Université Paris Cité
Ecole Doctorale 564 - Physique en Ile-de-France
Laboratoire Interdisciplinaire des Energies de Demain

Complex networks in entrepreneurial ecosystems : clustering methodologies and topological structure

Thèse de doctorat en Physique
Présentée par Théophile Carniel
Dirigée par José Halloy

Soutenue le 19 juin 2024

Jury :

Paola Tubaro	Directrice de Recherche	CNRS	Rapporteur
Joachim Henkel	Professeur	Technische Universität München	Rapporteur
Stéphane Douady	Directeur de Recherche	CNRS	Examinateur
José Halloy	Professeur	Université Paris Cité	Directeur de thèse
Jean-Michel Dalle	Professeur	Sorbonne Université	Membre invité

Contents

List of Figures

List of Tables

Summaries

Short summary

Title : Complex networks in entrepreneurial ecosystems : clustering methodologies and topological structure

Venture capital, through choices in allocating financial capital, has become an important driver of emerging technologies. The impact of venture capital funding on innovation has been evidenced, and recent works have started to more specifically study the resilience of VC-backed innovation, a matter of particular relevance in a world where crises have become more and more frequent. Furthermore, innovation networks (modeled through patent data) on the one hand and venture capital networks on the other have been the subject of quantitative investigation, but there has been no endeavor to investigate the structure of the network linking venture capital to the technologies they fund. To study its topological structure, we perform large-scale analysis of financial, startup and patent datasets obtained through commercial databases.

The network linking investors and patents is bipartite, with the two node classes being the investors and the patents. This network is very large and sparse (roughly 240 000 investors and 870 000 patents), making its analysis computationally difficult, and does not take into account the fact that investors belong to distinct types and patents to technological categories. We remediate this by regrouping homogeneous nodes in each class (clusters) in order to create a coarser-grained view of the network, resulting in a much smaller and denser network and facilitating the analysis by studying the behaviors of groups of similar actors rather than individuals (network of observations at the species level rather than at the individual level).

In the first part, we devised a novel clustering method for venture capital investors in entrepreneurial ecosystems. We computed 5 characteristic distributions for each individual investors based on their investments, and computed the pairwise similarity for all investors. We then detected clusters based on the similarity graph linking all investors, uncovering highly interpretable homogeneous investor clusters heterogeneous in size. We showed that this approach was robust to feature decimation, yielding similar high-level clusters when clustering only on a subset of the 5 characteristic distributions, suggesting underlying complex investment patterns. These results provided us with insights into the emergence of new actors of venture capital following events such as the 2008 financial crisis or the 2013 venture frenzy.

In the second part, since patent data is mostly textual information, we presented a topic modeling method that automatically extracts groups of thematically similar documents from a corpus of text documents. To validate it, we tested it on a smaller corpus of documents similarly-structured to patents : scientific articles. We used natural language processing models to automatically extract research topics from the titles and abstracts of the articles and analyzed the results.

In the third part, we presented a study of the startup-led innovation funding ecosystem. We built a bipartite network directly linking investors to patents owned by the startups they fund. We leveraged the approaches previously described to cluster investor nodes and patent nodes, creating a coarser-grained view of the network. Using structural metrics originally developed to study bipartite ecological networks, we found this network to be

topologically mutualistic, with a heterogeneous degree distribution, a high nestedness and a low modularity. This specific structure is due to the prevalence of links between generalist investors and general purpose technologies i.e. technologies with a broad spectrum of applications. This network structure implies non-linear response to crises, with the system weakly affected by negative events affecting specialist nodes and strongly affected by negative events targeting generalist nodes.

Keywords : *clustering, complex networks, complex systems, entrepreneurial ecosystems, entrepreneurship, innovation, scientometry, startups, venture capital*

Résumé succinct

Titre : Réseaux complexes dans les écosystèmes entrepreneuriaux : méthodes de partitionnements de données et structure topologique

Le capital-risque, par ses choix d'allocation de capital financier, est devenu un moteur important des technologies émergentes. L'impact du financement par le capital-risque sur l'innovation et la résilience de l'innovation soutenue par le capital-risque ont été récemment étudiés, une question particulièrement pertinente dans un monde où les crises sont de plus en plus fréquentes. De plus, les réseaux d'innovation d'une part et les réseaux de capital-risque d'autre part ont fait l'objet d'études quantitatives, mais la structure du réseau reliant le capital-risque aux technologies qu'il finance n'a jamais été étudiée. Afin d'étudier sa structure topologique, nous effectuons une analyse à grande échelle de données financières, de startups et de brevets obtenues à partir de bases de données commerciales.

Le réseau reliant les investisseurs et les brevets est bipartite, les deux classes de noeuds étant les investisseurs et les brevets. Ce réseau est grand et éparse, rendant son analyse computationellement difficile, et ne tient pas compte du fait que les investisseurs appartiennent à des types distincts et les brevets à des catégories technologiques. Nous y remédions en créant des groupes de noeuds homogènes dans chaque classe (clusters) afin de créer un réseau à plus gros grains, permettant de réduire la taille et de densifier le réseau.

Dans la première partie, nous présentons une nouvelle méthode de clustering pour les investisseurs en capital-risque dans les écosystèmes entrepreneuriaux. Nous calculons 5 distributions caractéristiques pour chaque investisseur individuel basées sur leurs investissements. En utilisant le graphe de similarité reliant tous les investisseurs, nous détectons des communautés d'investisseurs homogènes hautement interprétables. Nous montrons la robustesse de cette approche à la décimation des caractéristiques, suggérant des schémas d'investissement complexes sous-jacents. Ces résultats permettent également l'observation de l'émergence de nouveaux acteurs du capital-risque à la suite d'événements spécifiques et identifiables.

Dans la deuxième partie, du fait de la nature textuelle des données de brevets que nous souhaitons regrouper, nous présentons une méthode de topic modeling permettant l'extraction automatique de groupes de documents thématiquement similaires à partir d'un corpus de documents textuels. Pour la valider, nous l'appliquons à un corpus de plus petite taille composé de documents présentant une structure similaire à celle des brevets : des articles scientifiques. Nous utilisons des modèles de traitement du langage naturel pour extraire automatiquement les thèmes de recherche à partir des titres et résumés des articles,

et analysons les résultats.

Dans la troisième partie, nous présentons une étude de l'écosystème de financement de l'innovation menée par les startups. Nous construisons un réseau bipartite reliant directement les investisseurs aux brevets détenus par les startups qu'ils financent. Nous nous sommes appuyés sur les approches décrites précédemment pour regrouper les noeuds des investisseurs et les noeuds des brevets, créant ainsi un réseau à plus gros grain. En utilisant des métriques structurelles développées à l'origine pour étudier les réseaux bipartites en écologie, nous avons constaté que ce réseau est topologiquement mutualiste, avec une distribution des degrés hétérogène, une forte imbrication et une faible modularité. Cette structure spécifique est due à la prévalence des liens entre investisseurs généralistes et technologies génériques, c'est-à-dire des technologies ayant un large spectre d'applications. Cette structure de réseau implique une réponse non linéaire aux crises, avec un système faiblement affecté par les événements délétères touchant les noeuds spécialistes et fortement affecté par les événements négatifs ciblant les noeuds généralistes.

Mots-clés : *capital-risque*, *clustering*, *écosystèmes entrepreneuriaux*, *entrepreneuriat*, *innovation*, *réseaux complexes*, *scientométrie*, *startups*, *systèmes complexes*

Résumé substantiel

Les startups représentent aujourd'hui un moteur important d'innovation technologique, d'emploi et de croissance économique [87]. On peut par exemple citer le succès massif récemment obtenu par BioNTech, une startup basée à Mayence, en Allemagne [208]. L'entreprise, fondée en 2008, a participé au développement des vaccins Covid-19, permettant de sauver environ six millions de vies [291]. La ville de Mayence était à l'époque confrontée à de graves difficultés économiques, menant à un taux de chômage élevé et une dette municipale importante. Le succès de BioNTech a donné un nouveau souffle à l'économie locale, effaçant la dette de la ville grâce à des taxes exceptionnelles et permettant une réduction des taux d'imposition, attirant ainsi d'autres entreprises dans la région et suscitant la création d'une nouvelle génération de startups [232].

Définir précisément la notion de startup n'est pas simple [95, 245, 71]. Celle-ci est cependant intrinsèquement liée à l'expérimentation de nouvelles activités dans des marchés émergents où l'évaluation du risque est complexe, couplée à la recherche d'un modèle commercial industrialisable et rentable à fort potentiel de croissance[1]. Elle diffère donc fondamentalement des entreprises traditionnelles en ce sens que, là où les entreprises classiques optimisent un modèle commercial existant afin de maximiser les profits et de croître organiquement, les startups expérimentent plutôt différents modèles commerciaux à travers le test de marchés potentiels, évoluant de manière itérative au cours de ce processus jusqu'à trouver une adéquation entre leur produit et leur marché cible. La nature innovante de cette approche s'accompagne donc d'une grande incertitude [95], avec une rentabilité souvent atteinte plusieurs années après la création de l'entreprise. Ce mode de développement particulier se distingue ainsi des modèles traditionnels de création d'entreprise, dans lesquels l'évaluation des risques est relativement simple et systématisée. Cette différence a de fortes implications notamment en termes de financement des nouvelles entreprises.

[1] https://bpifrance-creation.fr/moment-de-vie/quest-ce-quune-startup

Les entreprises traditionnelles avec un modèle économique bien établi peuvent par exemple être financées au travers de fonds privés ou de prêts bancaires dans un premier temps, leur permettant ainsi d'amorcer leur activité jusqu'à ce qu'elle commence à générer des revenus en un temps relativement court. Les startups, en revanche, doivent obtenir un financement externe afin de financer leur croissance et leur développement pendant plusieurs années du fait des fortes incertitudes technologiques ou de marché qui les caractérisent. Ce type de financement spécifique est généralement fourni par des sociétés de capital-risque ou des particuliers qui apportent à une entreprise à haut risque et à haut rendement les fonds nécessaires pour soutenir sa croissance en échange de parts de propriété de l'entreprise. La notion d'entrepreneuriat, tout comme la notion de startup, peut ainsi également être défini de différentes façons [96], l'un de ses principes fondamentaux étant « la poursuite d'une opportunité au-delà des ressources contrôlées » [281]. Conformément à cette définition, nous nous concentrons dans le cadre de cette thèse sur l'entrepreneuriat dans le contexte de startups soutenues par le capital-risque, où les entreprises à fort potentiel mobilisent des ressources extérieures par le biais de financements de capital-risque.

Dans ce contexte, des approches systèmes complexes ont été appliquées aux écosystèmes entrepreneuriaux [259, 148, 212, 187], les études sur l'entrepreneuriat bénéficiant de l'incorporation de la myriade de facteurs qui affectent les écosystèmes entrepreneuriaux. Cette constatation a conduit les chercheurs à essayer d'intégrer un plus large éventail de facteurs dans leurs analyses afin de mieux comprendre les mécanismes sous-jacents à l'entrepreneuriat (tels que les facteurs économiques, socioculturels ou psychologiques [293]). Ces efforts, cependant, sont restés principalement théoriques ou résultent souvent d'études sur des échantillons restreints et spécifiques, appelant à des travaux supplémentaires pour les confronter à des données expérimentales afin de les transformer en conclusions exploitables dans un contexte plus général. Des avancées dans ce sens ont récemment été réalisées, conduisant notamment à la naissance de la *Science des startups* [209]. Bien qu'il ne s'agisse pas d'une rupture radicale avec les méthodes d'analyse existantes dans la littérature entrepreneuriale, cela représente un changement de paradigme significatif par rapport à la littérature économique existante, rappelant celui amorcé par la *Science de la science* dans le domaine de la métascience [110] : des perspectives nouvelles avec des applications directes peuvent être obtenues en étudiant l'entrepreneuriat de manière systémique à travers l'analyse de bases de données massives. Cette discipline requiert une méthodologie, des concepts et des outils spécifiques basés sur des concepts théoriques inspirés de disciplines connexes telles que l'économie, l'informatique, la gestion ou encore la psychologie couplés à des analyses quantitatives à grande échelle d'ensembles de données financières, de startups et de brevets obtenus à travers de bases de données commerciales existantes [85, 249].

Des analyses de réseaux complexes ont ainsi été appliquées à ces ensembles de données, avec par exemple l'analyse de réseaux d'interactions entre investisseurs ou entre brevets, mais un ensemble de réseaux liant entre elles ces différentes composantes n'ont –à notre connaissance– jamais été étudiés, notamment le réseau reliant les investisseurs de capital-risque aux technologies qu'ils financent. La caractérisation de la structure topologique de ces réseaux permettrait ainsi de bénéficier des résultats obtenus dans d'autres contextes tels que l'écologie ou d'autres systèmes socio-économiques où les liens entre la structure du réseau et ses conséquences sur le système considéré ont été étudiés, suggérant ainsi un ensemble de propriétés du système pour une structure donnée. L'analyse quantitative des

réseaux entrepreneuriaux présente cependant plusieurs difficultés : leur taille importante présente des contraintes computationnelles, leur forte éparsité requiert des outils spécifiques et l'analyse des données à l'état brut ne prend pas en compte l'appartenance des noeuds individuels à des espèces d'investisseurs et de technologies définies. Pour pallier cela, nous proposons de grouper les noeuds individuels de chaque classe en groupes homogènes (clusters) afin de créer une représentation à plus gros grain du réseau investisseurs-brevets de taille plus faible et de densité plus importante, permettant ainsi la caractérisation à l'échelle du réseau. Ce regroupement utilise les caractéristiques intrinsèques à la nature des noeuds, et différentes méthodologies sont ainsi utilisées pour regrouper les investisseurs similaires et les brevets similaires.

Dans la première partie, nous avons conçu une nouvelle méthode de partitionnement des investisseurs en capital-risque. Nous calculons des distributions caractéristiques pour chaque investisseur individuel suivant ses investissements, et nous mesurons la similarité par paire entre tous les investisseurs. Nous détectons ensuite des communautés basées sur des relations de similarité entre investisseurs. Nous avons montré que cette méthode permettait de découvrir des clusters d'investisseurs homogènes de taille hétérogène très interprétables. En outre, nous avons également montré que cette approche était robuste à la décimation des caractéristiques lors des calculs de similarité, avec des clusters similaires obtenus en prenant en compte l'intégralité des caractéristiques d'investisseurs ou en prenant simplement en compte une partie d'entre elles. Ces résultats suggèrent l'existence de comportements d'investissement complexes sous-jacents, permettant ainsi l'identification de communautés d'investisseurs spécialisées sectoriellement (par exemple dans le domaine de la santé), géographiquement (ciblant très majoritairement certains pays tels que la Chine) ou encore temporellement sans que ces facteurs ne soient pris en compte dans la caractérisation des investisseurs individuels. L'analyse de ces résultats nous a également permis de comprendre l'émergence de nouveaux acteurs du capital-risque à la suite d'événements spécifiques tels que la crise financière de 2008 ou la frénésie du capital-risque de 2013. En outre, cette approche de partitionnement fournit un outil méthodologique solide pour pallier la rareté des interactions et l'hétérogénéité des noeuds dans les réseaux de capital-risque, représentant ainsi une étape importante dans l'étude de leur dynamique.

Dans la deuxième partie, du fait de la nature textuelle des données de brevets, nous avons présenté une méthode de modélisation thématique qui crée automatiquement des groupes de documents thématiquement similaires à partir d'un corpus complet. Cette méthodologie se base sur des avancées récentes en terme d'apprentissage machine, notamment en traitement automatique du langage naturel. En appliquant des modèles de langage basés sur l'architecture Transformer spécifiquement entraînés sur des corpora en lien avec la nature des données à traiter (par exemple littérature scientifique ou brevets dans nos cas d'usage), il est possible d'obtenir, pour chaque document, une représentation vectorielle rendant compte du contenu textuel de chaque document. Cette représentation vectorielle permet ainsi d'appliquer une étape de réduction de dimensionnalité suivie d'un algorithme de clustering basé sur la densité dans l'espace vectoriel, extrayant ainsi automatiquement des clusters de documents traitant de thèmes similaires. Afin de valider cette méthodologie, nous l'avons testée sur un corpus de documents relativement restreint présentant une structure similaire à celle des brevets : les articles scientifiques. Nous avons extrait un corpus d'articles de journaux sur la bioinspiration et la biomimétique, un sous-ensemble de la littérature hautement interdisciplinaire. A travers la méthode de modélisation thématique,

nous avons extrait automatiquement les sujets de recherche contenus dans l'ensemble du corpus à partir des titres et des abstracts des articles de chacun des documents. Nous avons caractérisé et présenté chacun des thèmes de recherche découverts automatiquement dans chacune des sources à travers à la fois une étude manuelle et une étude quantitative, et analysé leurs intersections entre les différentes sources. Nous avons également examiné les tendances thématiques de publication en étudiant l'évolution du nombre d'articles dans chacun des thèmes de recherche. Nous avons ainsi obtenu un aperçu de l'état de l'art de la littérature sur la bioinspiration et la biomimétique, et validé la qualité de la méthode de modélisation thématique sur un corpus de taille modérée à travers une étude manuelle et quantitative en préparation de son application sur l'ensemble de données de brevets.

Dans la troisième partie, nous avons présenté une étude de l'écosystème du financement de l'innovation par les startups. Nous avons construit un réseau bipartite reliant directement les investisseurs aux brevets détenus par les startups qu'ils financent. Nous nous sommes appuyés sur les approches décrites précédemment pour détecter les communautés d'investisseurs et utiliser la modélisation thématique pour regrouper les brevets, créant ainsi une représentation du réseau à plus gros grain, réduisant sa taille, son éparsité et augmentant l'homogénéité des noeuds. En utilisant des mesures structurelles développées à l'origine pour étudier les réseaux écologiques bipartites, nous avons pu caractériser sa structure topologique et la relier à des structures communément étudiées en écologie : ce réseau est topologiquement mutualiste, présentant une distribution des degrés hétérogène, une forte imbrication et une faible modularité. La présence de cette structure spécifique dans le réseau investisseur-brevets, avec notamment un réseau significativement imbriqué, peut être expliquée par la prévalence de liens entre investisseurs généralistes et technologies à usage général, c'est-à-dire des technologies ayant un large spectre d'applications. L'analyse des réseaux mutualistes en écologie a montré que cette classe de réseaux présente une réponse non-linéaire aux crises (extinctions ou attritions), le système étant faiblement affecté par les événements négatifs touchant les noeuds spécialisés et fortement affecté par les événements négatifs ciblant les noeuds généralistes.

Introduction

Startups have become an important driver of technological innovation, employment and economic growth [87], attempting to solve a number of today's most challenging problems. One recent example is the massive success achieved by BioNTech, a startup based in Mainz, Germany [208]. The company, founded in 2008, was involved in the development of Covid-19 vaccines, saving an estimated six million lives [291]. At a time where the city of Mainz was faced with severe economic challenges, giving rise to unemployment and significant municipal debt, the success of BioNTech proved to be a strong agent of change : it breathed new life into the local economy, clearing the city's debt through windfall taxes and enabling a reduction of tax rates which attracted other businesses to the region and sparked the creation of a new generation of startups [232].

