights management, distributed computing, distributed revision control system, driver of Industry 4.0, drug delivery, educational software, educational technology, electrical element, email filtering, energy conservation technology, energy industry, error correction code, expert system, fictional technology, fictional uploaded consciousness, field of study, finance, forward error correction, free and open-source software, free software, genetic algorithm, grid computing, group or class of chemical substances, group or class of proteins, GNU package, hash function, history of technology, hypothetical technology, image processing, individual transportation, industry, information system, information technology, informed search algorithm, integrated development environment, interdisciplinary science, interface standard, invention, knowledge market, learning management system, light rail, linear code, list of manufacturing processes, machine learning, macroscopic quantum phenomena, marine propulsion, media, medical device type, medical diagnosis, medical imaging, medical journal, medical model,

medical speciality, medical test type, metaheuristic, method, mobile app, mobile phone generation, mobile phone network standard, mobile telecommunication technology, mode of transport, navigation system, network, noise reduction, non-classical state of matter, nuclear reactor generation, nuclear technology, online course, online service, open-source software, parody generator, patent classification, pathfinding algorithm, people mover, periodical, peripheral, photographic technique, physical media format, physical technological component, process, product, production process, programming language, programming paradigm, project, property, protocol stack, public-domain software, push technology, question-and-answer site, radar, radio program, reasoning, recording medium, research project, robot, robotics, scholarly article, science, science fiction theme, scientific journal, sensor, sequencing, server software, service, social networking service, software, software as a service, software feature, software library, space instrument, spaced repetition software, spaceflight, speciality, startup company, subject heading, superpower, symbol, system, Tactical communications system, technical sciences, technical standard, technical term, technique, technology, thermal weapon sight, thermographic camera, thought experiment, trademark, train and rail category, Transport system for persons, type of computer memory or storage, type of manufactured good, type of quantum particle, type of security, type of sport, type of technology, vaccine type, vehicle model, virtual reality, visual effects, volunteer computing, weapon functional class, web portal, web service, website, Wikimedia disambiguation page.

BIBLIOGRAPHY

Abercrombie, R. K., Udoeyop, A. W., & Schlicher, B. G. (2012). A study of scientometric methods to identify emerging technologies via modeling of milestones. *Scientometrics, 91*(2), 327–342.

Agenda, F., & News, F. S. (2023). *Top 10 emerging technologies of 2023* (tech. rep.). https://www.weforum.org/publications/top-10-emerging-technologies-of-2023/

Albert, R., & Barabási, A.-L. (2002). Statistical mechanics of complex networks. *Reviews of modern physics, 74*(1), 47.

Albert, R., Jeong, H., & Barabási, A.-L. (1999). Diameter of the world-wide web. *Nature (London), 401*(6749), 130–131.

Aldrich, H. E., & Wiedenmayer, G. (2019). From traits to rates: An ecological perspective on organizational foundings. In *Seminal ideas for the next twenty-five years of advances* (pp. 61–97). Emerald Publishing Limited.

AlShebli, B. K., Rahwan, T., & Woon, W. L. (2018). The preeminence of ethnic diversity in scientific collaboration. *Nature communications, 9*(1), 5163.

Alvarez-Aros, E. L., & Bernal-Torres, C. A. (2021). Technological competitiveness and emerging technologies in industry 4.0 and industry 5.0. *Anais da Academia Brasileira de Ciências, 93*.

Antretter, T., Blohm, I., & Grichnik, D. (2018). Predicting startup survival from digital traces: Towards a procedure for early stage investors.

Armstrong, C., Davila, A., & Foster, G. (2006). Venture-backed private equity valuation and financial statement information. *Review of Accounting Studies, 11*(1), 119–154.

Arnoux, P.-H., Xu, A., Boyette, N., Mahmud, J., Akkiraju, R., & Sinha, V. (2017). 25 tweets to know you: A new model to predict personality with social media. *Proceedings of the international AAAI conference on web and social media*, *11*(1), 472–475.

Arora, A., Belenzon, S., & Patacconi, A. (2018). The decline of science in corporate r&d. *Strategic Management Journal*, *39*(1), 3–32.

Arora, S. K., Porter, A. L., Youtie, J., & Shapira, P. (2013). Capturing new developments in an emerging technology: An updated search strategy for identifying nanotechnology research outputs. *Scientometrics*, *95*(1), 351–370.