Precisely defining what constitutes a startup has proven to be challenging [95, 245, 71]. The notion of startup, however, can intrinsically be linked to the experimentation of new activities in emerging markets where assessing risk is difficult, seeking an industrializable and profitable business model that allows for scalable growth[2]. It thus fundamentally differs from traditional companies in that, where regular companies optimize an existing business model in order to maximize profit and grow, startups instead experiment with different business models through market testing, iteratively evolving in the process until product-market fit is figured out. The innovative nature of this approach comes with high uncertainty [95], with profitability often being achieved several years after the company has been founded. This particular mode of development is markedly different from traditional models of company creation where risk assessment is relatively straightforward.

This has strong implications, notably in terms of financing new activities. Traditional companies that have an established business model can be financed by private funds or bank loans to kick-start their business, which will be able to start generating revenue relatively quickly. Startups, on the other hand, have to secure external funding to finance their uncertain growth and development for several years. This specific type of funding is generally provided by venture capital firms or angel investors that provide a high-risk high-reward company with the required funds to sustain its growth in exchange for equity, *i.e.* shares of ownership of the company.

Entrepreneurship can also be defined in a number of different ways [96], with one of its central tenets being "the pursuit of opportunity beyond resources controlled" [281]. In keeping with this definition, we will, in the course of this thesis, focus on entrepreneurship in the context of venture-capital backed startups, where high-potential companies mobilize outside resources through venture capital funding.

This "pursuit of opportunity beyond resources controlled" implies interactions between startups and other relevant actors (such as investors, human capital, lawyers, academia, addressable markets and policymakers), naturally introducing the notion of "entrepreneurial ecosystems" [278] which extends the ecological ecosystem concept to the entrepreneurial context where interactions in a geographically-defined location (such as a regional unit)

[2] Translated from https://bpifrance-creation.fr/moment-de-vie/quest-ce-quune-startup

are taken into account in order to characterize the potential of a region to foster startup-led innovation [46].

In entrepreneurship, success is not the norm[3], and cases such as the one of BioNTech are rather few and far between. As is the case with a large number of socio-economic systems, however, there is a strong non-linearity : a small number of young, high-growth firms are responsible for a large share of the economic outcomes [87]. Understanding these relatively new economic actors capable of quickly spurring significant growth and the circumstances that can help them succeed has thus attracted the attention of the scientific community, with entrepreneurship research recently gaining considerable prominence in leading management journals [321, 11].

This growing interest mostly comes from the economic and management scientific communities, but is increasingly attracting practitioners from varied disciplines. Machine learning methods, for instance [58, 114, 105], are now being used to predict startups' future outcomes [37], the end goal being the detection of successful ventures as early in a company's lifecycle as possible. Even though these predictive models have achieved significant progress and display promising results [248], there remains substantial room for improvement [113] due to a fundamental characteristic of the system : entrepreneurship, as the outcome of a large number of diverse interacting components [57], is complex [227], with venture successes depending on a large number of intangibles [122]. This is a natural consequence of the thousands of different agents (e.g. investors, startups or the human capital that compose them) directly and indirectly interacting within entrepreneurial ecosystems. Direct funding interactions between a company and a given investor in the investor-startup bipartite network, for instance, can also give rise to indirect interactions such as driving investors away from competing organizations. From the investors' point of view, improving the performance of predictive models pertaining to entrepreneurship thus requires a finer understanding of the complex relationships governing the environment startups evolve in. From the public policy makers' point of view, creating and supporting local entrepreneurial ecosystems which have *de facto* become major components for economic growth requires an understanding of the ingredients and relationships necessary for their success.

In this context, complex systems approaches have been applied to entrepreneurial ecosystems [259, 148, 212, 187] but have gained relatively little traction, in particular in terms of quantitative investigations. Entrepreneurship-related studies benefit from taking into account the myriad of factors that affect entrepreneurial ecosystem, effectively extending the system-environment boundary. This realization has led researchers to continuously try to incorporate a wider range of factors in their analyses in order to better understand the drivers of entrepreneurship (such as economic, socio-cultural or psychological factors [293]). These efforts, however, have mainly been theoretical or result from studies on small and specific samples, calling for further work that confronts them with experimental data in order to translate them into actionable insights. Progress in this direction is being made, recently leading to the birth of the so-called *Science of Startups* [209]. While it does not represent a drastic departure from existing methods of analysis in entrepreneurship, this represents a significant paradigm shift from the existing economic literature, reminis-

[3] 20% of new businesses fail in the first year following creation, and 50% in the five years according to the US Bureau of Labor Statistics (https://www.bls.gov/bdm/us_age_naics_00_table7.txt)

cent of the one pushed by the *Science of Science* in the field of metascience [110] : new and powerful insights can be gained by systematically studying entrepreneurship through massive data-based studies. This field of study requires specific methodology, concepts and tools, drawing from and validating theoretical concepts drawn from many other disciplines such as economy, management or psychology leveraging the power of large-scale data.

Indeed, innovation economics argue that economic development is the result of the diffusion of knowledge, innovation and entrepreneurship within an institutional environment of systems of innovation [78]. In this paradigm, entrepreneurship functions as an open and complex system where the determinants of success are varied : personal factors, government-related factors, education and training, access to finance, cultural factors and economic factors all play a role in setting entrepreneurs up for success [322]. In a certain way, this approach can be seen as trying to peer inside the economists' "black box" [257] where the process of innovation transforms the inputted resources into outputted products. Contrary to the – admittedly over-the-hill – view of [250], the connection between economics and scientific and technological progress is not treated as "hopelessly obscure", but rather as a fundamental component of the creation process : the exogenous element of science and technology flows into the black box, modifying the input-to-output ratio of the system [9]. A perfect understanding of the direct and indirect dynamic interactions resulting from the introduction of new technologies is thus possible only when the complete set of structural elements of the system into which new technologies are being introduced is properly accounted for, *i.e.* when the structure of underlying relations connecting all components of the system is taken into account [9]. In practice, this is only partly achievable due to a number of technical limitations, both in terms of methods and access to the required data; steps should – and have, to some extent – nevertheless be taken in that direction.

It is for instance well-known that the origin of many technological innovations developed by startups can be found in academic laboratories [226], where the process of technology transfer [313] brings technologies developed through fundamental research to applications. Indeed, the existence of fully disclosed, cutting-edge technologies represents an externally-financed bedrock for companies to iterate upon in order to create products that were previously out of technological reach. Disruptive waves of innovation [40] can be traced back for centuries and tend to happen when a particular technological lock is undone, such as portable electronics resulting from the development of lithium-ion batteries [330]). Even a tool that now seems relatively trivial such as the first commercial versions of the PageRank algorithm ubiquitously used by the general population through the Google search engine finds its origin in a research program at Stanford University [45]. The algorithm in its applied version is the outcome of – at the time – recent research on Web Search Engines and of decades of academic research in the fields of citation analysis. This is one of the many examples of now-omnipresent products finding their origin in academia. As [40] put it, "small, hungry organizations are good at agilely changing product and market strategies", making startups particularly suited to perform this task of technology transfer : when facing the uncertainty of an ill-defined product and target market, as is often the case in the first stages of pushing a technology from the laboratory to application, agility and adaptability are vital and necessary skills [277].

Scientific discovery and innovation thus form a cohesive and mutually beneficial rela-

tionship [305] : innovation through the application of these scientific advances will often face strong challenges, and the feedback collected as innovators manage – or fail – to get through these obstacles will feed academia with novel methods, tools and research questions, in turn leading to more potential applications. This feedback loop suggests that valuable insights into entrepreneurial dynamics could be gained through the analysis of scientific production and its links with innovation, observable for instance through patent filings and ownerships or human capital mobility between academia and industry. Although studies have analyzed the relationships between patents and articles [202, 150], the network of technology transfer – modeled by the network connecting startups, patents and academic publications – has not been subjected to direct, large-scale quantitative studies. The extent to which startups in different sectors rely on academic production, for instance, is therefore not well known, nor the manner in which the dynamics of academia percolate until the entrepreneurial sphere is reached. The study of entrepreneurship would thus benefit from a quantitative interdisciplinary approach drawing concepts, methods and tools from a number of disciplines (such as economics, innovation, complex networks and data science) that can yield novel and powerful insights complementary to the existing qualitative literature on the topic.

This approach has proven fruitful in the study of other areas of economics : in economic geography, for instance, the economic complexity approach [136] estimates the innovative capacity of geographical units based on their current assets. Using network-based approaches on large product export databases, the breadth and depth of a region's implicit knowledge is quantified through its local production and characteristics related to the products themselves. This embedded knowledge, in turn, has consequences on the region's ability to acquire new knowledge and enter new industries, in line with Kauffman's *adjacent possible* [163], "a kind of shadow future, hovering on the edges of the present state of things". This characterization of the economic complexity of a region thus gives measurable and actionable insights [4] into how the development of specific products can bring in new knowledge and helps estimate the possibilities and cost of bridging the gap between the current knowhow the location possesses and new and more complex products, providing decisionmakers with information that helps match the region's current production capacities with the potential growth opportunities best suited for it.

In the context of entrepreneurial ecosystems, venture capital firms, which play a crucial and well-defined role in fostering innovation, can provide companies with the means to reach beyond the adjacent possible : as nascent companies developing breakthrough technologies depend on outside financing during a large part of their early stages of research and development, investors have a direct role in supporting emerging technologies. Even though this role has recently been the subject of quantitative investigations[84, 145], there is –to our knowledge– no explicit study of the structure of venture-backed innovation as modeled by the network linking venture capital firms with the patents they indirectly finance. Due to the informationally incomplete nature of investment decision-making processes in a constantly and rapidly changing environment with relatively low legal regulation [43], this network emerges in a largely self-organized manner. As pointed out by several researchers on socio-economic systems [205, 184] and as has been extensively studied using the complex network framework in other disciplines [199], the structural properties

[4] https://atlas.cid.harvard.edu/

of a network have strong implications for the system's robustness to exogenous events, but this analysis has only been weakly applied in financial contexts. Given the increasing importance of entrepreneurship in innovation, a better understanding of the venture-backed innovation network's weaknesses and strengths on a structural level is desirable. This gap in the literature, however, is not necessarily surprising. First, the datasets allowing for the construction and analysis of large-scale interactions in private markets are relatively new compared to those in public markets [85], and their scale itself can prove a challenge when applying direct analysis on the graphs resulting from the interactions they collect. These representations give rise to specific bipartite network structures that are typically very large and sparse with hundreds of thousands of nodes in interaction, making direct applications of statistical methods technically challenging and sometimes sub-optimal [218] due to the fact that structure metrics for large networks are not necessarily well-suited for sparse representations [5].

Objectives

In the context of this thesis, we aim to take steps towards the study of entrepreneurial ecosystems using a complex systems approach, developing appropriate tools in the process. We address the following topics :

- building the networks : due to the availability of massive databases, large-scale interaction networks in entrepreneurial ecosystems can be directly studied. Some of them, such as the investor-investor or the investor-startup networks, have been scientifically investigated for over 2 decades whereas others, such as the investor-patent network, have –to our knowledge– never been studied. We propose to build the venture capital-backed innovation funding network through financial and patent databases in order to study its properties.

- dealing with interaction sparsity and large network size : the bipartite interaction networks in entrepreneurship are sparse and heterogeneous due to the high number of nodes on both sides and the comparatively low degree and high idiosyncrasies of individual actors. Quantitative analysis of these large networks (hundreds of thousands to millions of nodes) is thus difficult, both methodologically and computationally. Through clustering, we can create coarser-grained representations of these networks in order to facilitate quantitative analysis by reducing network sparsity and network size while still taking into account node idiosyncrasies. As domain-specific knowledge can be injected to create specific clustering methods that expressly account for the entrepreneurial nature of the data, we develop algorithms specific to this task.

- node class-dependent clustering : due to the different nature of the nodes and to avoid biasing analyses of the coarse-grained networks, we endeavor to perform clustering of the different node classes independently. Indeed, studying the dynamics of

[5] Indeed, sparse network analysis requires dedicated tools such as http://www.small-project.eu/ that typically only handle unipartite graphs

a coarser-grained network, for instance, where we explicitly use dynamical elements of the origin network to perform the coarse-graining can induce biases, and should thus be avoided as much as possible. Furthermore, the varied nature of the nodes suggests that a one-size-fits-all clustering method is not optimal as it would underutilize available information. We thus develop and apply specific clustering methods on the various node types in order to build coarser-grained networks.

- investor-patent network characterization : applying the methods developed in the course of this thesis, we can create a coarser-grained representation of the investor-patent network. This new representation allows for the application of metrics specifically designed to characterize networks at the structural scale, and to find common characteristics of the structure of our network with networks originating from other disciplines. What insights are gained through this structural analysis of the venture capital-backed innovation network ?

Outline

Chapter 1 gives a brief introduction to venture capital, complex systems and networks and existing applications of complex network analysis to venture capital.

Chapter 2 presents a novel clustering approach specifically designed to find homogeneous communities of venture capital investors. The methodology of the clustering is presented, and the communities uncovered are analyzed. We show that these communities are highly interpretable, and show that community allocation of investors is robust to feature subsampling. This feature subsampling analysis reveals communities similar to those detected through the clustering performed with all features, suggesting the presence of significant underlying complex investment patterns. The appearance of new investor communities in the wake of significant shifts in entrepreneurial ecosystems is also observed.

Chapter 3 presents a state-of-the-art topic modeling approach and its application to a subset of the scientific literature on biomimetism, an interdisciplinary field of research. We automatically discover meaningful topics from a diverse collection of journals and conference proceedings, and analyze how the research themes in the different publication media relate to each other. We finally examine the temporal evolution of the number of research papers per research theme to determine research trends.

Chapter 4 presents an analysis of the network linking investors to the technologies owned by the startups they fund through patent ownership. We leverage the methodologies presented in chapters 2 and 3 to create communities of investors and clusters of patents in order to reduce the size and sparsity of the graph. We analyze the structure of the network, and find it to be topologically mutualistic, sharing structural traits with mutualistic networks in ecology. We analyze and discuss the implications of these findings on the robustness of this network to perturbations, such as economic crises.

State of the Art

Chapter 1

State of the art

This section is divided into 3 parts :

- in section 1, the industry of venture capital and the related scientific literature will be presented
- in section 2, parts of the literature on complex systems and complex networks relevant to this thesis will be discussed
- in section 3, applications of complex networks to entrepreneurial ecosystems will be detailed

In relation to the interdisciplinary nature of this thesis, relevant subsets of the scientific production on venture capital and complex networks (and the interaction between the two) have thus been selected in order to give the necessary context to the works presented in this thesis.

1.1 Venture capital

1.1.1 What is venture capital ?

The role of venture capital

Startup companies often need to rely on external sources of capital in order to hold out until their business gets off the ground and starts generating enough revenue to reach sustainability. In order to obtain the necessary resources for their development through this dangerous period of their life cycle, companies can receive backing from the venture capital (VC) industry [117]. VC firms serve as capital providers to companies that could have trouble attracting financing from more standard institutions, such as banks [335]. Indeed, these companies are often young, do not possess much in terms of assets, and can face

uncertainty from a number of sources such as, just to name a few, market size estimation, founders' ability to properly lead the project, product-market fit or gauging existing competition potentially addressing the same market. Companies trying to attract VC financing are often high-risk, high-reward projects that sell equity shares (*i.e.* partial ownership) to VC firms in exchange for the cash needed to finance its development efforts. Investors, in addition to funding, also provide their portfolio companies with other benefits in order to help them succeed [121, 160]: networking opportunities, connections to other portfolio companies and to investment banks, advice to founders and access to service providers (such as human resources, legal, public relations firms). VC-backed companies have thus been shown to outperform non-VC-backed companies, not only due to sorting effects (VC investors only selecting the most promising companies) but also due to positive treatment effects [20, 27, 274]. The end goal for investors is to successfully *exit* the company by selling their shares when the company either raises subsequent funding, gets acquired or undergoes an Initial Public Offering (IPO), selling their shares for more than they initially bought them.

To synthesize, venture capitalists buy a stake in an entrepreneurial team's idea, support the entrepreneurial venture for a relatively short amount of time and exit (hopefully successfully) with the help of an investment banker [335].

The "series" funding

Fundraising for a startup company happens in incremental steps. Note that as each company faces very specific challenges, its circumstances are different : the steps described here are only guidelines and are by no means absolute rules. The mean number of investors and amounts for the different funding rounds are described in Table 1.1.

In their early stages, companies will try to attract *seed financing* or *angel investing* to start developing their projects. Investors will typically acquire 20-30% of the shares of the company. The goal of this fundraising is to hire a few key team members, buy the necessary equipment and start iterating on the product in order to develop a solid proof of concept. At the seed stage, investors have very little information on which to judge the company : the founders' profile and vision, as well as uniqueness of the proposed solution, are major criteria on which investors base their decisions [97]. The risk associated with this uncertainty is higher, and seed investors tend to be actors specializing in this funding stage.

Once progress has been achieved, companies will try to raise their *Series A*. This is a sizeable step-up in terms of amounts raised, and happens when a startup has demonstrated the potential to grow and generate revenue. The company needs to have a viable business model, and will typically use this new capital influx to undertake significant hires, facility and equipment purchases in order to quickly grow the company. Investors will typically acquire 20-25% of the shares of the company at this stage. The uncertainty is lower than during the seed stage due to companies already having elements to show such as a proof of concept, a tentative business model and potentially some revenue; investment decision making at this stage is hybrid, based on the existing assets of the company and its potential of growth (based on its addressed market, founders, product).

Stage of investment	Average amount (million USD)	Average number of investors	Number of rounds
Pre-Seed	0.46	1.89	18207
Angel	1.06	1.87	12299
Seed	1.89	2.88	64640
Series A	11.78	3.15	41048
Series B	24.86	3.74	22254
Series C	42.69	4.34	10496
Series D and above	88.72	4.74	7418

Table 1.1: **Descriptive table of the mean amount raised and mean number of investors for each stage of investment.** Note that the number of investors for the *Angel* stage of investment is not representative of the actual number of individual angel investors, as angels can create structures known as *angel groups* in order to handle their investment. An angel group will only count as a single investor, but in reality corresponds to several individual business angels. Values computed using data from Crunchbase.

Early-stage financing is comprised of funding rounds up to and including series A. Financing after this stage (series B and above) is termed *late-stage* financing.

A company raising *series B* financing is past the initial startup stage. It is becoming a mature company, and has already achieved a number of milestones in developing its business. The investors that fund companies during a series B round are usually specialized in late-stage funding, and acquire roughly 20% of the company. For late-stage investments, investors will base their decisions on concrete elements related to the company [97], such as market acceptance.

The subsequent funding rounds are completely company-specific, general guidelines for these development stages thus tends not be done.

1.1.2 Risk management in venture capital

As venture capital is an inherently risky endeavor, a number of strategies exist in order to mitigate the risk, but usually come at the cost of the potential reward of investments [290].

Syndication

During the funding round of a company, several investors can choose to invest together in order to reach the target funding amount of the company [149]. This process is called syndication, and allows investors to invest in companies when they otherwise would not necessarily be able or willing to by injecting lower total individual amounts. Even though the reward for each investor in the event of a syndicated deal is lower if the company succeeds down the line, the losses incurred by each investor are spread between syndicate members in the case of the subsequent failure of the company. This reduced exposure to risk can be desirable, as it is usually better to limit the negative impacts of a failure than to maximize the benefits of a success [149]. Syndication also allows investors to participate in

a larger number of deals, mitigating their risk through portfolio diversification [144]. Furthermore, syndication allows for stronger shared knowledge during deal selection and better managerial advice to funded companies due to the idiosyncratic skill sets of the various investors [144]. The benefits of syndication thus go beyond mere risk mitigation, improving deal quality and portfolio company support [42].

During a syndicated funding round, there can be one (and seldom several) lead investor that takes charge of the funding round. This is another way of sharing resources between investors during a funding round : a credible lead investor (through past success or personal connections) will be able to better convince other investors to participate in the funding round due to the trust they place in its due diligence process and expertise in selecting investees.

Early-stage vs. late-stage

As described in section 1.1.1, investors use different analysis criteria for potential investments depending on the maturity level of the companies. Coupled with the constraints imposed by the large amounts of cash required to invest in late stage ventures, venture capital firms often specialize in investment stages, either in early stage or late stage [97].

A specific type of early-stage funding model in venture financing is the accelerator model [73]. Accelerators target cohorts of startups in their early stages of development, exchanging company equity against 3-to-6 months intensive development programs in order to help companies achieve very specific objectives (for instance having a functional proof of concept at the end of the acceleration program). Most accelerator programs also offer key resources : small seed funding, personalized coaching from successful entrepreneurs, networking opportunities with other entrepreneurs and specific events where companies get the chance to pitch to investors and attract funding. The accelerator model appeared in 2005 with the creation of Y Combinator, and its massive repeated successes (such as Stripe, Airbnb, DoorDash, Coinbase or Dropbox[1] to cite just a few) led others to replicate the model, with over 8000 accelerators worldwide in 2020. Stage specialization was not found to be significantly positively or negatively related to fund returns [290].

Generalists vs. specialists

Companies can be founded in a variety of sectors, each with their own specific advantages and drawbacks. Hardware-based companies, for instance, will have to deal with the challenges that accompany physical goods, such as production, storage, supply chain and transport [116]. Biotech and medtech companies, on the other hand, are confronted to the difficulties specific to living organism, for instance stringent regulations, biological constraints and long development cycles due to the regulatory and technological complexity of the products [271]. Software-based companies have their own specific challenges, where barriers-to-entry can often be comparatively lower and thus put greater emphasis

[1] https://stripe.com/, https://www.airbnb.com/, https://www.doordash.com/, https://www.coinbase.com/, https://www.dropbox.com/

on founders' ability to execute, *i.e.* make the right decisions in order to grow quicker than their competition [190].

Due to the diversity of challenges, investors can draw benefits from developing expertise in specific sectors (specialization), thus increasing their ability to accurately gauge the viability of investment opportunities and offer better assistance to their portfolio companies. As always, specialization comes at a price : if the specific favored sector of investment sees a dearth of activity compared to the general entrepreneurial landscape, their existence can be in peril as novel opportunities would be scant. This gives rise to a trade-off between fund performance and risk management in terms of sectoral specialization. Sectoral portfolio diversification is thus a lever of action in order to modulate reward as a function of the risk [141]. Diversification in a limited number of industries was found to be positively correlated with fund performance [290].

Geographical specificity

The entrepreneurial ecosystem notion, which implies a systemic and geographically-constrained view of entrepreneurship, is not new and can even be dated as far back as the 1920s [201] under the identity of "industrial -districts", gaining significant traction in the second half of the 2000s [64]. Leading entrepreneurial ecosystems (such as the Silicon Valley on the West Coast of the United States, the Boston area in the East Coast or the Greater London area in Europe) are generally located around major cities able to attract all the key components necessary for the creation of a vibrant entrepreneurial ecosystem [294].

Due to the existence of a strong local geographical footprint in entrepreneurship, investors, particularly during early stage investments, generally invest in local companies in order to mitigate risk by having better access to their portfolio companies, leading to better advising and management [275]. Indeed, information on a generally spreads locally in social and geographical space [31], a phenomenon also found in the Venture Capital industry [275]. Access to information has strongly been linked with risk mitigation [171], making investments in spatially distant companies inherently more risky. In order to mitigate this risk, investors can leverage the social structure of the venture capital market in order to overcome the informational constraints resulting from geographical distance : investors boasting central positions in the syndication network (built through co-investments between investment firms) can extend their access to information through social ties [275], thus allowing them to invest globally by creating syndicates with trusted investors that they have previously interacted with in the geographical vicinity of the targeted company. Geographical diversification was not found to be significantly related to fund returns [290].

1.1.3 Venture capitalists' decision criteria

Due to the numerous different facets of a venture, investors often focus on specific venture aspects in order to drive investment decision making when selecting deals among all candidates [160, 118]. Some firms focus more on the "jockey" (the management team) and others on the "horse" (the product, technology or business model) [159]. Cross-sectional varia-

tions were still observed, as investors were found to focus on different aspects depending notably on the company's stage of development.

The horse

Business-side investors will prioritise investing in companies that they believe have high-quality products, offering novel and superior services often backed by strong intellectual property, and a strong fit with the market demands. Criteria of particular importance for these investors will be product- and market-related. Investment decisions will hinge more strongly on whether the product is proprietary or can be protected, whether it can be described as "high tech" (with for instance links between the founding team and academia leading to the product genesis) and whether the company has already demonstrated its feasibility and potential value (for instance by showing a minimum viable product) [194, 173]. These investors will also tend to favor ventures targeting markets they have expertise in, paying particular attention to market-based criteria : growth rate of the target market, existing competition, impact of the candidate venture on the market (such as stimulation of the existing market or creation of a new one), market-related experience of the founding team [194, 173, 203]. [118] found that 37% of all firms they surveyed rated business-related factors as the most important factor, but that these factors tended to be valued more highly by investors assessing late-stage ventures. This is coherent with the fact that more business- and market-related information is available for the deal screening process compared to early-stage ventures where information is scarcer and will be highly susceptible to undergo strong change.

The importance of business-related factors compared to management-related factors was also found to be higher in health care investors compared to IT investors, with 55% of the health care subsample selecting business-related factors as the most important compared to 32% selecting team-related factors, a result once again consistent with the particular emphasis put on intellectual property and non-human capital assets in the health care sector. Late-stage investors were also found to put significantly less emphasis on the product compared to early-stage investors, and significantly more emphasis on the company valuation [118]. When a company reaches the late-stage, it has typically gone through all the necessary steps in developing a product and demonstrating its potential, which makes the product quality less of a differenciating factor between deal opportunities : for a company to reach the late stage, a viable product is a necessary condition. Late-stage ventures also offer more financial metrics for venture funds to analyze, and are generally closer to a potential exit from the venture capitalist's point of view. The valuation therefore becomes a bigger focal point, both due to the presence of a larger number of concrete elements to evaluate it and due to it representing a –relatively– short-term goal. This is coherent with the observation that late-stage investors are more *structured* in their investment decision making, using a larger number of metrics (2.4 on average compared to 1.8 for early-stage investors) and making *gut investment decisions* more rarely [118].

The jockey

Some investors, on the other hand, will prioritise investment in companies with what they deem to be a strong foundation team regardless – to some extent – of market fit and product quality, with the belief that a quality founding team will manage to steer the company towards a relevant market and product, especially in the earlier stages of a company's life cycle where agility and experimentation is of particular importance [26]. These investors will pay particular attention to the abilities of the foundation team, such as the capacity to correctly evaluate and react to risk, to make sustained intense effort or to present and defend their venture in an articulate manner ([266], for instance, found that founders displaying high passion significantly increased investor neural engagement and thus interest in the presented venture). The past experiences of the founding team will also be strong drivers of investment decision making, favoring entrepreneurs that have demonstrated leadership ability in the past and that have venture-relevant track records [194, 173]. These factors also present the benefit of being always available for analysis at any point during the life cycle of the venture, starting from its creation. In a low information context that requires decision making such as pre-seed or seed stage investing, this provides the investors with key information on which to judge a candidate deal when data is scarce [58].

Indeed, the quality of the founding team has been consistently found to be of particular importance for early-stage investors. [118], for instance, found that 53% of early-stage investors rated the team as the single most important factor compared to 47% for the complete sample and 39% for late-stage investors. [26], using a randomized field experiment to identify startup characteristics that are most important to investors in early-stage firms, found a strong response from the average investor to information about the founding team, but not to the identities of lead investors or to firm traction.

The importance of the foundation team, even though more pronounced for early-stage investors, is not unique to them. Indeed, a strong foundation team is deemed as essential for most venture capital firms [118], independently of their specialty : [194] found that five of the ten criteria investors most commonly rated as essential are related to the entrepreneurs themselves (consistent with the survey responses of [118] where 47% of VC firms chose the management team as the most essential factor). Indeed, a recent study [195] has shown that, when performing cluster analysis on questionnaires where venture capitalists rated highly successful and unsuccessful ventures on a number of screening and performance criteria, 3 classes of unsuccessful and 4 classes of successful ventures emerged. All 3 unsuccessful classes had a successful pendant, with the major criteria of differentiation between the two being a flaw in the venture team [195].

This impact of founder personality on startup success, while long known to be a major factor of positive outcomes [146], has recently experienced significant progress due to the availability of new methods and data allowing for the large-scale quantification and analysis of founder personality traits and their entrepreneurial outcomes [112, 208]. Using public Twitter data, both studies built startup founders' Big Five psychological profiles and, estimating the founded startups' success, measured the impact of the personality traits on the companies' entrepreneurial outcomes. [112] found startup personality traits were found to be significant across all stages of the company's life cycle, and [208] found that startups

demonstrating larger, personality-diverse teams showed increased likelihood of success. This psychological analysis has also been recently applied to investors themselves [153], with two personality traits (neuroticism and openness) standing out in terms of explanatory power for equity investments, with high neuroticism and low openness being associated with low equity shares *i.e.* the amounts and fraction of their total investments invested in equities.

Financial characteristics

Finally, all investors look at the financial characteristics of a candidate venture. Investors will try to target ventures that can yield a high return of investment in a relatively short timeframe (*i.e.* 10 times the investment within 5-10 years [194]), and that can easily be exited, for instance through an acquisition or an IPO [173]. These findings are congruent with a recent study [195] that found, using factor analysis, that investors tend to screen out ventures where there is a high risk of competitive attack or profit erosion before cash-out and ventures where the investment is locked up for a long period of time. The financial characteristics that investors pay attention to can also vary with the company's stage of development for several reasons : first, as discussed previously, early-stage companies are surrounded with more uncertainty, with fewer financial metrics allowing for analysis. The risk associated with investment in early-stage ventures tends to be higher than for late-stage ventures, but so does the return on investment as shares are acquired for a lower valuation (Peter Thiel, for instance, made a roughly 2200x return by investing around $500k in Facebook and selling it back for $1.1 billions). The portfolio performance of early-stage investment firms tends to be driven by a few very successful investments due to the power law distribution of returns in venture capital [180], leading to the creation of funds specifically targeting such high-risk high-reward ventures (termed *moonshots* in entrepreneurial ecosystems)[2].

1.2 Complex networks

1.2.1 What is a complex system ?

A complex system can be defined as a system composed of many components in interaction with each other [228]. This includes both natural and socio-economic systems, ranging from ecological ecosystems to financial systems, transportation systems, the Internet, the brain or the Earth's climate. Due to their high number of components and the potentially complex nature of their interactions, their study requires specific tools and frameworks as they are - for the most part - not analytically solvable. Complex systems often present two very important properties :

- nonlinearity : cause and effect are disproportionate *i.e.* small changes in the system can have large effects, and large changes in the system can have little to no effect

[2] https://medium.com/leadership-prevails/understanding-the-vc-power-law-why-fund-size-matters-in-venture-capital-returns-b3dcc268150

- emergence : the system as a whole exhibits macroscopic properties that its components do not possess on their own, such as connections between neurons in the brain giving rise to human consciousness

Furthermore, and perhaps most importantly, complex systems are thermodynamically open [143], *i.e.* they are *very* difficult to bound : they interact with a lot of elements not directly studied, and accurately delimiting their perimeter is an arduous task. The scope of a complex system is thus by nature imperfectly defined, and approximations (with the possible limitations that entails) are a necessary component of its study. Indeed, one cannot accurately understand, for instance, the world trade network (*i.e.* imports and exports between countries) without understanding transportation networks and the idiosyncracies of the complex societies making up the various trading partners. Said societies and transportation networks will, in turn, be the consequence of a number of local path dependencies due to historical, social and political reasons that, in order to achieve holistic characterization, would also need to be taken into account, leading to the incorporation of more and more elements into the study of the system of origin (here, the world trade network) as more and more interactions are accounted for. Complex systems are often represented using networks, which try to bring the very properties that make a system complex to the fore to furnish new forms of explanation, rather than using idealization techniques to simplify the system [244].

1.2.2 Networks

Formally, networks are described using the mathematical field of graph theory. A network can be represented by a graph $G = (V, E)$, defined as a set of n_V vertices (or nodes) V and a set of n_E edges (or links) E where $e(i, j) = \{e_1(i, j),\ ...,\ e_n(i, j)\} \in E$ is the set of edges joining vertices $i, j \in V$. Note that, with this definition, $n_E = \sum_{i,j \in V} \text{len}(e(i, j))$ where $\text{len}(e(i, j))$ is the number of elements in set $e(i, j)$. Vertices i and j are said to be *interacting* if there exists an edge linking vertices i and j, and are said to be the *endpoints* of the edge. Here, we will present network theory concepts and tools relevant for the rest of this manuscript.

Different types of graphs

A number of different graphs exist, such as :

- unweighted vs. weighted : all edge weights (the strength of the interaction between the nodes joined by the edge) are the same *i.e.* $e_n(i, j) = 1\ \forall n$ vs. different edge weights are allowed *i.e.* $e_n(i, j) = w$ with $w \in \mathbb{R}$ the weight of the interaction

- undirected vs. directed : all edges are non-directed *i.e.* $e(i, j) = e(j, i)\ \forall i, j \in V$ vs. edges connecting two given nodes are directed and can have different values *i.e.* $\exists\ i, j \in V \mid e(i, j) \neq e(j, i)$

- regular graphs vs. multigraphs : only one edge can join two nodes *i.e.* $e(i,j) = \{e_1(i,j)\}$ vs. multiple edges can join two nodes *i.e.* $e(i,j) = \{e_1(i,j),\ ...,\ e_n(i,j)\}$

This list is, of course, but a tiny fraction of the diverse bestiary of graph classes (see for instance [77] for a more complete overview).

Adjacency matrix

A graph G is described by its associated adjacency matrix $\mathbf{A}_{n_V \times n_V}$, where $\mathbf{A}_{ij} = 0$ if no edge connects vertices i and j and $\mathbf{A}_{i,j} = e(i,j)$ otherwise. In the case of a multigraph, $\mathbf{A}_{ij} = \sum_{n=1}^{N} e_n(i,j)$.

Incidence matrix

A graph G can also be described by its incidence matrix $\mathbf{B}_{n_V \times n_E}$.

For undirected graphs,

$$\mathbf{B}_{ij} = \begin{cases} w \text{ if vertex } v_i \text{ is incident with edge } e_j \text{ with weight } w \\ 0 \text{ otherwise} \end{cases} \tag{1.1}$$

For directed graphs,

$$\mathbf{B}_{ij} = \begin{cases} w \text{ if edge } e_j \text{ exits vertex } v_i \\ -w \text{ if edge } e_j \text{ enters vertex } v_i \\ 0 \text{ otherwise} \end{cases} \tag{1.2}$$

Note that the sign convention is arbitrary, and can be reversed for directed graphs.

Bipartite networks

A specific class of networks is the bipartite (or 2-mode) network. In a bipartite network, nodes fall into one of two defined classes (or guilds) : the *top* and *bottom* nodes, which can only interact with nodes of the other class (*i.e.* top nodes can only interact with bottom nodes and inversely) as a consequence of the *forbidden interactions* in the system [120] (interactions prevented by the specific traits of the nodes, such as physiological or phenological constraints of biological species). Bipartite networks are the natural representation of a number of natural and socio-economic systems, such as for instance plant-pollination interactions [17] (where pollinators can only interact with plants), actor-movie networks [316] (where actors are linked to the movies they acted in) or gene-pathway networks [134] (where genes can only be linked to the specific pathways they were involved in).

Let $\mathbf{A}_{n_V \times n_V}$ be the adjacency matrix of a bipartite weighted graph with $n_V = n_t + n_b$ where n_t is the number of top nodes and n_b the number of bottom nodes. $\mathbf{A}$ is block-diagonal, i.e.

$$\mathbf{A}_{n_V \times n_V} = \begin{pmatrix} 0_{n_t \times n_t} & \tilde{\mathbf{A}}_{n_t \times n_b} \\ \tilde{\mathbf{A}}^T_{n_b \times n_t} & 0_{n_b \times n_b} \end{pmatrix} \tag{1.3}$$

where $\tilde{\mathbf{A}}$ is called the *biadjacency matrix* of the network and $\tilde{\mathbf{A}}^T$ denotes its transpose.

Multilayer networks

Due to the complex nature of the relationships between agents, their interactions can take the form of distinct type. As an example, if we study a telecommunications graph where nodes are individuals and edges are created when two nodes interact through sending each other a message, the channel through which messages are sent can be relevant : email communications might happen in a more professional setting, whereas messages sent through social networks or Short Message Service (SMS) texts tend to happen in a more personal setting. Explicitly discriminating between the different interaction types can thus be valuable, and will be modeled using a specific type of graph called the multilayer graph [168]. In a multilayer network, nodes are fixed but several representations (*layers*) of the graph exist where edges are created between the nodes depending on the interaction type *i.e.* one layer of the graph could link individuals through their email interactions, another layer through their social network interactions and a third layer through their SMS interactions.

1.2.3 Networks in ecology

Beyond the ecosystem concept, a number of network tools developed in ecological research have also been applied to socio-economic systems [199].

Bipartite networks in ecology

Different classes of interactions between species are commonly found in ecological networks throughout habitats and ecosystems. These classes of interactions, amongst which most notably *competition*, *predation* and *symbiosis*, have been widely studied in the ecological literature and fall in one of two categories : *positive interactions*, where both species derive benefits from the interaction, and *negative interactions*, where one of the interacting species is benefited and the other is harmed.

These interactions are defined as follows :

- *Competition* is defined as the negative interaction between individuals vying for a common resource present in limited supply.
- *Predation* is defined as the negative interaction where a predator from a given species kills and eats a prey from the same (*cannibalism*) or a different species.

- *Symbiosis* is defined as an interaction between two species purposefully living in contact with each other. This interaction can be negative (*parasitism*), beneficial for one species involved and neutral for the other subset (*commensalism*), or positive (*mutualism*), with both species deriving benefits from the interaction.

The ecosystems within which these interactions take place can often be described and modeled using a bipartite structure due to the physiological and phenological constraints between the interacting species. An important topic of study in quantitative ecology is the characterization of the structure of these networks (in the mathematical sense). A significant part of bipartite network tools, both analytical and methodological, have thus historically been developed by members of this community [92]. One particular area of interest of this literature is the study of how structural metrics at a network level are linked with their robustness, defined as their stability in response to perturbations [199, 108, 124].