Arundel, A., & Kabla, I. (1998). What percentage of innovations are patented? empirical estimates for european firms. *Research policy*, *27*(2), 127–141.

Balakrishnan, N., Cheng, G., & Patel, M. (1979). Discussion of emerging technologies for air pollution control. *Pollut. Eng.;(United States)*, *11*(11).

Barrick, M. R., & Mount, M. K. (1991). The big five personality dimensions and job performance: A meta-analysis. *Personnel psychology*, *44*(1), 1–26.

Beauchamp, M. A. (1965). An improved index of centrality. *Behavioral Science*, *10*(2), 161–163.

Bekar, C., Carlaw, K., & Lipsey, R. (2018). General purpose technologies in theory, application and controversy: A review. *Journal of Evolutionary Economics*, *28*, 1005–1033.

Bettencourt, L., Kaiser, D., Kaur, J., Castillo-Chavez, C., & Wojick, D. (2008). Population modeling of the emergence and development of scientific fields. *Scientometrics*, *75*(3), 495–518.

Blank, S., et al. (2013). Why the lean start-up changes everything. *Harvard business review*, *91*(5), 63–72.

Block, J., & Sandner, P. (2009). What is the effect of the financial crisis on venture capital financing? empirical evidence from us internet start-ups. *Venture Capital*, *11*(4), 295–309.

Boldi, P., Furia, F., & Vigna, S. (2022). Spectral rank monotonicity in undirected networks. In R. M. Benito, C. Cherifi, H. Cherifi, E. Moro, L. M. Rocha, & M. Sales-Pardo (Eds.), *Complex networks & their applications x* (pp. 234–246, Vol. 1014). Springer.

Boldi, P., Furia, F., & Vigna, S. (2023). Monotonicity in undirected networks. *Network Science*, *11*(3), 351–373.

Boldi, P., & Vigna, S. (2014). Axioms for centrality. *Internet Math.*, *10*(3-4), 222–262.

Bollobás, B., Borgs, C., Chayes, J. T., & Riordan, O. (2003). Directed scale-free graphs. *SODA*, *3*, 132–139.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., et al. (2021). On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Bonaccorsi, A., Chiarello, F., Fantoni, G., & Kammering, H. (2020). Emerging technologies and industrial leadership. a wikipedia-based strategic analysis of industry 4.0. *Expert Systems with Applications*, *160*, 113645.

Bonaventura, M., Ciotti, V., Panzarasa, P., Liverani, S., Lacasa, L., & Latora, V. (2020). Predicting success in the worldwide start-up network. *Scientific reports*, *10*(1), 345.

Bono, J. E., & Judge, T. A. (2004). Personality and transformational and transactional leadership: A meta-analysis. *Journal of applied psychology*, *89*(5), 901–910.

Börner, K. (2010). Atlas of science.

Branch, B. (1974). Research and development activity and profitability: A distributed lag analysis. *Journal of political economy*, *82*(5), 999–1011.

Breschi, S., Malerba, F., & Orsenigo, L. (2000). Technological regimes and schumpeterian patterns of innovation. *The economic journal*, *110*(463), 388–410.

Bresnahan, T. F., & Trajtenberg, M. (1995). General purpose technologies 'engines of growth'? *Journal of econometrics*, *65*(1), 83–108.

Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *Computer networks and ISDN systems*, *30*(1), 107–117.

Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A., & Wiener, J. (2000). Graph structure in the web. *Computer networks (Amsterdam, Netherlands : 1999)*, *33*(1-6), 309–320.

Brown, B. (1968). *Delphi process: A methodology used for the elicitation of opinions of experts* (tech. rep.). RAND Corporation. United States.

Bruckman, A. S. (2022). *Should you believe wikipedia?: Online communities and the construction of knowledge*. Cambridge University Press.

Brush, C., Edelman, L. F., Manolova, T., & Welter, F. (2019). A gendered look at entrepreneurship ecosystems. *Small business economics*, *53*(2), 393–408.

Caliński, T., & Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods*, *3*(1), 1–27.

Callon, M., Courtial, J.-P., Turner, W. A., & Bauin, S. (1983). From translations to problematic networks: An introduction to co-word analysis. *Social science information*, *22*(2), 191–235.

Carroll, L. S. L. (2017). A comprehensive definition of technology from an ethological perspective. *Social Sciences*, *6*(4), 126.