Links with economy and finance

Due to the susceptibility of economic and financial systems to crises and the far-reaching social impacts of their disruption, practitioners of complexity have pushed for regulation increasing their structural robustness, drawing inspiration from ecological systems due to the desirable properties they display and the fundamental similarities they share [183, 184, 205].

Current biological systems are the product of millions of years of evolution and natural selection, surviving through a large and varied range of challenges throughout their history. This repeated selection process has led to the current solutions that are - to a certain extent - robust by virtue of their continued existence through continuous global and local change [205]. These systems, when analyzed through their networks, show remarkable universality in structure. Characterizing the structural attributes shared by these systems and drawing inspiration from their characteristics that provide a high degree of robustness can help design new regulatory frameworks for economic and financial systems, which can face similar challenges. Indeed, a number of fundamental concepts are shared between economic sciences and biological sciences, notably ecology and evolutionary biology.

One of these shared concepts is the existence of a trade-off between exploration and exploitation [214]. Indeed, this trade-off between the exploration phase, which is the process of searching for new optima, and the exploitation phase, which corresponds to the implementation of the current best solutions, is a fundamental component of evolution via natural selection, but also of the behaviours of investors, companies, and financial institutions that need to find a good balance in order to generate sufficient revenue while staying ahead of their competition. In economy and finance as in biology, finding the right balance between exploration and exploitation will depend upon context and is done through trial and error, trying to best navigate the variability and uncertainty caused by endogenous as well as exogenous sources. Due to the capacity of economic agents to predict and adapt to changing conditions, this trade-off is not a simple optimization problem. This is, of course, not a novel finding [191] and has been one of the main arguments against Keynesian economics [165] that ignore the ability of rational economic agents to anticipate, influence and

optimally respond to policy changes. In a different framework, however, this issue can be addressed : the regulators and the regulated do not interact in a vacuum, but are instead part of a much larger ecosystem. Their strategies evolve in response to each other, as a result of the changes they themselves generate.

Indeed, another fundamental concept shared between biological and financial systems is that they are adaptive [188]. Evolutionary insights permit a natural alternative to the concept of the efficient markets hypothesis which has been debated within the economic and finance literature [295]. Rather than prices being the sole driver of investor behaviour, markets can be considered as evolutionarily adaptive systems where the fitness function is related to factors characterizing the local financial ecosystem, such as the makeup of local investors, the available profit opportunities and the local socio-economic context. The adaptive market hypothesis thus provides an explanation for why suboptimal investment strategies may persist in a market over time, and why market conditions may change over relatively brief periods [188].

In order to create a stable and sustainable economy, it is thus necessary to take into account the interactions - and the resulting feedback loops - in the system when devising regulations. Doing so in a proper, meaningful manner is, of course, a tall order, and requires frameworks, data and tools specifically designed to deal with this complexity.

Indeed, the question of complexity is central in socio-economic systems. The continuous cycles of booms, busts and financial crises are as many signs of the unsteady foundation on which our economy is built, of the irrationality of investors and of the fallibility of markets, with catastrophic consequences at a global level [100]. As is typical of complex adaptive systems, a many individual agents collectively perform a large number of actions in the pursuit of their self-interest, whether biological or financial, often giving rise to self-organization and emergent phenomena with unpredictable consequences on the system as a whole. Even though they take different forms, the main ingredients of evolution - reproduction, innovation, selection - also exist in economy and finance. As in biological systems, evolutionary consequences will necessarily follow.

In spite of the high complexity of economic and financial systems, significant progress can be made : using principles and results from complex adaptive systems with equivalent (or higher) complexity such as those found in ecology and evolutionary biology (disciplines which have found successful applications of complexity science [176]) can thus help with promoting economic growth while still maintaining robustness and stability, leveraging their similarities to socio-economic systems in mechanisms and structure.

1.2.4 Complexity, the economy and economic complexity

Even though the economy has long been thought of as complex system [142], empirical investigation in this direction has only been undertaken relatively recently, due notably to the recent increased availability of fine-grained economic data and the computing power and tools required for its study. Under the paradigm of complexity, the multitude of interactions between agents in a given economic system are directly studied through the network they give rise to rather than their aggregates (such as GDP) that are more traditionally used.

Each agent is considered a separate entity with its own idiosyncrasies, contribution and response to systematic change. These methodologies can thus yield results fundamentally different from (and complementary to) those from more traditional methods in economic sciences [213, 263].

Several forays in this direction have already been taken. In [139], the concept of relatedness is applied to country-product export networks in order to measure the compatibility between a given activity and location. This compatibility gives strong insights into the potential of a region to start exporting new products based on the number of related products it already exports. Similarly, an industry will be more likely to develop in a region that already has a number of related industries [225]. The concept of relatedness that was originally developed to study export networks has been shown to be general, yielding insights across different spatial scales, economic activities and institutions [138].

Another application of complexity science to economic systems is the field of economic complexity [137] that was first developed to explain the disparity in economic development between countries. This difference is posited to be partially explained by the presence of non-tradable activities, such as specific human capital, regulation or infrastructure, with the productivity of a country (measured through its export data) being intrinsically linked to the diversity and availability of its capabilities as measured through the economic complexity index [136] (ECI). One significant departure from the existing literature is that, rather than estimating complexity from averaged indicators, the ECI is directly measured from the production network of the geographically-embedded socio-economic system studied (such as product export, employment or innovation), escaping the need to rely on the identification of individual factors of development and economic growth. The presence of the combined factors, regardless of their nature, is instead directly learned from production matrices using Singular Value Decomposition-based [169] dimensionality reduction techniques, which present the strong advantage of being a general method applicable in a wide variety of regions and domains where data is present without requiring any *a priori* assumptions. This index has been shown to accurately capture information about the set of capabilities available in a region and to be predictive of both future growth and of the complexity of a country's future exports [137].

Economic complexity is strongly linked to endogenous growth theory [255, 256], which argues that economic growth is tied to investments in human capital, innovation and the associated knowledge produced. Indeed, knowledge in economy holds a special role, being neither a conventional nor a public good but instead being a non-rival, partially excludable good [255]. In spite of these facilitating properties, spreading knowledge is difficult due to its - almost by definition - complex nature : the acquisition of new knowledge requires a base on which to build upon, and knowledge dissemination is both performed formally by specific institutions (such as universities) and informally through local expertise. For these reasons, expertise in a given region gives rise to path dependencies as new activities depend on the existing production which determines the potential economic futures. Economic complexity methods can in a way be seen as indirectly inferring knowledge carried in a geographical unit by learning the combined factors from the fine-grained input data.

1.2.5 Community detection on networks

Analysis of networks with a large number of heterogeneous nodes can be difficult, as, on the one hand, one cannot account for each node individually but, on the other hand, network-level measures can often be too coarse, washing away the idiosyncrasies of the individual nodes. In spite of this heterogeneity, subgroups of nodes can still display commonalities. In order to reduce the size of the network while still retaining its structure, one could find a way to group similar nodes together, thus creating *communities* comprised of nodes, and analyze both the constitution of these communities and the coarser-grained network resulting from the interactions between these communities based on the interactions of their members.

More formally, community detection is a task where the goal is to find an optimal partition of the network nodes by maximizing a suitable objective function. The scalability of such methods is particularly important, as network sizes can reach thousands or even millions of nodes.

Community detection on unipartite networks

The community detection algorithm introduced in [34] has become one of the standards of community detection on networks due to a number of desirable properties, such as being modularity-based (*i.e.* directly measuring the quality of the partition), strongly scalable and unsupervised. After an initialization where each node is allocated to its own community, the algorithm works iteratively, with each iteration composed of two phases.

In the first phase, for each node i, its neighbors j are considered and the associated potential gain of modularity Q if node i was placed in the community of j is computed following eq. 1.4. Node i is then placed in the community of the neighbor that maximizes the modularity gain if the gain is positive, and remains in its community otherwise. This is done repeatedly and sequentially for all nodes until no further improvement of the network modularity can be achieved.

$$Q = \frac{1}{2m} \sum_{i,j} \left[\mathbf{A}_{ij} - \frac{k_i k_j}{2m} \right] \delta(c_i, c_j) \tag{1.4}$$

where $\mathbf{A}$ is the adjacency matrix of the graph, $k_i = \sum_j \mathbf{A}_{ij}$ is the degree of vertex i, c_i is the community to which vertex i is allocated, δ is the Kronecker delta *i.e.* $\delta(\alpha, \beta) = 1$ if $\alpha = \beta$ and 0 otherwise, and $m = \frac{1}{2} \sum_{ij} \mathbf{A}_{ij}$. The modularity $Q \in [-1, 1]$ measures the density of links inside communities (modules) compared to links between communities.

In the second phase, a new network is built where the nodes are the communities found during the first phase (yielding a coarser-grained view of the network of origin), and the edges between the nodes correspond to the total weight of the edges between all nodes of each community in the original graph. Note that self-loops are allowed in this graph and are used to represent intra-community edges. Once this new network has been built, the first phase of the algorithm is applied on the new network, merging communities together

until the modularity can no longer be increased.

These two phases (local maximization of the modularity, and coarse-graining of the graph) are applied recursively until there are no more changes in the graph, resulting in a maximally modular partition. The algorithm is stopped, and the community allocation of all nodes are returned.

Community detection on bipartite networks

Bipartite networks have a specific structure that must be taken into account in order to design effective community detection algorithms. Indeed, the 2-mode topology of the network adds additional constraints to the partitioning as communities must be comprised of nodes from both modes.

Here, we will discuss LPAwb+ [21] and its variant DIRTLPAwb+ [21], an algorithm specifically designed for community detection on bipartite networks. In the context of this algorithm, a bipartite network can have at most $F = \min(n_t, n_b)$ communities as module is comprised of both top and bottom nodes in a bipartite network (where n_t and n_b is the number of top and bottom nodes as defined in section 1.2.2). The algorithm is initialized by giving a unique community label to each node in the smallest of the two sets.

The first stage of the LPAwb+ algorithm is the label propagation stage, where top and then bottom community labels are asynchronously updated by locally maximizing modularity Q_W (eq. 1.5).

$$\begin{aligned} Q_W &= \frac{1}{2m}\sum_{i=1}^{n_t}\sum_{j=1}^{n_b}(\tilde{\mathbf{A}}_{ij} - \frac{k_i k_j}{2m})\delta(c_i, c_j) \\ &= \frac{1}{2m}\text{tr}(R(\tilde{W} - \tilde{E})C) \end{aligned} \tag{1.5}$$

where tr is the trace, F the number of communities in the network, $R_{F\times n_t}$ the top label matrix, $C_{n_b\times F}$ the bottom label matrix. R and C are binary matrices with non-zero values indicating to which community each top and bottom node belongs; the matricial form allows for more efficient vectorized computation.

For top node i, this can be written as choosing a new label c_i^{new} by finding the label maximizing the condition shown in eq. 1.6.

$$c_i^{new} = \underset{g}{\text{argmax}}\left(\sum_{j=1}^{n_b}\tilde{\mathbf{A}}_{ij}\delta(g, c_j) - \sum_{j=1}^{n_b}\frac{k_i k_j}{2m}\delta(g, c_j)\right) \tag{1.6}$$

The condition maximized for the bottom nodes is similar but summed over all top nodes, as top nodes only use information about the bottom nodes to update their labels and, inversely, bottom nodes only use information about the top nodes.

The updating rules for the top and bottom nodes are applied asynchronously, *i.e.* top labels are updated, then bottom labels, then top labels, and so on until modularity as shown in eq. 1.5 can no longer be increased.

The second stage of the algorithm is the agglomeration stage. Much like in the unipartite graph case outlined above, stage 1 ensures that a local modularity maximum is reached, but does not ensure that the global maximum is found. To avoid getting stuck in a local maximum, two communities are fused together if the fusion would result in an increase in modularity and if there is no other community whose fusion with either of the two communities would result in a larger modularity increase. Stages 1 and 2 are performed until network modularity is maximized, *i.e.* there are no more changes to the modules.

LPAwb+ was found to get stuck in suboptimal solutions with a larger number of modules due to the nature of the algorithm [21]. To remediate this issue, an improved version of the algorithm was designed that makes use of LPAwb+'s sensitivity to node label initialization. DIRTLPAwb+ computes LPAwb+ multiple times with different random initializations which have an increasing number of modules, and returns the solution with the greatest modularity score, resulting in the optimal community partition of a bipartite network.

1.3 Data-driven approaches of entrepreneurial ecosystems

1.3.1 Venture capital networks

The application of complex network theory to entrepreneurial ecosystems is a growing field, due in no small part to the greater availability of relevant data [85]. The choice of articles briefly presented here, while very far from exhaustive, aims to give a sense of the different research directions that have been studied. Network-based analyses of entrepreneurial ecosystems generally deal with three main themes : the impact of the reputation of venture capital firms, prediction of companies' future outcomes using network-based metrics, and robustness of the entrepreneurial network to perturbations.

Syndication networks

Syndication networks are built based on the co-investment interactions during funding rounds. Nodes represent investors, with edges linking two nodes if they took part in the same funding round. These syndication networks represent some of the first and most studied entrepreneurship-related networks as they represent, as [140] put it, a natural starting point for the application of complex networks to venture capital due to several advantages : they are easy to build (the data is public and easy to parse) and play a role in two of the main drivers behind a VC's performance (the deal screening and value-add process selecting the companies and providing support).

In [140], authors study a US-based syndication network of over 3 000 investment firms between 1980 and 1999 using centrality measures borrowed from graph theory, with the hypothesis that better-networked venture capital firms would have higher centrality values. It was found that the more central nodes in the VC syndication network had better performance in terms of Internal Rate of Return (rate of return on an investment) and

in terms of exit rates amongst their portfolio companies. From the company point of view, a well-connected lead VC firm amongst its investors significantly boosted its performance, strongly raising its probability of successfully exiting or raising a subsequent funding round.

In [186], authors study the temporal evolution of the yearly syndication network between VC institutions in the Chinese market. The temporal evolution of the influence of VC institutions in this co-investment graph is evaluated using a k-shell (or k-core) decomposition algorithm that allows for the discovery of higher-order network structures based simply on node degrees. Each node has an associated temporal series of k-shell values computed for each year between 1990 and 2013, where k-shell values are used as an evaluation of the influence of investors in the syndication network. Clustering is then performed on the time series of k-shell values, resulting in 5 distinct groups in terms of financial performance and investment behaviors. Financial investment performance can thus be indirectly estimated simply based on topological features of the syndication network, with the best classification performance (corresponding to a smaller intra-group and larger inter-group distance) achieved using k-shell decomposition rather than other network centrality measurements such as degree centrality, betweenness centrality, h-index or eigenvector centrality.

Company success

Recent works build startup-related networks allowing for the estimation of factors linked to company success. Metrics are then computed on these network and are related to successful outcomes. A prediction criterion can then be defined, which is used to compare predictions to null models and real venture capital firms' performance.

In [99], the characterization of the performance of VC-backed firms is approached through the study of a bipartite network linking companies in the health sector and investors. Authors restrict their analysis to the healthcare sector to homogeneize their sample (sector-specific effects have been observed on investment patterns) and to reduce the impact of market oscillations on the observed outcomes. Here, a successful company is defined following [36] but here excluding acquisitions performed by the company and adding mergers (a successful company is defined as either having been acquired, gone public or merged with another firm). The bipartite network is also a multigraph since investors can invest several times in a single company, and accounts for the temporal dimension with links persisting through time (*i.e.* the graph representation for year y_f contains all links from years $y \leq y_f$). This graph is projected on the firm (two firms are linked if they receive an investment from the same investor in a 7-year timespan) and investor (two investors are linked if they have co-invested in a firm) layers, yielding two representations of this bipartite graph used to compute network metrics of interest such as average neighbor degree, clustering coefficient and a myriad of centralities. Funding trajectories (*i.e.* cumulative funding amounts raised by a company over time) are then computed and clustered into high and low investment regimes, and the links between values of the network metrics and allocation of a company to the high or low investment regime are investigated. Firms' closeness and PageRank centralities are found to be positively significantly linked to the probability of the com-

pany belonging to the high-regime class. The agreement between standard (IPO, acquired or merged) and trajectory-based (high- or low-regime) success was also found to be high, with an accuracy of 0.71 (correctly predicted samples over total number of samples), a precision of 0.57 (true positives over total number of positive predicted instances) and a recall of 0.31 (true positives over total number of positive instances). As different centrality measures capture different aspects of the social relationships in the temporal network, their varying importance gives insights into the important factors at different points of a venture's life cycle : the results suggest that being connected to important and well-connected investors is important in the early stages of the company (positive effects of the PageRank centrality and investors' average neighbor degrees increase in the early life of a firm), whereas being in a far-reaching portfolio of investors (small clustering coefficient and high eigenvector centrality) has a stronger impact in later stages.

In [36], the time-varying network of information and knowledge flow is approximated using the worldwide network of human capital moving between companies. The assumption is made that, as employees move from one company to another, the new company gains access to the knowledge (both technical and business-related) developed in the employee's old firms, giving rise to a network approximating information flow between companies and increasing the likelihood of success of companies central in this network. Startup outcomes were then measured to link company success (as defined by the company either acquiring another company, being acquired or undergoing an IPO) with network-related characteristics of the companies. Using network centrality measures at an early stage to assess the likelihood of long-term success, this knowledge transfer network was found to hold strong predictive power for companies in the pool of *open deals* (firms which have not yet received funding, been acquired or gone public). This work thus underlines the links between centrality in the professional network and long term economic success of companies in a knowledge-intensive industry.

Robustness

Finally, some applications of network analysis to entrepreneurial ecosystems try to assess the robustness of these networks through the impact of extinctions of certain node classes or the study of shock propagation through the network.

In [104, 103], the specific contributions of VC firms to the robustness and innovative capabilities of the Silicon Valley ecosystem are studied. It is argued that the presence of VC firms in the cluster of innovation facilitates specific interactions between companies and other members of the innovative cluster, such as universities, large companies and laboratories. Innovative clusters are characterized by their capabilities to generate breakthrough innovations, creating new industrial domains and nurturing startups that develop disruptive technologies rather than incrementally improving on existing industrial sectors. By framing the innovativeness of the Silicon Valley as an economic phenomenon embedded in a complex network, complex network theory methods are used in order to analyze the interactions of the numerous economic agents (such as institutions, companies, investors) that give rise to innovation and entrepreneurship. This analysis demonstrates the various roles of VC firms beyond simply funding startups : their selection of the most promising

projects of the region acts as a signal to the business community, they accumulate and help spread entrepreneurial knowledge in the innovative cluster and embed the interdependent agents of the network (*i.e.* they link economic and non-economic institutions). The removal of VC nodes from the network would also weaken the system as a whole due to the specificity of the competencies they hold, as the entire system is weakened if one type of agent is not present.