Castells, M. (2011). *The rise of the network society*. John wiley & sons.

Caviggioli, F. (2016). Technology fusion: Identification and analysis of the drivers of technology convergence using patent data. *Technovation*, *55*, 22–32.

Chan, L. K., Lakonishok, J., & Sougiannis, T. (2001). The stock market valuation of research and development expenditures. *The Journal of finance*, *56*(6), 2431–2456.

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785

Chiarello, F., Trivelli, L., Bonaccorsi, A., & Fantoni, G. (2018). Extracting and mapping industry 4.0 technologies using wikipedia. *Computers in Industry*, *100*, 244–257.

Clauset, A., Shalizi, C. R., & Newman, M. E. (2009). Power-law distributions in empirical data. *SIAM review*, *51*(4), 661–703.

Cobb-Clark, D. A., & Schurer, S. (2012). The stability of big-five personality traits. *Economics Letters*, *115*(1), 11–15.

Coccia, M. (2019). Why do nations produce science advances and new technology? *Technology in society*, *59*, 101124.

Costa Jr, P. T., & McCrae, R. R. (1995). Domains and facets: Hierarchical personality assessment using the revised neo personality inventory. *Journal of personality assessment*, *64*(1), 21–50.

Cozzens, S., Gatchair, S., Kang, J., Kim, K.-S., Lee, H. J., Ordóñez, G., & Porter, A. (2010). Emerging technologies: Quantitative identification and measurement. *Technology Analysis & Strategic Management*, *22*(3), 361–376.

Dahlman, C. (2007). Technology, globalization, and international competitiveness: Challenges for developing countries. *Industrial development for the 21st century: Sustainable development perspectives*, 29–83.

Damian, R. I., Spengler, M., Sutu, A., & Roberts, B. W. (2019). Sixteen going on sixty-six: A longitudinal study of personality stability and change across 50 years. *Journal of personality and social psychology*, *117*(3), 674.

Davies, D. L., & Bouldin, D. W. (1979). A cluster separation measure. *IEEE transactions on pattern analysis and machine intelligence*, (2), 224–227.

Davila, A., Foster, G., & Gupta, M. (2003). Venture capital financing and the growth of startup firms. *Journal of business venturing*, *18*(6), 689–708.

Davila, A., Foster, G., He, X., & Shimizu, C. (2015). The rise and fall of startups: Creation and destruction of revenue and jobs by young companies. *Australian Journal of Management*, *40*(1), 6–35.

den Besten, M. (2017). Using crunchbase for economic and managerial research. *OECD science, technology and industry working papers*.

DeYoung, C. G., Quilty, L. C., & Peterson, J. B. (2007). Between facets and domains: 10 aspects of the big five. *Journal of personality and social psychology*, *93*(5), 880–896.

Duggan, M., Ellison, N. B., Lampe, C., Lenhart, A., & Madden, M. (2015). Demographics of key social networking platforms.

https://www.pewresearch.org/internet/2015/01/09/demographics-of-key-social-networking-platforms-2/

Dworak, D. (2022). Analysis of founder background as a predictor for start-up success in achieving successive fundraising rounds.

Eberhart, A. C., Maxwell, W. F., & Siddique, A. R. (2004). An examination of long-term abnormal stock returns and operating performance following r&d increases. *The journal of finance*, *59*(2), 623–650.

Eikelboom, M. E., Gelderman, C., & Semeijn, J. (2018). Sustainable innovation in public procurement: The decisive role of the individual. *Journal of Public Procurement*, *18*(3), 190–201.

Eilers, K., Frischkorn, J., Eppinger, E., Walter, L., & Moehrle, M. G. (2019). Patent-based semantic measurement of one-way and two-way technology convergence: The case of ultraviolet light emitting diodes (uv-leds). *Technological Forecasting and Social Change*, *140*, 341–353.

Ernst, H. (2003). Patent information for strategic technology management. *World patent information*, *25*(3), 233–242.

Fan, J. S. (2022). Startup biases. *UC Davis Law Review*.

Fisch, C., & Block, J. H. (2021). How does entrepreneurial failure change an entrepreneur's digital identity? evidence from twitter data. *Journal of Business Venturing*, *36*(1), 106015.

Fortunato, S., Bergstrom, C. T., Börner, K., Evans, J. A., Helbing, D., Milojević, S., Petersen, A. M., Radicchi, F., Sinatra, R., Uzzi, B., et al. (2018). Science of science. *Science*, *359*(6379), eaao0185.