In [332], the channels propagating risk and loss in venture capital markets were studied using a multilayer network analysis in order to better understand the contagion risk mechanism through which negative effects (such as the collapse of the *dot-com bubble* of the early 2000s) propagate from the failure of specific market participants into the rest of the system. A multilayer network was built with VC firms embedded in one layer, startups in another, internal links between VCs in the VC layer (common capital provider, usually limited partners) and startups in the startup layer (business reliance such as collaboration between small companies) and external links between VCs in one layer and startups in the other (an investor invests in a startup). All links are undirected and unweighted. Each node also has a weight representing its cash position (amount of money owned by the agent at a given moment in time). Once the cash position of a node reaches below a certain threshold, this node is deactivated. Two channels of risk contagion are then defined : direct liquidity shocks propagated via external links (a failed VC node, for instance, transmits a shock to its portfolio companies which reduces the cash positions of all involved actors) and indirect risk contagion propagated via internal links where failure of venture capital firms in a limited partner's portfolio lead to capital supply shrinkage, which can in turn lead to venture capital firm failure, and startup failure who will impact their connected business partners. This internal failure is driven by a resilience parameter where an agent fails if a sufficient number of its graph neighbors fail. Simulations are then run, with one agent is selected as the initial failure node and the evolution of the graph is computed following the two risk contagion dynamics which directly influence the cash positions of the nodes (and thus their failure). The multilayer graph is built based on venture capital investment data obtained from Bureau van Dijk covering the complete year 2017, and the initial failure corresponds to the startup or venture capital firm with the highest degree of external connections. When losses propagate only via external links, the system was found to remain robust to perturbation. When both direct and indirect risk contagion channels were taken into account, however, the whole market was found to display an abrupt transition between a stable and unstable state. Market robustness was also found to be linked both with network connectivity (more connected networks were more robust), and to the initial distributions of cash positions (heterogeneous distributions were found to be more susceptible to global collapse).

1.3.2 Machine learning-based prediction models

The recent availability of large datasets and algorithmic developments have also attracted the attention of other communities, in particular with regards to the task of automatic data-based prediction of successful startups. This spike in interest is in no small part driven by the venture capital firms themselves in order to help solve a major problem : a surge in

potential investment opportunities that they are not equipped to properly deal with. Indeed, the venture capital business has limited scaling capabilities [67], with large fixed costs for evaluating candidate ventures [118] due to the manual and empirical nature of the traditional deal screening process. Furthermore, human investors are unable to consistently make the right choices based on intuition alone due to their inherent biases [82]. Both academia [37] and venture capital professionals point to this surge of investment opportunities as a significant reason behind the adoption of data-driven technologies in deal screening. This subfield of entrepreneurship-related research is not directly related to the core topic of this thesis, but we will briefly present its general concepts given its current relevance and methodological proximity.

General concept

Following [58], we will give a brief rundown of the steps followed when building a machine learning based pipeline for startup prediction from the point of view of a venture capital firm. The goal is to find common patterns in inputted startup-related data and approximate a function(*training*) linking these patterns to the defined outcomes of the companies. After training is done and given similar data on a new startup, the model estimates the probability of success of the new venture (*prediction*).

- definition of the prediction problem : due to constraints on the type of deals a fund can participate in, the prediction problem is more complex than simply finding successful companies using machine learning models. Indeed, compatibility between the company and the investment thesis of the fund (such as geographical location, sector focus, investment mandate and exit opportunities) needs to be taken into account when assessing ventures. Compatibility can be accounted for in two main ways in a machine learning pipeline. One can either have the model predict, given a single venture, both its success probability and compatibility with the investor's thesis or first filter out all companies that do not match the investment thesis and then predict success probabilities on the remaining sample. The second method is preferred as it simplifies the model, increases flexibility (no re-training is required if the investment thesis changes), and is more intuitive.
- definition of the success criteria/criterion : as discussed in the previous subsection, the question of what constitutes success for a company is open. It is, however, necessary to have a formal and measurable definition of success in order to annotate startup outcomes to allow for model training and prediction, with the definition of success depending on the activity of the investor (late-stage funds can be more interested in success being defined as the company going public or getting acquired, while early-stage investors can be more interested in defining success as the company managing to raise subsequent funding rounds).
- gathering the data : model performance is intrinsically linked to the quality of the data it is being fed to describe companies. In gathering data, it is thus of value to obtain data approximating information obtained from both modes. Due to constraints on data obtainability, modality or size, choices must be made : multimodal data (*i.e.* data

that comes in a mix of different types such as text, video, audio or numerical), for instance, can have stronger predictive power but is usually harder to obtain and to parse. This is particularly relevant as human analysis is theorized to be the result of 2 parallel modes of information processing : the*intuitive* and *analytical* modes [98], where soft signals (such as founder personality or ability to convince an audience) are processed through the intuitive mode and hard facts (such as financial data) through the analytical mode. During deal screening, a VC firm evaluates candidate ventures through both modes, with intuitive features usually being much harder to measure. Some data (often related to the approximation of intuitive features) can prove to be disproportionately hard to acquire or manipulate compared to the marginal predictive performance gain of the model.

- preprocessing the data : in order to have a balanced model that is less susceptible to biases in the input data (such as overrepresentation of specific classes or differences in scale between the various features), feature preprocessing is required. Furthermore, as most machine learning models only accept numerical features as inputs, modal data that is not in a numeric format needs to be transformed into a numerical representation (using for instance embedding techniques, encoding or statistical descriptions).

- splitting the data : it is necessary to reserve part of the data for model evaluation, ensuring that the model correctly performs on "unknown" data. This data splitting step can vary in complexity, with the simplest way being simply holding out a random percentage of the total dataset (usually between 20% and 30% of the dataset), training the model on the remainder of the data and evaluating the performance of the trained model on the held-out data.

- choosing model architecture and training parameters : choosing the actual model can be a tricky affair, in particular for deep-learning models where the architecture can quickly become very complex. As the "No free lunch" theorem [325] famously states, there is no one-size-fits-all solution to this question, and choosing the correct model will depend on each use case, such as the prediction problem and modality of the data. Evaluating multiple appropriate models and choosing the best performer *a posteriori* is the approach usually followed, with a bonus given to simplicity and explainability.

- evaluating the performance of the model after training : using evaluation metrics on the sample held-out before training, it is possible to get a general appreciation of the model performance by comparing model predictions (model-predicted outcomes) to the ground-truth labels (actual outcomes). Several evaluation metrics are usually used in tandem, with their selection driven by what constitutes a good model for the use case. This model evaluation will help determine which model architecture and training parameters should be favored, and ascertain the capability of the model to generalize (*i.e.* correctly make predictions on situations not present in the training data).

- evaluating model predictions : the high complexity (in terms of parameters) of machine learning models can lead to highly nonlinear relationships between input features and output predictions, which makes understanding why the outputs are as

they are very difficult. As these models are used as tools to aid decision making during deal screening processes, the ability to explain the factors behind the final prediction is crucial for investment professionals. The field of explaining the inference processes and final results of deep learning models is called explainable AI [327]. Understanding the relationships between input features and final outcomes is important for several reasons. First, it increases trust in the models by allowing users to judge if the features they consider important are reflected in the model. Second, it enables hypotheses-mining, which consists in seeing how feature values relate to the predicted outcomes without requiring *a priori* hypotheses. Third, it helps with model troubleshooting by allowing users to analyze the relationships between features and predictions. Two separate levels of explainability are considered relevant : the global-level explainability, where the importance of individual features is related to the predicted outcomes across all observations (*i.e.* high values for feature **A** are generally linked to positive outcomes) and the instance-level explainability that provides fine-grained interpretability where, given a single observation, the impact of each feature is shown on the predicted outcome (*i.e.* for a given successful outcome, the main drivers behind the predicted value are features **A** and **C**).

- deploying the model : once a model displays satisfying performance, it has to be deployed into production to screen real inbound deals. As humans and machine learning models perform well in different situations (with humans being better at evaluating outliers and machine learning models being faster, and less susceptible to biases). Hybrid decision making models (human-in-the-loop) where the model prediction and its drivers are treated as additional information about the venture which will then be assessed by a human are usually preferred. Human oversight also allows for performance monitoring of the level on several fronts : first, to assess that model performance is in line with the evaluation and testing stage and second, to guard against degradation of the model over time (due to, for instance, data drift or changes in user context due to the evolution of the fund's investment thesis over time).

Potential pitfalls

It has been argued that the very idea of venture capital is at odds with data-driver deal selection [37]. Indeed, on the one hand venture capital investments tend to finance novel ideas that rarely succeed but achieve major success when they do [164] while, on the other hand, algorithms rely on massive datasets to learn patterns from historic data. Furthermore, algorithms, while not biased as humans can be in their decision making, are prone to biases in the data : female founders have, for instance, been found to raise less funds when launching companies in male-dominated industries [158]. These biases could be learned and reproduced by an algorithm taking as inputs, amongst a number of others, features pertaining to the industry of the company and the founders' gender. The risk of data-driven investing, then, is two-fold : novel ideas could find themselves unable to attract financing, and past biases could find themselves perpetuated by virtue of their presence in the dataset.

The impact of a VC firm becoming "data-driven" (defined as hiring their first data-related employee) was studied in [37] using a worldwide investment database sourced from

Crunchbase. This analysis is based on the backward-similarity of a startup computed as the textual similarity between the company and all previous VC-funded companies operating in the same industry (a high backward-similarity corresponds to a startup whose business is similar to many others). The main findings of this article are that data-driven VCs, compared to regular VC firms, deploy more capital towards backward-similar startups, significantly increase their number of investments and assets under management, are significantly more likely to select startups that survive and receive follow-on investments and exhibit a comparative advantage at screening backward-similar startups but show no significant difference for other companies. Data-driven VCs, however, become significantly less likely to invest in startups that achieve scarce major success (such as an IPO or a profitable acquisition), and invest in less innovative startups as measured by the number of future patents filed and citations obtained. The performance of data-driven VCs, however, was found to remain similar, with data technologies potentially enabling VC firms to increase their assets under management without harming their performance.

General conclusions

Venture capital has been considered a field of study in its own right since the end of the 20^{th} century at the very least [65, 265], gaining traction ever since [278]. A number of research questions such as identifying the drivers behind successful entrepreneurship or understanding investment behaviors of venture capital firms have historically been approached through qualitative, small-scale analysis using surveys or manually collected datasets. These studies have yielded major insights into a myriad of facets of entrepreneurship, but present limitations and biases (such as western overrepresentation) due to the relatively small sample sizes and to the nature and identity of both survey designers and respondents. Data-based studies have in recent years gained steam due notably to the growing availability of large-scale, worldwide databases, quantitatively estimating both "soft" and "hard" aspects of entrepreneurship and linking them with entrepreneurial outcomes. Even though the advances in this direction have been numerous, the complex nature of entrepreneurial ecosystems has required simplifications in how interactions are modeled, discarding information in the process.

Indeed, even though network studies of venture capital have been a topic of interest for over 2 decades, coinciding with the general rise in available data and interdisciplinary applications of network science, most quantitative analyses tend to ignore the bipartite nature of interactions between venture capital and startups by either studying unipartite graphs (such as syndication networks [140]) or collapsing the bipartite investor-startup graph into separate unipartite graphs [99]. Studying this bipartite graph as is proves challenging due to its size and sparsity, with tens or even hundreds of thousands of highly heterogeneous nodes in each guild interacting with only –compared to the total size of the graph– few other nodes. Furthermore, drawing parallels with ecological sciences, networks can be studied at different levels of organization, such as the individual level, the species level or at broader spatiotemporal scales. Each of those scales yields different information about the characteristics of the item of study, with individual-based networks allowing for the study of variations in niches among individuals, and species-based networks allowing for

the description of the architecture of the ecological communities [129]. As investors are empirically known to belong to different types, building *investor species* from the *individual investors* to better understand the architecture of our economic communities is particularly relevant in the context of this thesis. This in turn allows for the study of *species*-level networks *i.e.* coarser-grained representations where investors are aggregated into homogeneous communities.

Contribution

Chapter 2

Investor clustering

This chapter is based on Carniel, T., Halloy, J., & Dalle, J. M., 2023 : *A novel clustering approach to bipartite investor-startup networks* (Plos one, 18(1), e0279780).

We propose a novel similarity-based clustering approach to venture capital investors that takes as input the bipartite graph of funding interactions between investors and startups and returns clusterings of investors built upon 5 characteristic dimensions. We first validate that investors are clustered in a meaningful manner and present methods of visualizing cluster characteristics. We further analyze the temporal dynamics at the cluster level and observe a meaningful second-order evolution of the sectoral investment trends. Finally, and surprisingly, we report that clusters appear stable even when running the clustering algorithm with all but one of the 5 characteristic dimensions, for instance observing geography-focused clusters without taking into account the geographical dimension or sector-focused clusters without taking into account the sectoral dimension, suggesting the presence of significant underlying complex investment patterns.

2.1 Introduction

Within the active field of entrepreneurship research [65], quantitative analyses of the structural properties of investor-startup interactions have been conducted so far on a simplified version of the investor-startup network, namely, on the network of investor-investor relationships, through the construction of syndication networks where two investors are linked if they either invested jointly in a startup or have a common startup in their portfolios [128, 275, 140].

These limitations are typically manifest when trying to address and account for the important and structural heterogeneity between investors: startup investors have marked differences, with respect to sectoral specialization, to the average amounts invested (from hundreds of thousands of dollars to hundreds of millions), or else to their geographical focus, to name but a few relevant dimensions. Ignoring this heterogeneity or failing to address it appropriately results in biased, if not misleading, conclusions, and certainly makes

the observation and characterization of larger-scale collective phenomena with respect to entrepreneurial ecosystems and of their temporal dynamics an impossible task. Community detection algorithms [109, 253] have been applied to traditional syndication networks but have either failed to incorporate explicit information about investment stages [155], which typically results in overestimating actors who invest early in startups and are therefore linked to numerous subsequent investors according to syndication links, or have relied on a semi-supervised approach [326] that relies on ex ante and partly subjective and/or largely unavailable segmentation of investors, or else have been structurally limited by the definition of the networks studied: [49], using a modularity-based community detection algorithm, identifies communities of investors based on their interactions, but cannot do so based on their similarity and therefore are unable to address the heterogeneity of structural investors. Syndication networks, as one-mode projections, cannot capture the complex and multi-layered interactions characteristic of bipartite venture networks, and therefore relevant aspects of entrepreneurial ecosystems are lost.

More recent methods such as multi-view data clustering [314, 185, 328] are promising, but are not able to deal with our specific constraints : our data is fundamentally bipartite, with each of the views containing different types of data (numerical vs. categorical vs. logarithmic) that are either node-based or edge-based. Specific clustering algorithms incorporating domain-specific knowledge to cluster similar investors through their position and representation along the various axes of the complex bipartite multilayer multigraph are thus necessary in order to study investment dynamics in the investor-startup network.

New analytical tools are required to take advantage of the distinctive structure of these networks and to extract more information, associated with more complete datasets that would allow to build both sides of the bipartite networks and the interactions between them. Fortunately, the use of databases giving both large-scope and in-depth data on investor and startup companies and on their interactions is now rapidly becoming standard [85] while, following notably the ecological literature, methods for bipartite graph analysis have recently become more and more developed and accessible [77]. In this context where both tools and materials have become available, we initiate in this chapter an enriched analysis of interactions in entrepreneurial networks and ecosystems, with a direct look at the funding events rather than at the syndication shadow they project.

2.2 Objectives

We propose a novel, unsupervised investor clustering approach for entrepreneurial investors that mitigates some of the difficulties described earlier. It was developed both as a direct tool to probe and characterize the typology of actors in venture capital ecosystems and as a methodological building block with respect to the quantitative analysis of the dynamics of entrepreneurial ecosystems. Our method is based on an unsupervised community detection algorithm using a Hellinger-based similarity measure, computed over all pairs of investors, and accounting for 5 well-defined characteristic dimensions to describe investors. As a consequence, the similarity between investors is easily quantifiable and interpretable, compared to traditional clustering method based on machine learning techniques - and although significant progress has been made in terms of interpretability [220].

The similarity graph pruning threshold is the only parameter, and the number of outputted classes is freely determined by the clustering algorithm and is not constrained. As it happens, this method also allows for a controlled modification of the clustering parameters and features, which results in the identification of unexpected community-level patterns that help better understand the dynamics of the different classes of investors.

2.3 Materials and methods

2.3.1 Dataset

The dataset used for this study was extracted through the Crunchbase API on October 7$^{\text{th}}$, 2020. It contains information on $1\ 156\ 085$ startups (name, creation date, headquarter location, sectors of activity), $348\ 020$ funding events (target startup, date, investors involved, amount, investment stage), $159\ 585$ investors (name, creation date, investor type, investor location) and $1\ 067\ 089$ individuals (name, past and current professional experiences, level and sectors of education, company board memberships and advisory roles). We removed the *Software* sector from all startups' sectors of activity as this tag is overly represented (occurs in roughly 25% of startups, almost twice as frequent as the second most frequent tag) and is relatively non-descriptive.

2.3.2 Investor-startup network

We create a temporal bipartite multigraph where top nodes are the investors, bottom nodes are the startups and edges correspond to funding events between the investor and the startup (see Fig. 2.1 for a schematic representation of the graph). As an investor can fund a startup at several points in time, two nodes can be linked through several temporal edges. We removed nodes for which the geographical information was not available and edges where the financing event was not an investment event (grants, debt financing, etc.), and afterwards removed isolated nodes as they do not take part in the network interactions studied. This process resulted in a network with $65\ 653$ top nodes, $95\ 329$ bottom nodes and $392\ 204$ edges linking these two sets.

2.3.3 Hellinger distance and investor similarity

The Hellinger distance h [90] and the associated similarity θ between two normalized discrete probability distributions P and Q are defined as :

$$h(P,Q) = \frac{1}{\sqrt{2}}\left\|\sqrt{P} - \sqrt{Q}\right\|_2 \tag{2.1}$$

$$\theta(P,Q) = 1 - h(P,Q) \tag{2.2}$$

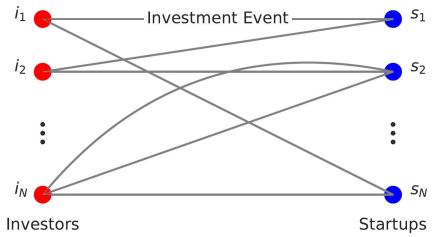

Figure 2.1: **Schematic representation of the investor-startup multigraph.** The red nodes on the left represent investor nodes, the blue nodes on the right represent startup nodes. The edges between investor node i and startup node s represent a funding interaction where investor i invested in startup s at a given time. As an investor can invest in a startup several times, multiple edges can connect two given nodes as shown on the figure.

where $||.||_2$ is the Euclidean (or L2) norm [170] and $\sqrt{P}$ is the vector with elements the square root of the elements of P. By definition, $0 \leq h(P,Q) \leq 1$ and thus $0 \leq \theta(P,Q) \leq 1$ with $\theta = 0$ corresponding to minimal similarity (maximal distance) and $\theta = 1$ to maximal similarity (minimal distance) between two distributions. The Hellinger distance is used as the probability distributions are low-dimensional and it has been shown to be more suitable than Minkowski distances for probability vector comparisons [179, 334, 273].

The similarity Θ between two investors $\vec{i_a}$ and $\vec{i_b}$ is then defined as follows :

$$\Theta(\vec{i_a}, \vec{i_b}) = \left| \prod_{k=1}^{k=n} \theta(i_a^k, i_b^k) \right|^{1/n} \tag{2.3}$$

where i_a^k is the distribution characterizing investor a along the k-th dimension and n the total number of dimensions characterizing an investor.

2.3.4 Investor characterization

We characterize investors along $n = 5$ dimensions related to their investments in startups, each of which being associated with a frequency distribution, chosen in order to collectively exhaustively describe investment portfolios and therefore to allow to accurately characterize investors. Within the bipartite graph, these dimensions depend both on **edges** linking an investor to startups (for instance the date of the investment, as several different temporal edges can link an investor and a startup) or on the **startup nodes** (e.g. the geographical location of an investment made by investor ***i*** targeting startup ***s*** will be the geographical location of startup ***s***). These characteristic dimensions can be measured for all investors, are public enough so that the information is available for most transactions and are linked to common descriptors used by practitioners of the domain to characterize investors (for instance *early-stage* vs. *late-stage* [97], *domestic* vs. *international* [88], *specialized* vs. *generalist* [141], *historical* vs. *emergent* [93]).

- **Temporal investment distribution :** the frequency of investments per year of the investor (Fig. 2.2). This is an edge attribute.

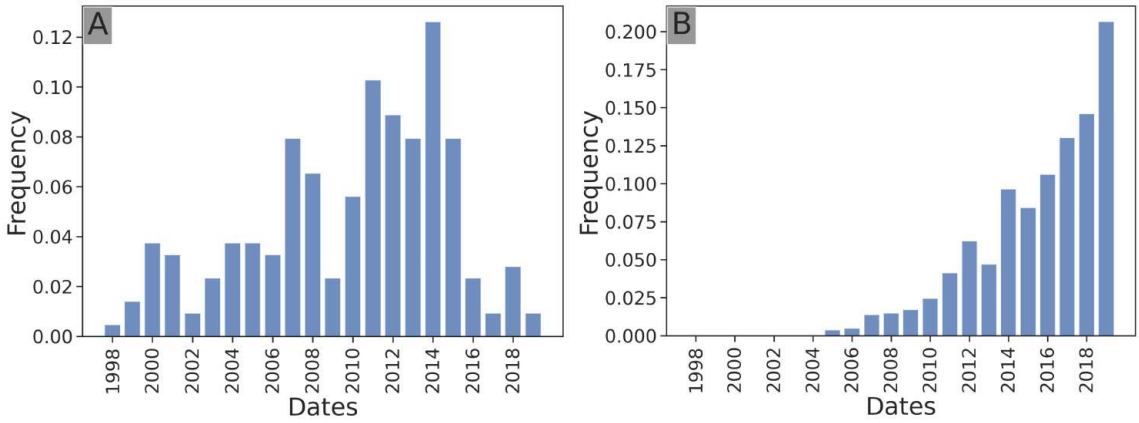

Figure 2.2: **Temporal investment distribution.** Temporal investment distribution of *Softbank Capital* (**A**), a telecom-focused US-based venture capitalist that stopped its activity in 2017, and of *Y Combinator* (**B**), a US-based startup accelerator founded in 2005. The two temporal patterns of actvitity are quite different between the two structures, as Softbank Capital stops investing near the end of the period whereas Y Combinator's activity steadily grows throughout the whole period.

- **Geographical investment distribution :** the frequency of investments of the investor in each country (an investor invests in a country if the target startup's headquarters are located in the country) (Fig. 2.3). This is a startup node attribute.

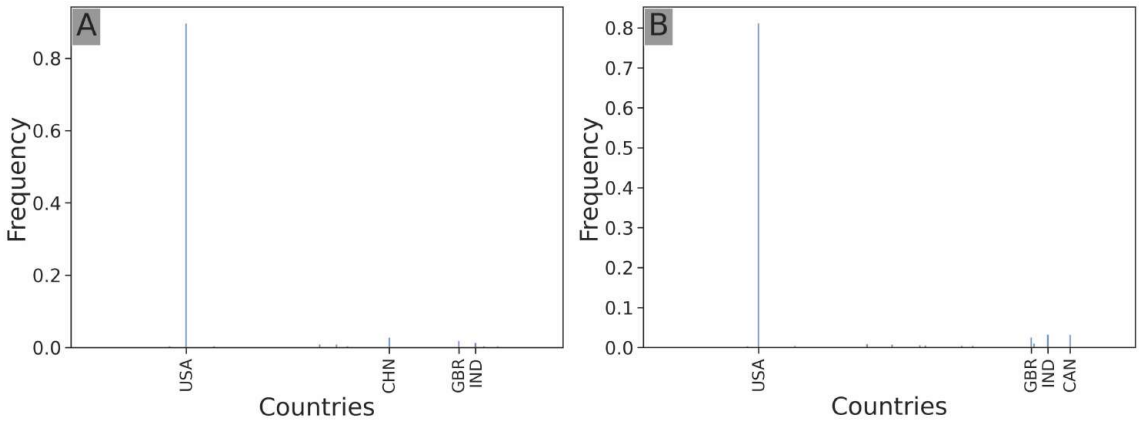

Figure 2.3: **Geographical investment distribution.** Geographical investment distribution of *Softbank Capital* (**A**), and *Y Combinator* (**B**). Only the top 4 target countries in terms of frequency of investment are labeled. Both structures heavily target US-based ventures.

- **Sectoral investment distribution :** the frequency of investments of the investor in each sector of activity (an investor counts as investing in a sector if the target startup of the investment is labeled in this sector) (Fig. 2.4). This is a startup node attribute.

- **Stage investment distribution :** the frequency of investments of the investor in each stage of the venture capital cycle (Fig. 2.5). This is an edge attribute.

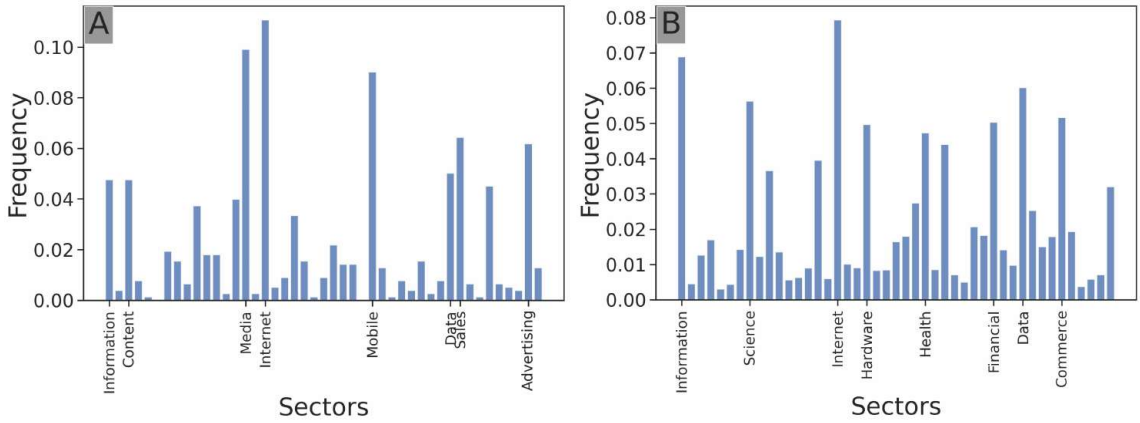

Figure 2.4: **Sectoral investment distribution.** Sectoral investment distribution of *Softbank Capital* (**A**) and *Y Combinator* (**B**). Only the top 8 sectors of investment are labeled. Softbank Capital shows a strong focus on IT-related ventures whereas Y Combinator shows a wider sectoral breadth.

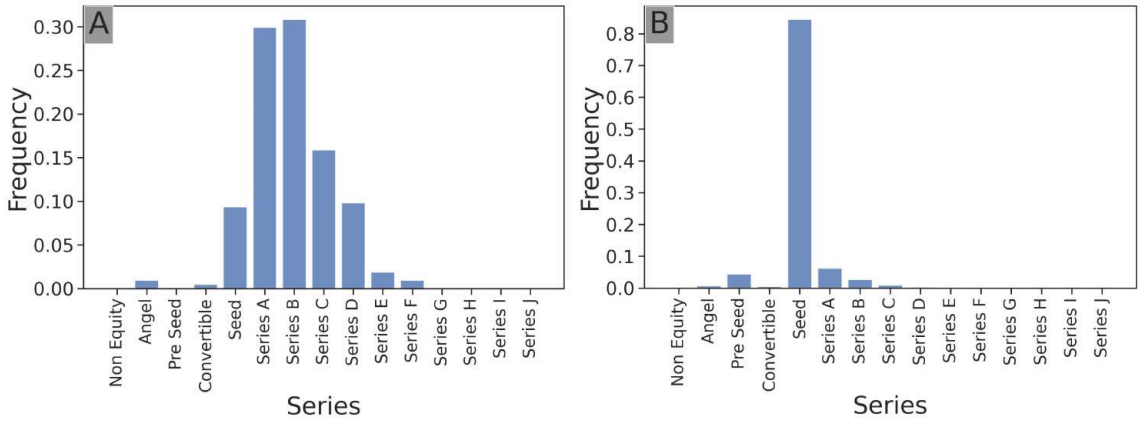

Figure 2.5: **Stage investment distribution.** Stage investment distribution of *Softbank Capital* (**A**) and *Y Combinator* (**B**). Softbank Capital shows a strong focus in late-stage investment (most of its investments are in Series B or later) whereas Y Combinator shows a very strong early-stage specialization (over 80% of its investments in Seed stage).

- **Amount investment distribution :** log-binned distribution of the funding amounts of all investments of the investor in USD (Fig. 2.6). Logarithmic binning was used because the amounts of start-up financing rounds follow a power-law type distribution [81]. This is an edge attribute.

2.3.5 Self-difference index

For each community g and each year t in the period of study, the set of the top p sectors $k_t^g = \{m_1, m_2, ..., m_p\}$ in terms of number of investment is computed. The self-difference

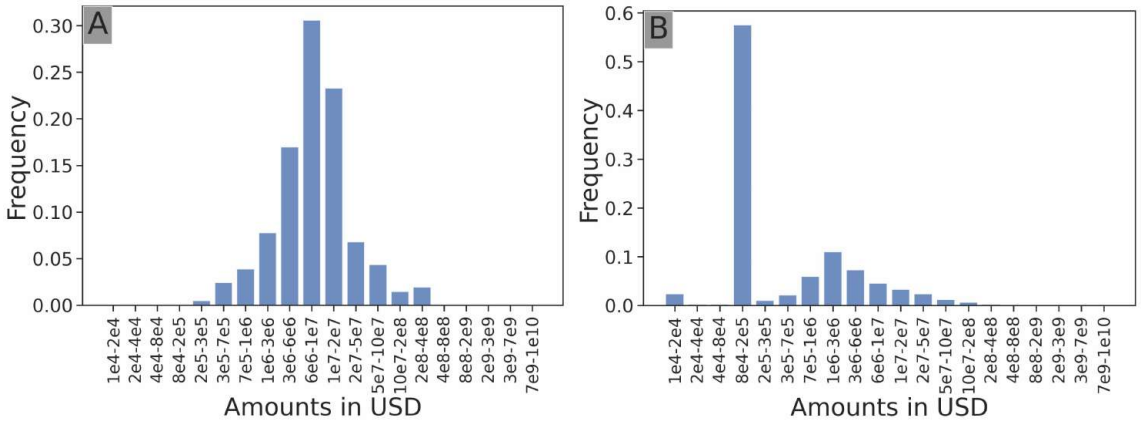

Figure 2.6: **Amount investment distribution.** Amount investment distribution of *Softbank Capital* (**A**) and *Y Combinator* (**B**). In line with Fig. 2.5, we see that Softbank Capital invests relatively high amounts (peak frequency of investment between 6 million USD and 10 million USD) whereas Y Combinator invests smaller amounts in a very systematic manner (peak frequency of investment between 80 000 USD and 200 000 USD). This is in line with the accelerator model where accelerators invest a set amount in all ventures they decide to support. Furthermore, Y Combinator has also developed funds such as Y Combinator Continuity dedicated to investing in its alumni companies after their initial investment. This can be seen in the small bump in the funding amount distribution between 700 000 USD and 10 million USD.

index $d \in [0, 1]$ between years t_1 and t_2 for community g is defined as follows :

$$d(k^g_{t_1}, k^g_{t_2}) = \frac{k^g_{t_1} \Delta k^g_{t_2}}{2 \min(P - p, p)} \tag{2.4}$$

where Δ is the symmetric difference between both sets and P is the total number of sectors. This self-difference index ranges from 0 (identical sets) to 1 (no overlap between the top p sectors of investment at year t_1 and the top p sectors of investment at year t_2). As there is a natural inflation in terms of number of investment rounds due to an increase in venture capital activity during the latter part of the period of study, the index takes into account the ordering of the sectors in terms of number of investments rather than the raw number of investments.

2.4 Results

2.4.1 Investor Communities

Clustering

We reduce the set of top nodes (investors) worldwide to top nodes with degree $d \geq 60$ investments throughout the 1998-2019 period (a low number for a professional investor over this time frame) to ensure a sufficient number of observations for each dimension characterizing an investor. Note that the same clustering results hold for a graph reduced to investors with $d \geq 100$ or more investments. This procedure results in 1014 investor nodes in the final graph with $159\ 353$ edges connecting them to startup nodes, isolate nodes being removed (see previous section). We compute the pairwise similarity Θ as defined in Eq. 2.3 between all investors in our sample and then define a complete weighted similarity graph with investors as nodes and the similarity between two investors as edge weights. We prune the graph by retaining for each investor the 1% edges with the highest similarity, yielding a k-nearest neighbor graph. Indeed, the k-nearest neighbor graph presents several interesting properties for clustering applications : the resulting adjacency matrix is sparse, it can connect nodes on different scales on the graph, and is generally less vulnerable to unsuitable parameter choices [312]. In our case, since all investors are linked to all other investors, this pruning procedure reduces the possible fluctuations of the community detection due to weak links, thus strengthening the community information present in the similarity graph [166, 16]. We then run the *best_partition* community detection algorithm from the Python *community* package [34] resulting in an investor clustering with 11 different communities.

For each of the communities, a theoretical *representative investor* defined as the barycenter of the communities' investors in the 5-dimensional probability space is computed: in each dimension, the distribution of the representative investor of a given community is the average of the distributions of all investors in the community. This representative investor allows for a compact visualization and understanding of each community, yielding some relevant understanding as to how the communities are formed. Fig. 2.7 for instance shows the representative investor for community **A6** and shows that investors in community **A6** have an obvious China-focused geographical bias since over 84% of the cluster's investments target China-based startups. As another example, Fig. 2.22 in the supplementary material shows a similar sectoral focus on Health Care-related investments in community **A7**, with around 27%, 30% and 26% of investments in *Science and Engineering*, *Health Care* and *Biotechnology* respectively.

Fig. 2.8 shows the similarity graph pruned as described previously without (left) and with (right) the results of the clustering superimposed on the individual nodes. In light of these observations, we further characterize each of the resulting communities as described in column **A** of Table 2.1 by analyzing the representative investors of each of the 11 communities, which can be found in the supplementary material (Figures 2.15 to 2.25)), and referring also to the identity of individual investors in the clusters (see Table 2.2 for a sam-

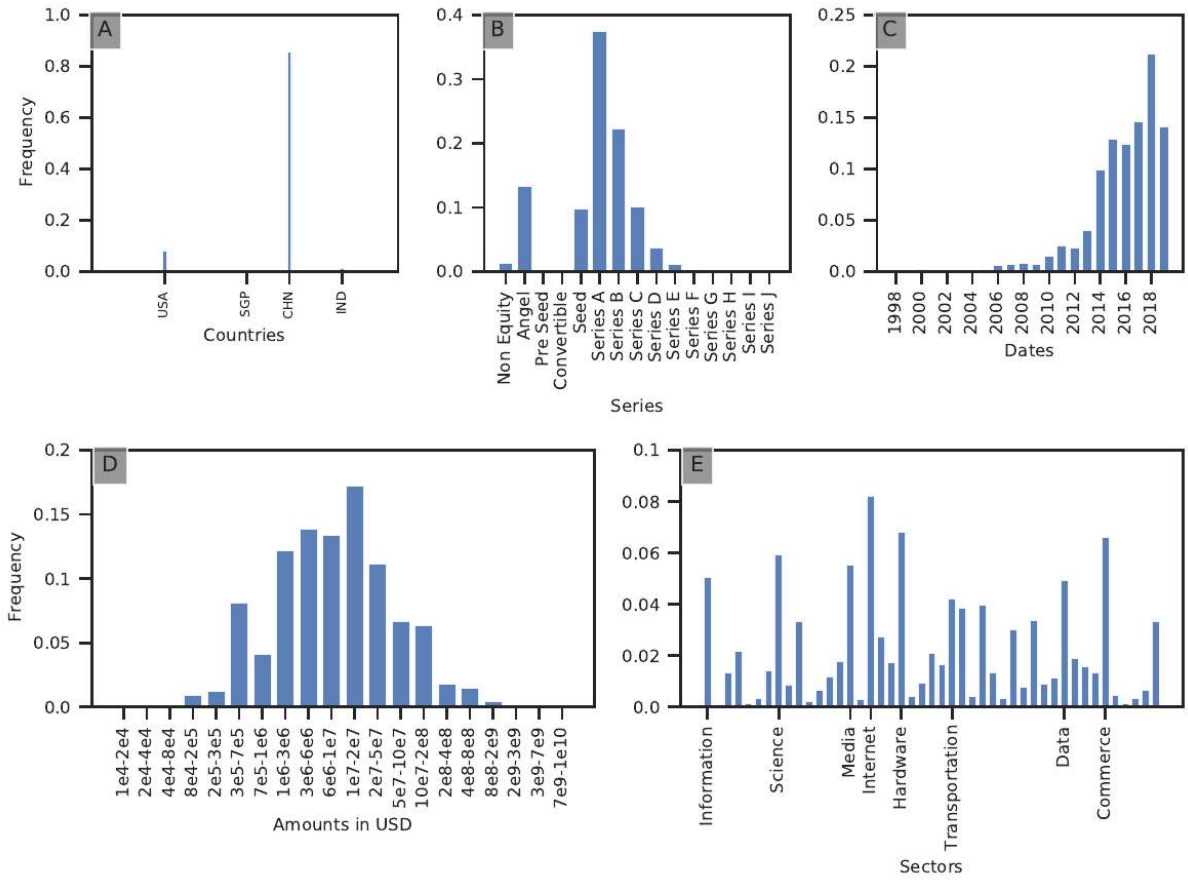

Figure 2.7: **Representative investor of community A6.** Community **A6** appears comprised of investors targeting China-based ventures during the second half of the 2010s with no clear sectoral specialization. Panel **A** shows the representative geographical investment distribution of community **A6**, panel **B** the distribution of the series of investment, panel **C** the temporal distribution of investments, panel **D** the distribution of the amounts of investment and panel **E** shows the sectoral distribution of investment.

ple of individuals from each cluster). We observe that each community corresponds to a strong and specific pattern: a specific geographical area of investment, a specific sector of investment, investing at specific startup development stages, or displaying a specific temporal pattern notably in relation to the 2008 financial crisis i.e. grouping investors that were either active throughout the whole period, or that belonged to older or newer generations of investors typically active either before or after the 2008 crisis.