Fracassi, C., Garmaise, M. J., Kogan, S., & Natividad, G. (2016). Business microloans for us subprime borrowers. *Journal of Financial and Quantitative Analysis*, *51*(1), 55–83.

Freeman, L. C. (1979). Centrality in social networks: Conceptual clarification. *Social networks*, *1*(3), 215–239.

Freiberg, B., & Matz, S. C. (2023). Founder personality and entrepreneurial outcomes: A large-scale field study of technology startups. *Proceedings of the National Academy of Sciences*, *120*(19), e2215829120.

Friedman, H. S., & Kern, M. L. (2014). Personality, well-being, and health. *Annual Review of Psychology*, *65*(1), 719–742.

Gartner, W. B. (1988). "who is an entrepreneur?" is the wrong question. *American journal of small business*, *12*(4), 11–32.

Gartner, Inc. (2024). 30 emerging technologies that will guide your business decisions.

Ghosh, S., & Nanda, R. (2010). Venture capital investment in the clean energy sector. *Harvard Business School Entrepreneurial Management Working Paper*, (11-020).

Goldberg, L. R. (1990). An alternative "description of personality": The big-five factor structure. *Journal of personality and social psychology*, *59*(6), 1216–1229.

Gompers, P. A., & Lerner, J. (2004). *The venture capital cycle*. MIT press.

Goodwin, R. D., & Friedman, H. S. (2006). Health status and the five-factor personality traits in a nationally representative sample. *Journal of health psychology*, *11*(5), 643–654.

Gourevitch, A., Portincaso, M., de la Tour, A., Goeldel, N., & Chaudhry, U. (2021). Deep tech and the great wave of innovation. *Boston Consulting Group: Boston, MA, USA*.

Gowda, K. C., & Krishna, G. (1978). Agglomerative clustering using the concept of mutual nearest neighbourhood. *Pattern recognition*, *10*(2), 105–112.

Graham, P. (2013). Do things that don't scale. *Paul Graham*, *531*.

Grant, K. A., Croteau, M., & Aziz, O. (2019). The survival rate of startups funded by angel investors. *I-Inc White Paper Series: Mar*, *2019*, 1–21.

Griliches, Z. (1991). The search for r&d spillovers. *National Bureau of Economic Research Working Paper Series*, (w3768).

Grossman, G. M., & Helpman, E. (1993). *Innovation and growth in the global economy*. MIT press.

Guzman, J., & Stern, S. (2015). Where is silicon valley? *Science*, *347*(6222), 606–609.

Hallen, B. L., & Eisenhardt, K. M. (2012). Catalyzing strategies and efficient tie formation: How entrepreneurial firms obtain investment ties. *Academy of Management Journal*, *55*(1), 35–70.

Hamilton, B. H., Papageorge, N. W., & Pande, N. (2019). The right stuff? personality and entrepreneurship. *Quantitative Economics*, *10*(2), 643–691.

Hamilton, W., Ying, Z., & Leskovec, J. (2017). Inductive representation learning on large graphs. *Advances in neural information processing systems*, *30*.

Henrekson, M., & Johansson, D. (2010). Gazelles as job creators: A survey and interpretation of the evidence. *Small business economics*, *35*(2), 227–244.

Ho, J. C., Saw, E.-C., Lu, L. Y., & Liu, J. S. (2014). Technological barriers and research trends in fuel cell technologies: A citation network analysis. *Technological Forecasting and Social Change*, *82*, 66–79.

Hochberg, Y. V., Ljungqvist, A., & Lu, Y. (2007). Whom you know matters: Venture capital networks and investment performance. *The journal of finance*, *62*(1), 251–301.

Holloway, T., Bozicevic, M., & Börner, K. (2007). Analyzing and visualizing the semantic coverage of wikipedia and its authors. *Complexity*, *12*(3), 30–40.

Hsu, D. H. (2006). Venture capitalists and cooperative start-up commercialization strategy. *Management Science*, *52*(2), 204–219.

Hussain, O. A., Zaidi, F., & Rozenblat, C. (2019). Analyzing diversity, strength and centrality of cities using networks of multinational firms. *Networks and Spatial Economics*, *19*, 791–817.

Insights, C. (2019). The top 20 reasons startups fail [research brief].

Jaeger, H., Knorr, D., Szabó, E., Hámori, J., & Bánáti, D. (2015). Impact of terminology on consumer acceptance of emerging technologies through the example of pef technology. *Innovative Food Science & Emerging Technologies*, *29*, 87–93.

Jemielniak, D. (2019). Wikipedia: Why is the common knowledge resource still neglected by academics? *GigaScience*, *8*(12), giz139.

John, O. P., & Srivastava, S. (1999). The big five trait taxonomy: History, measurement, and theoretical perspectives. In L. A. Pervin & O. P. John (Eds.), *Handbook of personality: Theory and research* (2nd ed., pp. 102–138). Guilford Press.

Jovanovic, B., & Rousseau, P. L. (2005). General purpose technologies. In *Handbook of economic growth* (pp. 1181–1224, Vol. 1). Elsevier.

Kanze, D., Huang, L., Conley, M. A., & Higgins, E. T. (2018). We ask men to win and women not to lose: Closing the gender gap in startup funding. *Academy of management journal*, *61*(2), 586–614.

Kaplan, S. N., & Lerner, J. (2010). It ain't broke: The past, present, and future of venture capital. *Journal of Applied Corporate Finance*, *22*(2), 36–47.

Kern, M. L., McCarthy, P. X., Chakrabarty, D., & Rizoiu, M.-A. (2019). Social media-predicted personality traits and values can help match people to their ideal jobs. *Proceedings of the National Academy of Sciences*, *116*(52), 26459–26464.

Kerr, S. P., Kerr, W. R., Xu, T., et al. (2018). Personality traits of entrepreneurs: A review of recent literature. *Foundations and Trends® in Entrepreneurship*, *14*(3), 279–356.

Khan, L. M. (2016). Amazon's antitrust paradox. *Yale lJ*, *126*, 710.

Kile, C. O., & Phillips, M. E. (2009). Using industry classification codes to sample high-technology firms: Analysis and recommendations. *Journal of Accounting, Auditing & Finance*, *24*(1), 35–58.

Kim, J. (2022). New technologies and stock returns. *Available at SSRN 4299577*.

Kim, S., Park, I., & Yoon, B. (2020). Sao2vec: Development of an algorithm for embedding the subject–action–object (sao) structure using doc2vec. *Plos one*, *15*(2), e0227930.

Kitchin, R. (2014). *The data revolution: Big data, open data, data infrastructures and their consequences*. Sage.

Klotz, A. C., Hmieleski, K. M., Bradley, B. H., & Busenitz, L. W. (2014). New venture teams: A review of the literature and roadmap for future research. *Journal of management*, *40*(1), 226–255.

Kose, T., & Sakata, I. (2019). Identifying technology convergence in the field of robotics research. *Technological Forecasting and Social Change, 146*, 751–766.

Kostoff, R. N., Boylan, R., & Simons, G. R. (2004). Disruptive technology roadmaps. *Technological Forecasting and Social Change, 71*(1-2), 141–159.

Kousha, K., Thelwall, M., & Rezaie, S. (2011). Assessing the citation impact of books: The role of google books, google scholar, and scopus. *Journal of the American Society for information science and technology, 62*(11), 2147–2164.

Langville, A. N., & Meyer, C. D. (2006). *Google's pagerank and beyond : The science of search engine rankings*. Princeton University Press.

Lavassani, K. M., & Movahedi, B. (2021). Firm-level analysis of global supply chain network: Role of centrality on firm's performance. *International Journal of Global Business and Competitiveness, 16*(2), 86–103.

Lee, R. P., & Grewal, R. (2004). Strategic responses to new technologies and their impact on firm performance. *Journal of Marketing, 68*(4), 157–171.

Leutner, F., Ahmetoglu, G., Akhtar, R., & Chamorro-Premuzic, T. (2014). The relationship between the entrepreneurial personality and the big five personality traits. *Personality and individual differences, 63*, 58–63.

Lev, B., & Sougiannis, T. (1996). The capitalization, amortization, and value-relevance of r&d. *Journal of accounting and economics, 21*(1), 107–138.

Malizia, E. E., & Ke, S. (1993). The influence of economic diversity on unemployment and stability. *Journal of Regional Science, 33*(2), 221–235.

Malouff, J. M., Thorsteinsson, E. B., & Schutte, N. S. (2005). The relationship between the five-factor model of personality and symptoms of clinical disorders: A meta-analysis. *Journal of psychopathology and behavioral assessment, 27*(2), 101–114.