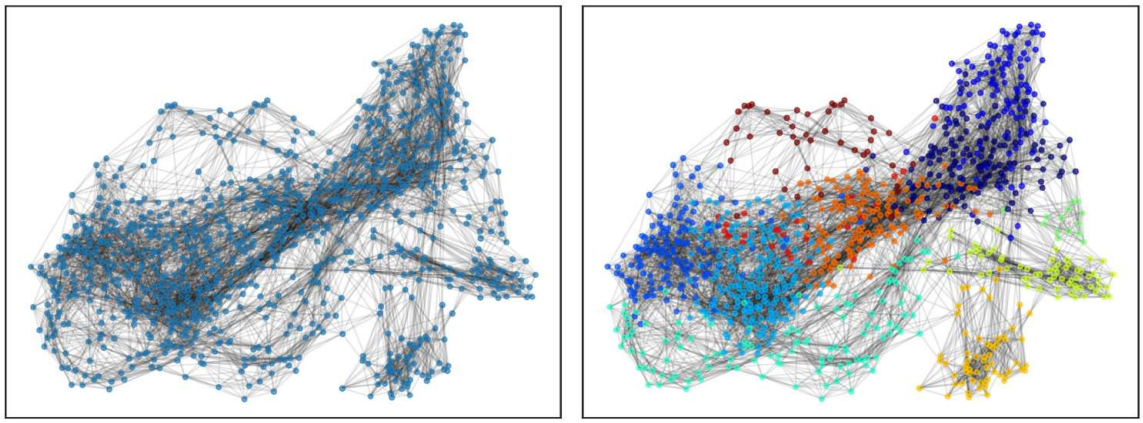

Figure 2.8: **Similarity graph and community assignment.** Pruned similarity graph without (left) and with (right) community assignment of the nodes as characterized in column **A** of Table 2.1. The neon yellow community corresponds to China-focused venture capital firms (**A6**), the dark red community to India and Japan-focused venture capital firms(**A10**), the gold community to Health Care specialists (**A7**), the blue community (far left) to accelerators (**A2**).

Temporal evolution patterns

Based on this investor clustering, figures 2.9 and 2.10 reveal the temporal evolution of two communities in terms of target sectors of investment over the 2010-2019 period. Community **A0**, composed of general investors active over the whole period studied, typically shows a relatively slow evolution in terms of sectoral trends, with a gradual shift (Fig. 2.9) in preferred sectors of investment towards so-called *deeptech* sectors (shift from sectors such as *Media and Entertainement*, *Mobile* towards sectors such as *Science and Engineering*, *Health Care*). Community **A7**, composed of health-care focused investors, shows a very strong dominance of Health Care-related sectors throughout the whole period (Fig. 2.10, **A**), but where the top 10 sectors have significantly evolved over the 10-year period of study (Fig. 2.10, **B**). A closer look at the non-health related sectors reveals a clear shift from *Manufacturing* and *Hardware*-related investments towards *Data Science and Analytics* and *Artificial Intelligence*-related investments, in line with the widespread adoption of these technologies in Health Care-related sectors during recent years [317].

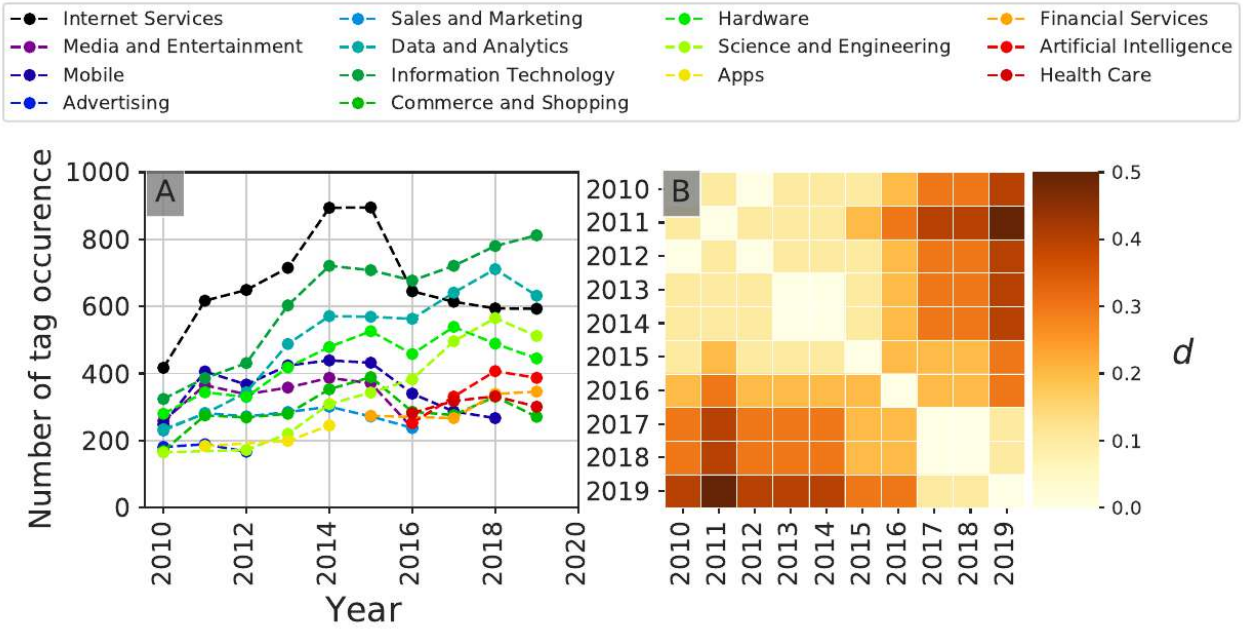

Figure 2.9: **Temporal evolution of the investment patterns of community A0.** Temporal community investment patterns of the target startups' sectoral tags for each year aggregated at the community level. Community **A0** is comprised of large, historical, rather late-stage focused venture capital firms. Panel **A** shows for each year the ten tags that received the most investments, panel **B** shows the community self-difference index described in Eq. 2.4. We see a gradual but consequent shift in the target industries of community **A0** throughout the period of study as evidenced in panel **B**, notably with the disappearance of relatively low-tech sectors such as the *Mobile*, *Apps* and *Advertising* sectors.

Table 2.1: **Descriptive table of the communities for the different clusterings.** Each clustering is denoted by a letter and each community by a number (i.e. community **B4** corresponds to community **4** for the clustering **without the geographical dimension**). The second line in each cell denotes the community from clustering **A** that is most similar and the associated similarity value. The similarity value is computed between the representative investors of said community and all communities of the complete clustering following eq. 2.3.

Community	Complete Clustering (**A**)	Clustering Without Countries (**B**)	Clustering Without Sectors (**C**)
0	General investors active whole period	General investors active whole period Similarity with community A0 : 0.931	General investors active whole period Similarity with community A0 : 0.956
1	General investors active pre-2008 crisis	General investors active pre-2008 crisis Similarity with community A1 : 0.960	General investors active pre-2008 crisis Similarity with community A1 : 0.96
2	Accelerators [72, 73]	Accelerators and incubators Similarity with community A2 : 0.92	Accelerators Similarity with community A2 : 0.915
3	Early-stage investors post-2008 crisis	Early-stage investors low amounts post-2014 Similarity with community A3 : 0.857	Early-stage investors post-2008 crisis Similarity with community A3 : 0.935
4	EU-focused investors	Early-stage investors low amounts post-2008 crisis Similarity with community A3 : 0.887	Canada-focused investors Similarity with community A9 : 0.956
5	Late-stage investors	Late-stage investors Similarity with community A5 : 0.882	General investors active post-2008 crisis Similarity with community A8 : 0.964
6	China-focused investors	China-focused investors Similarity with community A6 : 0.954	EU-focused investors Similarity with community A4 : 0.870
7	Health Care-focused investors	Health Care-focused investors Similarity with community A7 : 0.988	Health Care-focused investors Similarity with community A7 : 0.813
8	General investors active post-2008 crisis	General investors active post-2008 crisis Similarity with community A8 : 0.877	China-focused investors Similarity with community A6 : 0.989
9	Canada-focused investors	"Next-generation" post-2014 general investors Similarity with community A8 : 0.874	Japan and India-focused investors Similarity with community A10 : 0.978
10	Japan and India-focused investors		"Next-generation" post-2014 general investors Similarity with community A3 : 0.863

Community	Clustering Without Time (**D**)	Clustering Without Series (**E**)	Clustering Without Amounts (**F**)
0	General investors active whole period Similarity with community A8 : 0.899	General investors active whole period Similarity with community A0 : 0.969	General investors active whole period Similarity with community A0 : 0.956
1	Middle-stage investors active whole period Similarity with community A0 : 0.881	Early-stage investors active post-2008 crisis Similarity with community A3 : 0.904	General investors active pre-2008 crisis Similarity with community A1 : 0.988
2	General investors active post-2008 crisis Similarity with community A8 : 0.908	UK-focused early-stage investors Similarity with community A4 : 0.791	North America-focused incubators Similarity with community A9 : 0.805
3	North America-focused incubators Similarity with community A9 : 0.814	General investors active pre-2008 crisis Similarity with community A1 : 0.976	Accelerators Similarity with community A2 : 0.888
4	EU-focused investors Similarity with community A4 : 0.885	EU-focused investors Similarity with community A4 : 0.886	UK-focused early-stage investors Similarity with community A4 : 0.791
5	Very early-stage investors active post-2008 crisis (UK and US) Similarity with community A2 : 0.868	"Next-generation" post-2014 general investors Similarity with community A8 : 0.899	EU-focused investors Similarity with community A4 : 0.898
6	Early-stage investors active post-2008 crisis Similarity with community A3 : 0.949	China-focused investors Similarity with community A6 : 0.908	General investors active post-2008 crisis Similarity with community A8 : 0.923
7	General investors active pre-2008 crisis Similarity with community A1 : 0.933	Health Care-focused investors Similarity with community A7 : 0.969	Early-stage investors active post-2008 crisis Similarity with community A3 : 0.901
8	Israel-focused investors Similarity with community A0 : 0.829	Canada-focused investors Similarity with community A9 : 0.977	"Next-generation" post-2014 general investors Similarity with community A3 : 0.859
9	China-focused investors Similarity with community A6 : 0.992	Accelerators Similarity with community A2 : 0.929	China-focused investors Similarity with community A6 : 0.992
10	Health Care-focused investors Similarity with community A7 : 0.990	Japan and India-focused investors Similarity with community A10 : 0.953	Health Care-focused investors Similarity with community A7 : 0.973
11	Japan and India-focused investors Similarity with community A10 : 0.973		Japan and India-focused investors Similarity with community A10 : 0.976

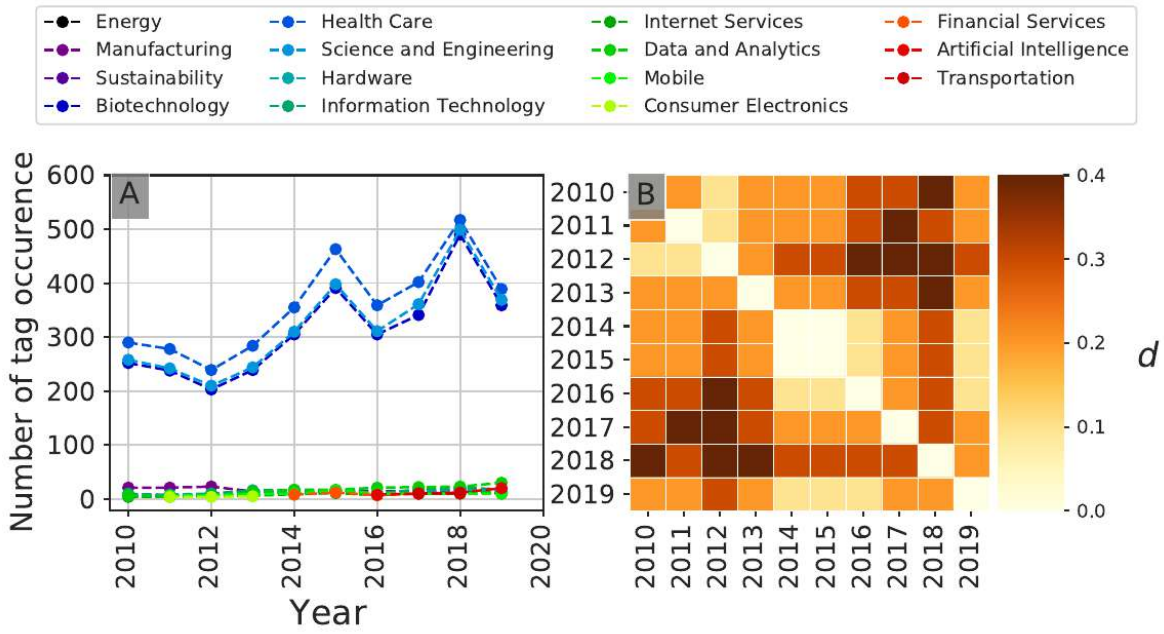

Figure 2.10: **Temporal evolution of the investment patterns of community A7.** Temporal community investment patterns of the target startups' sectoral tags for each year aggregated at the community level. Community **A7** is comprised of Health Care-specialized venture capitalists. Panel **A** shows for each year the ten tags that received the most investments, panel **B** shows the community self-difference index described in Eq. 2.4, with two markedly different areas of coherence, before and after 2014-2015.

2.4.2 Clustering factor analysis highlights underlying investment patterns

Since the 5 characteristic dimensions are based on domain knowledge, we ran the clustering algorithm 5 additional times, each time using only 4 of the 5 dimensions previously defined, computing the representative investors of all communities for each of these alternative clusterings in order to understand the characteristics of the new communities. Fig. 2.26 shows the representative investor of community **B6** resulting from a clustering without the geographical investment dimension. Surprisingly, the community shows a strong focus on the Chinese startup market, with around 80% of all investments targeting China-based startups although the geographical dimension was not taken into account, therefore suggesting the existence of an underlying structure: the existence of an investment pattern according to the 4 other investment dimensions that is actually characteristic of investors investing mostly in China. Similarly, Fig. 2.27 shows the representative investor of community **C7** resulting from a clustering without the sectoral dimension, but shows a community strongly focused on Health Care startups (around 17%, 18% and 15% of investments in *Science and Engineering*, *Health Care* and *Biotechnology* respectively) not unlike the community shown in Fig. 2.22, even though sectors were not taken into account in this clustering.

Following these observations, we systematically investigate the bivariate distributions

for all pairwise combinations for each alternative clustering, with the discrete bivariate distribution f of group g at coordinates (m, n) defined as :

$$f_g(m, n, k_1, k_2) = \frac{\sum_{\epsilon=1}^{\epsilon=T} i_\epsilon^{k_1}(m)\ i_\epsilon^{k_2}(n)}{\sum_{v=1}^{v=V}\sum_{w=1}^{w=W}\sum_{\epsilon=1}^{\epsilon=T} i_\epsilon^{k_1}(v)\ i_\epsilon^{k_2}(w)} \tag{2.5}$$

where investor distribution k_1 has dimension V and k_2 has dimension W with group g being comprised of T investors.

Geographical

Fig. 2.11 shows the resulting bivariate distribution for all pairs of dimensions for community **B6**, here presented as heatmaps. It shows that **B6** investors take part mostly in series A investments between $10M and $20M after 2015, which could correspond to a pattern characteristic of China-focused investors in our sample. For all bivariate distributions shown in Fig. 2.11 (community **B6**) and Fig. 2.12 (community **A6**), both communities display virtually identical behaviors : most likely due to this underlying investment pattern, taking into account the geographical dimension is *not* necessary to characterize this cluster despite its very strong geographical footprint.

Sectoral

Similarly, Fig. 2.13 shows the resulting bivariate distribution for all pairs of dimensions for community **C7**. It shows that **C7** investors invest mainly in series B rounds between $20M and $50M in North American ventures, which appears to be an investment pattern for investors specialized in Health Care in our sample. Fig. 2.14 shows community **A7** resulting from the complete clustering. Fig. 2.13 and Fig. 2.14 show a strong agreement in terms of *Series* and *Amounts* of investments but still display slight differences as community **A7** has been active for a longer time than community **C7**. We therefore observe different *generations* of Health Care-focused investors with the newer generations associated with a wider scope of investment in terms of sectors. These new investors tend to invest in Health Care-oriented companies with a stronger IT component in the latter part of the 2010s (see Fig. 2.13), a pattern not found in Fig. 2.14. This suggests that the current shift in Health Care venture funding (linked notably to the use of Artificial Intelligence solutions) could on a global level not be the result of a shift of focus of traditional Health Care-focused investors but rather the outcome of the emergence of a new group of investors in the domain.

Temporal

Again in a similar manner, and analyzing this time the clustering computed without the temporal dimension, Fig. 2.28 shows the representative investor of community **D7**, associated with a very specific temporal pattern of investment that appears markedly similar to community **A1** from the complete clustering (see Fig. 2.16), even though the temporal

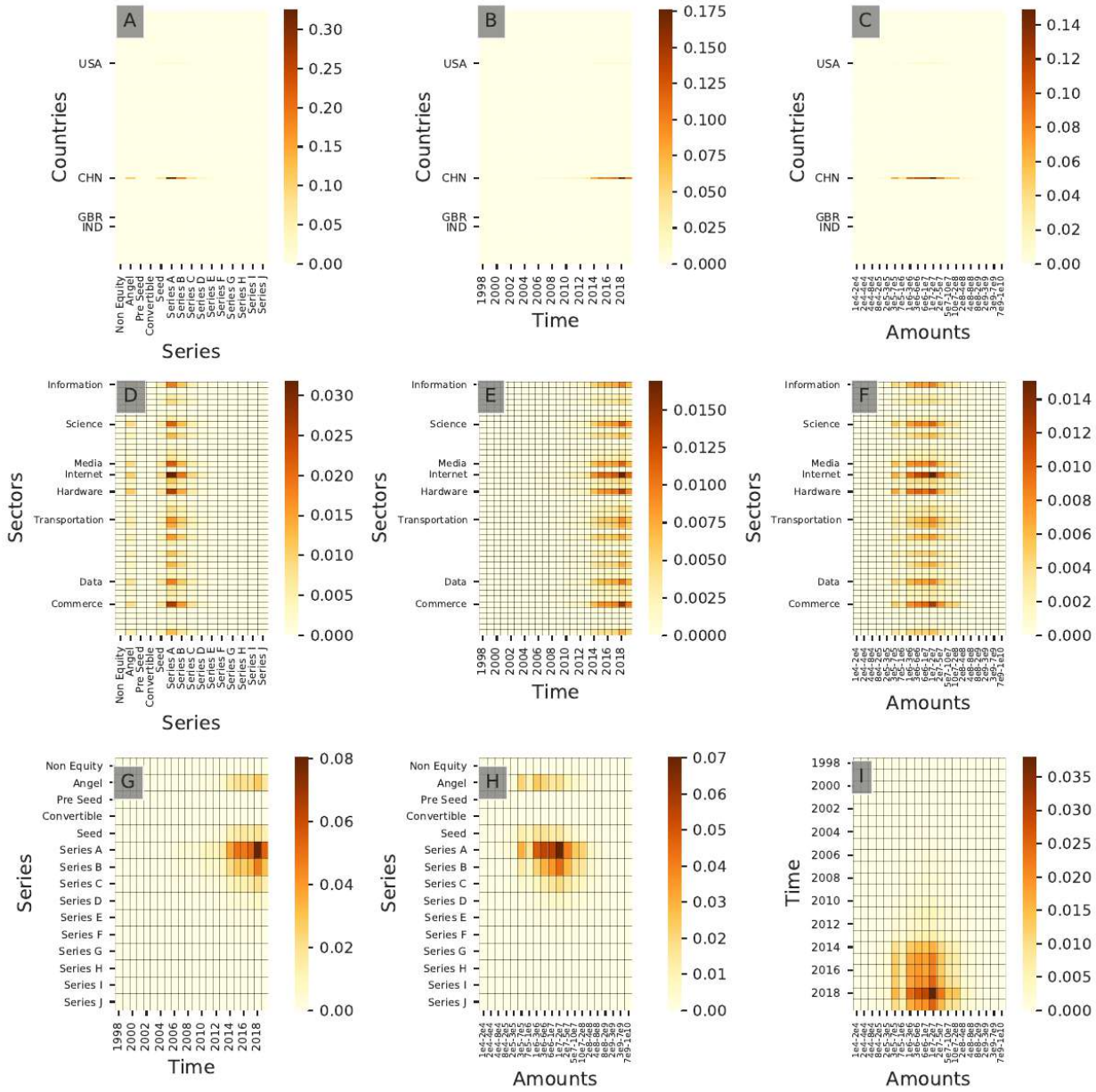

Figure 2.11: **Cross-interaction heatmaps for community B6.** This community corresponds to China-focused investors. Only the top 8 sectors and the top 4 countries in terms of frequency of investments are labeled for readability purposes.

dimension was excluded in the case of **D7**. This observation therefore again suggests the existence of underlying investment patterns associated with investors. Here, historical, older generation investors appear to have been clustered together independently of their temporal activity, and rather on the basis of a qualitatively specific investment pattern that differs from those of newer generation venture capital firms.