Martin, B. R. (1995). Foresight in science and technology. *Technology analysis & strategic management, 7*(2), 139–168.

McCarthy, P. X. (2015). *Online gravity: The unseen force driving the way you live, earn, and learn*. Simon; Schuster.

McCarthy, P. X., Kern, M. L., Gong, X., Parker, M., & Rizoiu, M.-A. (2022). Occupation-personality fit is associated with higher employee engagement and happiness.

McCarthy, P. X., McFarland, C., **Gong, Xian**, & Griffith, C. (2023). The atlas of australia online 2023.

McCarthy, P. X., **Gong, Xian**, Braesemann, F., Stephany, F., Rizoiu, M.-A., & Kern, M. L. (2023). The impact of founder personalities on startup success. *Scientific Reports*, *13*(1), 17200.

McCarthy, P. X., **Gong, Xian**, Coetzer, M., Rizoiu, M.-A., Kern, M. L., Johnson, J. A., Holden, R., & Braesemann, F. (2025). The economics of global personality diversity. *arXiv preprint arXiv:2503.19388*.

McCarthy, P. X., **Gong, Xian**, Eghbal, S., Falster, D. S., & Rizoiu, M.-A. (2021). Evolution of diversity and dominance of companies in online activity. *Plos one*, *16*(4), e0249993.

McCrae, R. R., & John, O. P. (1992). An introduction to the five-factor model and its applications. *Journal of personality*, *60*(2), 175–215.

McFarland, C., McCarthy, P. X., Gong, X., & Griffith, C. (2023). Atlas of australia online. *.au Domain Administration Limited (auDA)*.

McKinsey & Company. (2024). Mckinsey technology trends outlook 2024.

Mesgari, M., Okoli, C., Mehdi, M., Nielsen, F. Å., & Lanamäki, A. (2015). "the sum of all human knowledge": A systematic review of scholarly research on the content of wikipedia. *Journal of the Association for Information Science and Technology*, *66*(2), 219–245.

Mestyán, M., Yasseri, T., & Kertész, J. (2013). Early prediction of movie box office success based on wikipedia activity big data. *PloS one*, *8*(8), e71226.

Meusel, R., Vigna, S., Lehmberg, O., & Bizer, C. (2015). The graph structure in the web–analyzed on different aggregation levels. *The Journal of Web Science*, *1*.

Mikolov, T., Chen, K., Corrado, G. S., & Dean, J. (2013). Efficient estimation of word representations in vector space. *International Conference on Learning Representations*. https://api.semanticscholar.org/CorpusID:5959482

Milojević, S. (2015). Quantifying the cognitive extent of science. *Journal of Informetrics*, *9*(4), 962–973.

Nann, S., Krauss, J., Schober, M., Gloor, P. A., Fischbach, K., & Führes, H. (2010). Comparing the structure of virtual entrepreneur networks with business effectiveness. *Procedia-Social and Behavioral Sciences*, *2*(4), 6483–6496.

Newman, M. (2018). *Networks*. Oxford university press.

Nindl, E., Confraria, H., Rentocchini, F., Napolitano, L., Georgakaki, A., Ince, E., Fako, P., Tuebke, A., Gavigan, J., Hernandez Guevara, H., et al. (2023). The 2023 eu industrial r&d investment scoreboard, publications office of the european union, luxembourg. *DOI*, *10*, 506189.

Obschonka, M., Schmitt-Rodermund, E., Silbereisen, R. K., Gosling, S. D., & Potter, J. (2013). The regional distribution and correlates of an entrepreneurship-prone personality profile in the united states, germany, and the united kingdom: A socioecological perspective. *Journal of personality and social psychology*, *105*(1), 104–122.

OECD. (2016). *Oecd science, technology and innovation outlook 2016*. https://doi.org/https://doi.org/10.1787/sti_in_outlook-2016-en

Oltermann, P. (2021). Pfizer/biontech tax windfall brings mainz an early christmas present. *The Guardian*. https://www.theguardian.com/world/2021/dec/27/pfizer-biontech-tax-windfall-brings-mainz-an-early-christmas-present

Ortega, J. M. E., & Eballe, R. G. (2022). Harmonic centralization of some graph families. *arXiv preprint arXiv:2204.04381*.

Ozer, D. J., & Benet-Martínez, V. (2006). Personality and the prediction of consequential outcomes. *Annual review of psychology*, *57*(1), 401–421.

Page, L., Brin, S., Motwani, R., & Winograd, T. (1998). *The PageRank citation ranking: Bringing order to the web* (tech. rep. No. SIDL-WP-1999-0120). Stanford Digital Library Technologies Project, Stanford University.

Paunonen, S. V., & Ashton, M. C. (2001). Big five factors and facets and the prediction of behavior. *Journal of personality and social psychology*, *81*(3), 524–539.

Plank, B., & Hovy, D. (2015). Personality traits on twitter—or—how to get 1,500 personality tests in a week. *Proceedings of the 6th workshop on computational approaches to subjectivity, sentiment and social media analysis*, 92–98.

Porter, A., & Rafols, I. (2009). Is science becoming more interdisciplinary? measuring and mapping six research fields over time. *scientometrics*, *81*(3), 719–745.

Porter, A. L., & Cunningham, S. W. (2004). *Tech mining: Exploiting new technologies for competitive advantage*. John Wiley & Sons.

Pósfai, M., & Barabási, A.-L. (2016). *Network science* (Vol. 3). Cambridge University Press Cambridge, UK:

Pratt, A. C. (2006). Advertising and creativity, a governance approach: A case study of creative agencies in london. *Environment and planning A*, *38*(10), 1883–1899.

Rantanen, J., Metsäpelto, R.-L., Feldt, T., Pulkkinen, L., & Kokko, K. (2007). Long-term stability in the big five personality traits in adulthood. *Scandinavian journal of psychology*, *48*(6), 511–518.

Rao, D., Ye, Y., Lin, Z., & Ding, W. (2022). Efficient approximate calculations and application of network centrality. *International Conference on Frontier Computing*, 7–18.

Review, M. T. (2023). Top 10 breakthrough technologies 2023.

Riccaboni, M., Wang, X., & Zhu, Z. (2021). Firm performance in networks: The interplay between firm centrality and corporate group size. *Journal of Business Research*, *129*, 641–653.

Roberts, B. W., Caspi, A., & Moffitt, T. E. (2001). The kids are alright: Growth and stability in personality development from adolescence to adulthood. *Journal of personality and social psychology*, *81*(4), 670.

Roberts, B. W., Kuncel, N. R., Shiner, R., Caspi, A., & Goldberg, L. R. (2007). The power of personality: The comparative validity of personality traits, socioeconomic status, and cognitive ability for predicting important life outcomes. *Perspectives on Psychological science*, *2*(4), 313–345.

Roll, U., Mittermeier, J. C., Diaz, G. I., Novosolov, M., Feldman, A., Itescu, Y., Meiri, S., & Grenyer, R. (2016). Using wikipedia page views to explore the cultural importance of global reptiles. *Biological conservation, 204*, 42–50.

Rotolo, D., Hicks, D., & Martin, B. R. (2015). What is an emerging technology? *Research policy, 44*(10), 1827–1843.

Salmony, F. U., & Kanbach, D. K. (2022). Personality trait differences across types of entrepreneurs: A systematic literature review. *Review of managerial science, 16*(3), 713–749.

Schot, J., & Steinmueller, W. E. (2018). Three frames for innovation policy: R&d, systems of innovation and transformative change. *Research policy, 47*(9), 1554–1567.

Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., Shah, A., Kosinski, M., Stillwell, D., Seligman, M. E., et al. (2013). Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one, 8*(9), e73791.

Shane, S., & Venkataraman, S. (2000). The promise of entrepreneurship as a field of research. *Academy of management review, 25*(1), 217–226.

Sick, N., & Bröring, S. (2022). Exploring the research landscape of convergence from a tim perspective: A review and research agenda. *Technological Forecasting and Social Change, 175*, 121321.

Small, H., & Upham, P. (2009). Citation structure of an emerging research area on the verge of application. *Scientometrics, 79*(2), 365–375.

Smith, K. (2001). Comparing economic performance in the presence of diversity. *Science and Public Policy, 28*(4), 267–276.

Soldz, S., & Vaillant, G. E. (1999). The big five personality traits and the life course: A 45-year longitudinal study. *Journal of research in personality, 33*(2), 208–232.

Sougiannis, T. (1994). The accounting based valuation of corporate r&d. *Accounting review*, 44–68.

Sun, B., Yang, X., Zhong, S., Tian, S., & Liang, T. (2024). How do technology convergence and expansibility affect information technology diffusion? evidence from the

internet of things technology in china. *Technological Forecasting and Social Change*, *203*, 123374.

Gong, Xian, McCarthy, P. X., Griffith, C., McFarland, C., & Rizoiu, M.-A. (2025). Cosmos 1.0: A multidimensional map of the emerging technology frontier. *Scientific Data **(accepted)***.

Gong, Xian, McCarthy, P. X., Rizoiu, M.-A., & Boldi, P. (2024). Harmony in the australian domain space. *Proceedings of the 16th ACM Web Science Conference*, 92–102.

Gong, Xian, McFarland, C., McCarthy, P. X., Griffith, C., & Rizoiu, M.-A. (2023). Informing innovation management: Linking leading r&d firms and emerging technologies. *Proceedings of the International Society for Professional Innovation Management (ISPIM)*, 1–16.

The Economist. (2022). Which vaccine saved the most lives in 2021?: Covid-19. *The Economist (Online)*. https://www.economist.com/graphic-detail/2022/07/13/which-covid-19-vaccine-saved-the-most-lives-in-2021

Thornton, P. H. (1999). The sociology of entrepreneurship. *Annual review of sociology*, *25*(1), 19–46.

Tshitoyan, V., Dagdelen, J., Weston, L., Dunn, A., Rong, Z., Kononova, O., Persson, K. A., Ceder, G., & Jain, A. (2019). Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature*, *571*(7763), 95–98.

Vaishnav, B., & Ray, S. (2023). A thematic exploration of the evolution of research in multichannel marketing. *Journal of Business Research*, *157*, 113564.

Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-sne. *Journal of machine learning research*, *9*(11).

Viet, N. T., Kravets, A., & Duong Quoc Hoang, T. (2021). Data mining methods for analysis and forecast of an emerging technology trend: A systematic mapping study from scopus papers. *Artificial Intelligence: 19th Russian Conference, RCAI 2021, Taganrog, Russia, October 11–16, 2021, Proceedings 19*, 81–101.

Watson, D., Hubbard, B., & Wiese, D. (2000). Self-other agreement in personality and affectivity: The role of acquaintanceship, trait visibility, and assumed similarity. *Journal of personality and social psychology, 78*(3), 546–558.

Watts, D., & Strogatz, S. (1998). Collective dynamics of 'small-world' networks. *Nature, 393*(6684), 440–442.

World Economic Forum. (2024). Top 10 emerging technologies of 2024.

Xian, G., Paul X., M., Colin, G., Claire, M., & Marian-Andrei, R. (2025). Cosmos 1.0: A multidimensional map of the emerging technology frontier.

Yamada, I., Asai, A., Sakuma, J., Shindo, H., Takeda, H., Takefuji, Y., & Matsumoto, Y. (2020, October). Wikipedia2Vec: An efficient toolkit for learning and visualizing the embeddings of words and entities from Wikipedia. In Q. Liu & D. Schlangen (Eds.), *Proceedings of the 2020 conference on empirical methods in natural language processing: System demonstrations* (pp. 23–30). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-demos.4

Yasseri, T., Sumi, R., & Kertész, J. (2012). Circadian patterns of wikipedia editorial activity: A demographic analysis. *PloS one, 7*(1), e30091.

Youyou, W., Kosinski, M., & Stillwell, D. (2015). Computer-based personality judgments are more accurate than those made by humans. *Proceedings of the National Academy of Sciences, 112*(4), 1036–1040.

Zahra, S. A., & Wright, M. (2016). Understanding the social role of entrepreneurship. *Journal of management studies, 53*(4), 610–629.

Zhang, Y., Porter, A. L., Hu, Z., Guo, Y., & Newman, N. C. (2014). "term clumping" for technical intelligence: A case study on dye-sensitized solar cells. *Technological Forecasting and Social Change, 85*, 26–39.

Zhao, H., & Seibert, S. E. (2006). The big five personality dimensions and entrepreneurial status: A meta-analytical review. *Journal of applied psychology, 91*(2), 259–271.

Zhong, L., Wu, J., Li, Q., Peng, H., & Wu, X. (2023). A comprehensive survey on automatic knowledge graph construction. *ACM Computing Surveys, 56*(4), 1–62.