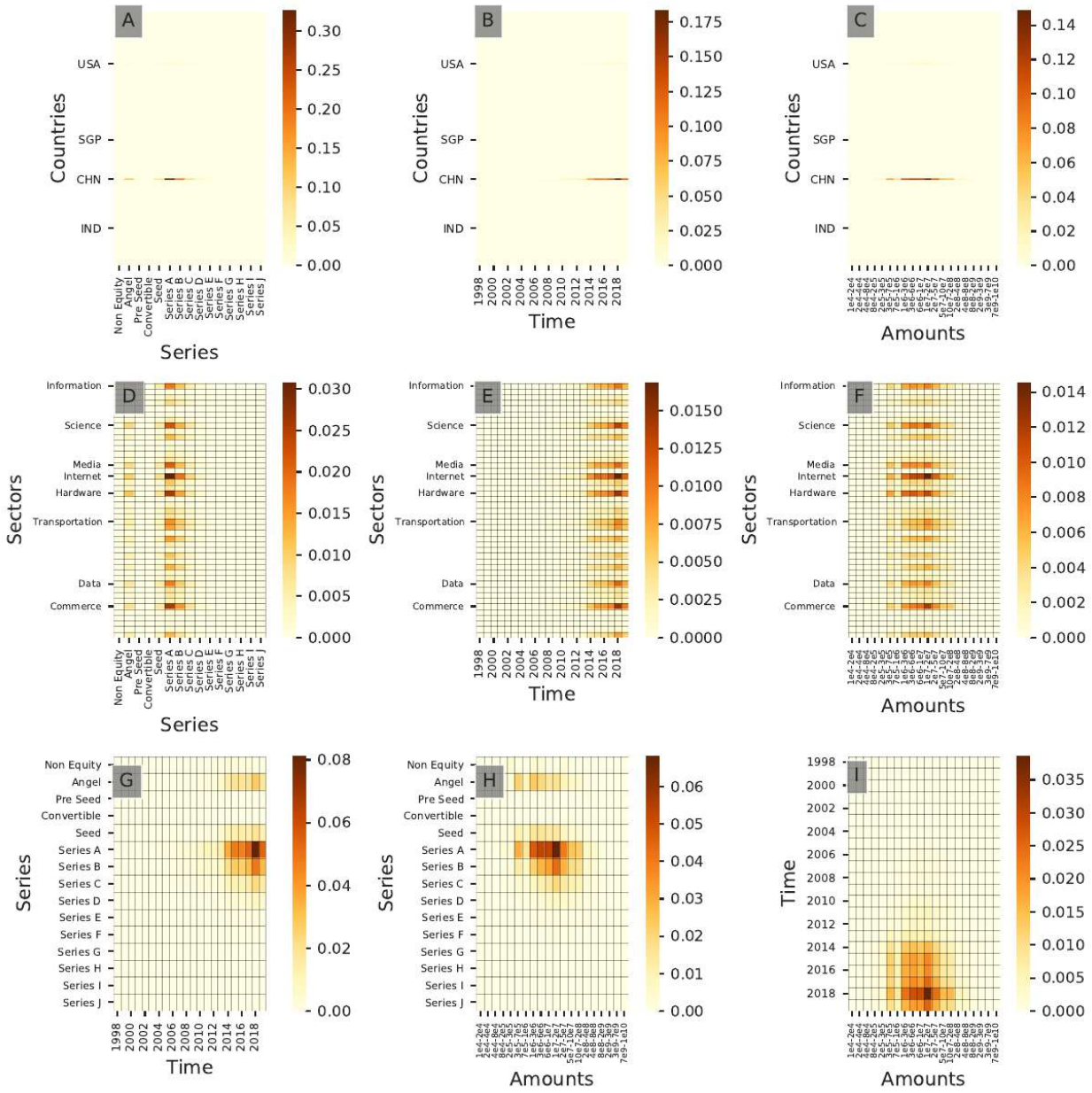

Figure 2.12: **Cross-interaction heatmaps for community A6.** This community corresponds to China-focused investors.

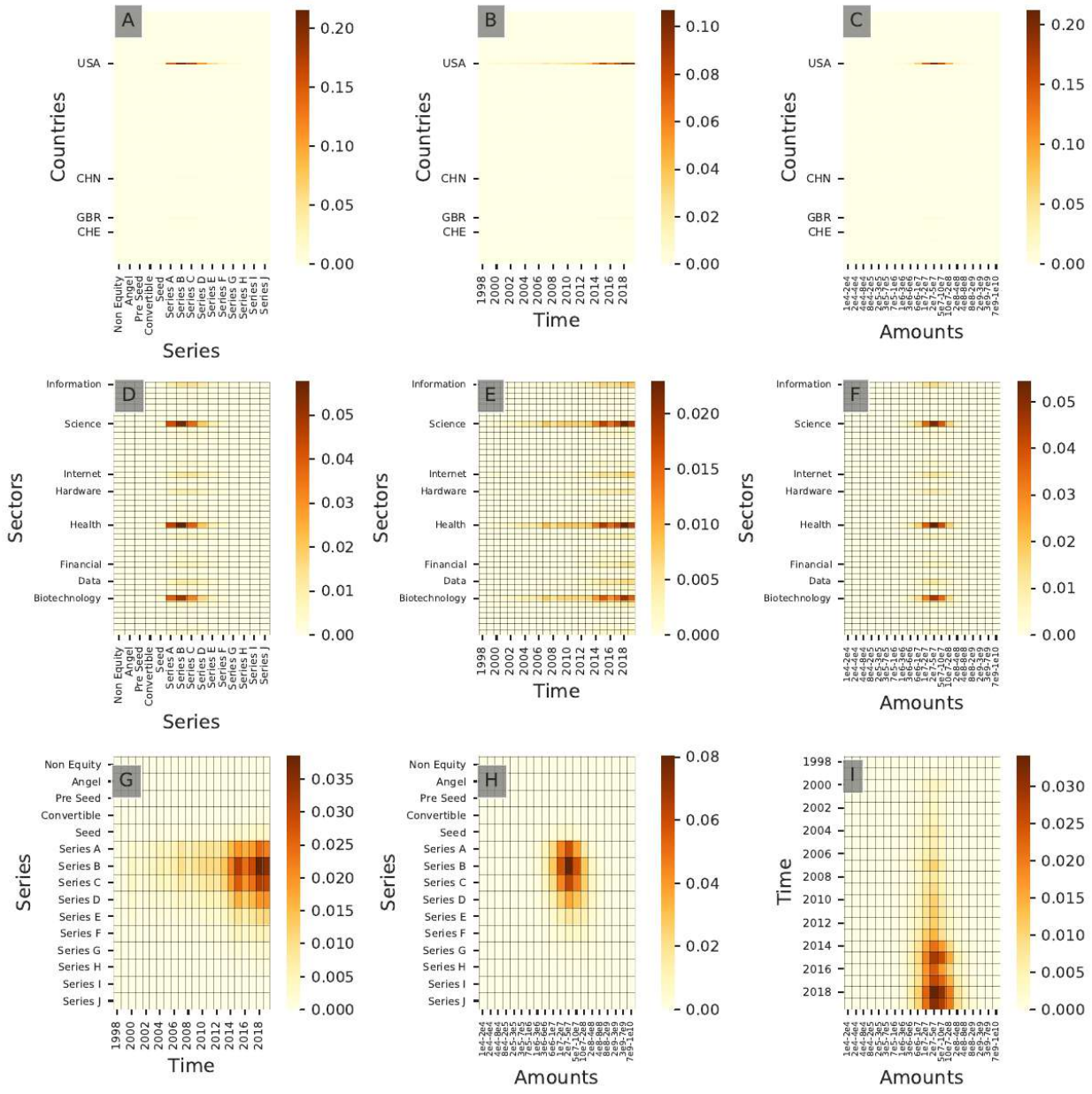

Figure 2.13: **Cross-interaction heatmaps for community C7.** This community corresponds to a Health Care-focused community of investors. Only the top 8 sectors in terms of total number of investments and the top 4 countries of investment are labeled for readability purposes.

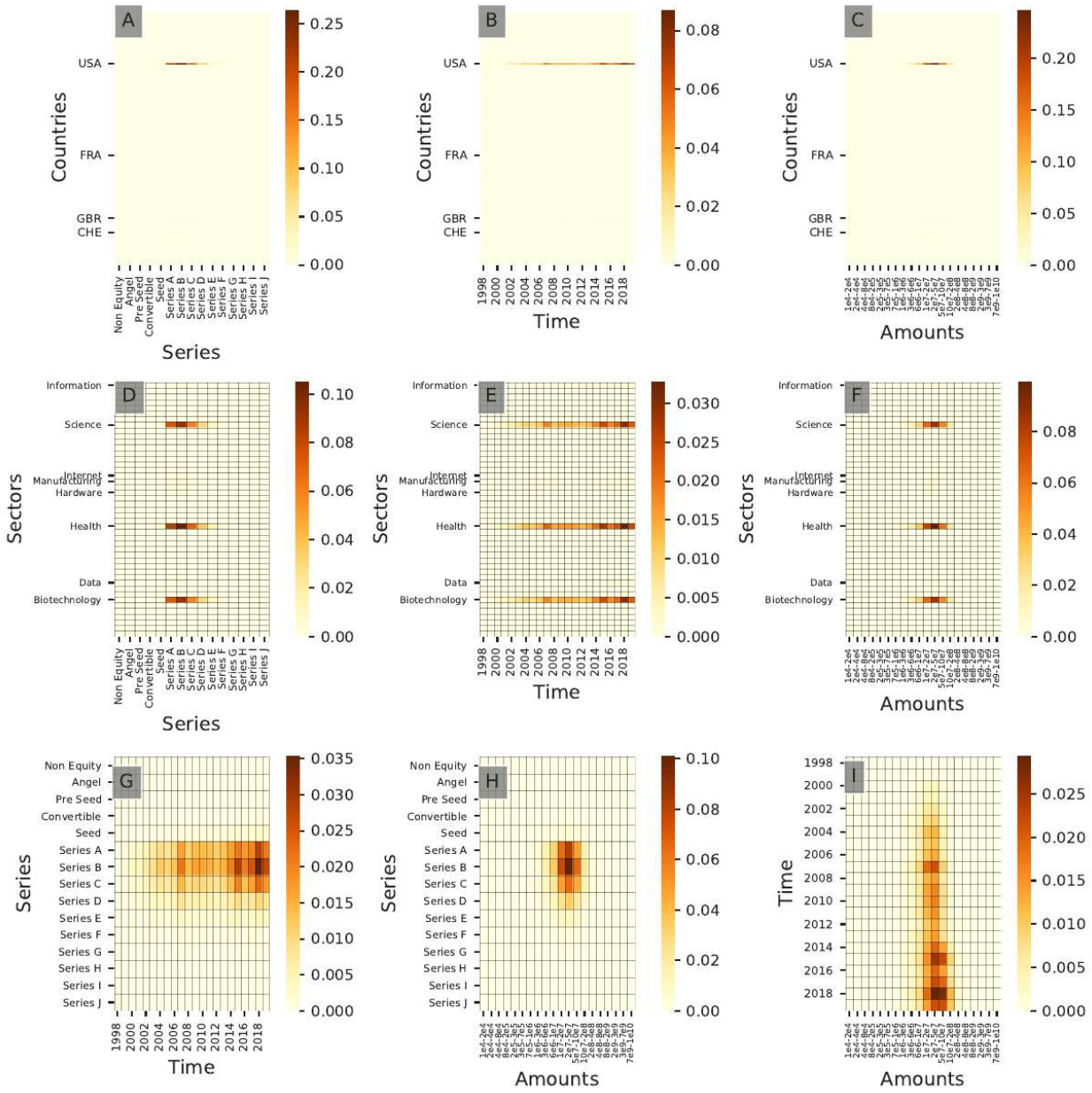

Figure 2.14: **Cross-interaction heatmaps for community A7.** These distributions correspond to Health Care specialists.

2.5 Conclusion

In this chapter, we approached investors through clustering methods in order to help us and fellow researchers make a better sense of the "venture capital community", perhaps in the sense of advocating for the end of their analysis as that of an homogeneous community. We thus described a novel approach to quantitatively group startup investors based only on the characteristics of their investments, as gathered from a bipartite investor-startup network. This clustering approach results in interpretable and homogeneous subgroups of investors with markedly different profiles, which we hope could prove helpful for the community of researchers interested in studying venture capital communities and networks by allowing them to differentiate *among* venture capitalists. In that sense, "the" venture capital community, as often referred to, might actually be composed of several venture capital communities whose investment behaviors and in particular whose co-investment behaviors might considerably differ. As a consequence, we would plead for some of the literature on venture networks to be assessed again on each of the venture communities separately, for instance with respect to the relationship between network position and centrality and the profitability of venture investments.

In addition, and by allowing the conditions under which investors are clustered according to our approach to vary, notably by reducing the number of characteristic dimensions taken into account, we were able to observe the presence of relatively surprising underlying and robust investment patterns characteristic of certain clusters of startup investors. For instance, the fact that some investors specialize as Health Care specialists seems to have consequences with respect to their other investment patterns notably in terms of funding amounts or funding rounds : we did observe a cluster of Healh Care-focused investors even when the sectoral dimension was not accounted for in the clustering. Similarly, the fact that some investors focus on investments in China also results in the existence of patterns with respect to their investment behaviors, once again in terms of funding amounts and funding rounds in particular : we indeed observed a cluster of investors focused on China even when the geography of investments was not taken into account. From a research point of view, these observations raise the issue of whether they would be the result of a behavioral phenomena or rather market outcomes. More broadly, the existence of such underlying patterns could also result in modifying how financial actors directly interpret and evaluate opportunities, compared then to such benchmarks.

Furthermore, similar underlying investment patterns were also observed to characterize different generations of investors, notably in relation to the 2008 financial crisis. We notably observed a cluster of investors mostly active before the 2008 crisis even when the temporal distribution of their investments was not taken into account. In our sample, this observation is particularly striking with respect to the aforementioned crisis, but we also observed preliminary evidence of a similar phenomenon in the case of Health Care focused investors with 2014 as a breaking point, which we can relate to the significant increase in startup investment activity that occurred around that date. Altogether, and adding also that the cluster of so-called accelerators (**A2**) also corresponds to a completely new "species" of investors that appeared in the late 2000s, these preliminary observations might suggest a mechanism that would evoke the notion of *speciation* in ecology: whenever the "financial

environment" would change, newer "species" of investors could appear in an evolutionary way, by seizing the newer opportunities offered by the new environment, while existing investors might either adapt or stay locked in their previous patterns even though these patterns might eventually not represent an adaptive advantage in a new financial environment. Rather than simply suggesting an evolutionary perspective, these observations could also shed more light on the determinants of success for so-called "Limited Partners" [182], i.e. investors in venture capital funds, by potentially providing a supplementary explanation of why returns would differ systematically across limited partners [63]. They could also provide limited partners and other actors in the finance community themselves with a new understanding of the dynamics of innovation in the venture capital market.

Cluster 0	Cluster 1	Cluster 2	Cluster 3
CRV	Threshold Ventures	Masschallenge	Marc Cuban
Greylock	Venrock	Skydeck Berkeley	Band of Angels
Battery Ventures	Sigma Partners	MIT Media Lab	SV Angel
RRE Ventures	Fidelity Ventures	500 Startups	Scott Banister
Bain Capital Ventures	H.I.G. Capital	Techstars	Fabrice Grinda
GGV Capital	ABS Ventures	Y Combinator	Alexis Ohanian
Goldman Sachs	Polaris Partners	Kima Ventures	Betaworks
Kleiner Perkins Caufield Byers	Baird Capital	Start-Up Chile	Angelpad
Sequoia Capital	Cedar Fund	SOSV	Kickstart Seed Fund
Benchmark	Enterprise Partners	Chinaccelerator	Lerer Ventures

Cluster 4	Cluster 5	Cluster 6	Cluster 7
Seedcamp	Tiger Global	IDG Capital Partners	Sofinnova Ventures
Amadeus Capital Partners	Temasek	Ceyuan Ventures	Abingworth Management
Balderton Capital	KKR	SIG China	Frazier Healthcare Ventures
Index Ventures	T. Rowe Price	Shenzhen Capital Group	Sante Ventures
Partech	General Atlantic	Sequoia Capital China	SV Life Sciences
Alven Capital	Wellington Management	Vertex Ventures China	Orbimed Advisors
Xange Private Equity	Coatue	Qingsong Fund	Life Sciences Partners
IDInvest Partners	Iconiq Capital	Zhenfund	Oxford Bioscience Partners
Bayern Kapital	Google Capital	Baidu	Lilly Ventures
Iris Capital	Softbank Vision Fund	Matrix Partners China	Deerfield Management Company

Cluster 8	Cluster 9	Cluster 10
Silverton Partners	Celtic House Venture Partners	Mitsubishi UFJ Capital
First Round Capital	BDC Venture Capital	Mizuho Capital
Greycroft	Fonds de Solidarite FTQ	SMBC Venture Capital
Andreessen Horowitz	Inovia Capital	Omidyar Network
Ridge Ventures	Relay Ventures	Sequoia Capital India
GE Ventures	Innovacorp	East Ventures
Foundry Group	Anges Quebec	Mumbai Angels
Miramar Venture Partners	Founderfuel	Innovation Network Corp of Japan
Lux Capital	Venture Alberta	Nissay Capital
IA Ventures	Creative Destruction Lab	Itochu Technology Ventures

Table 2.2: ***Complete clustering*: Sample investors from each community.** Ten investors are manually chosen from each community to provide insights about the typology of investors.

2.6 Appendix

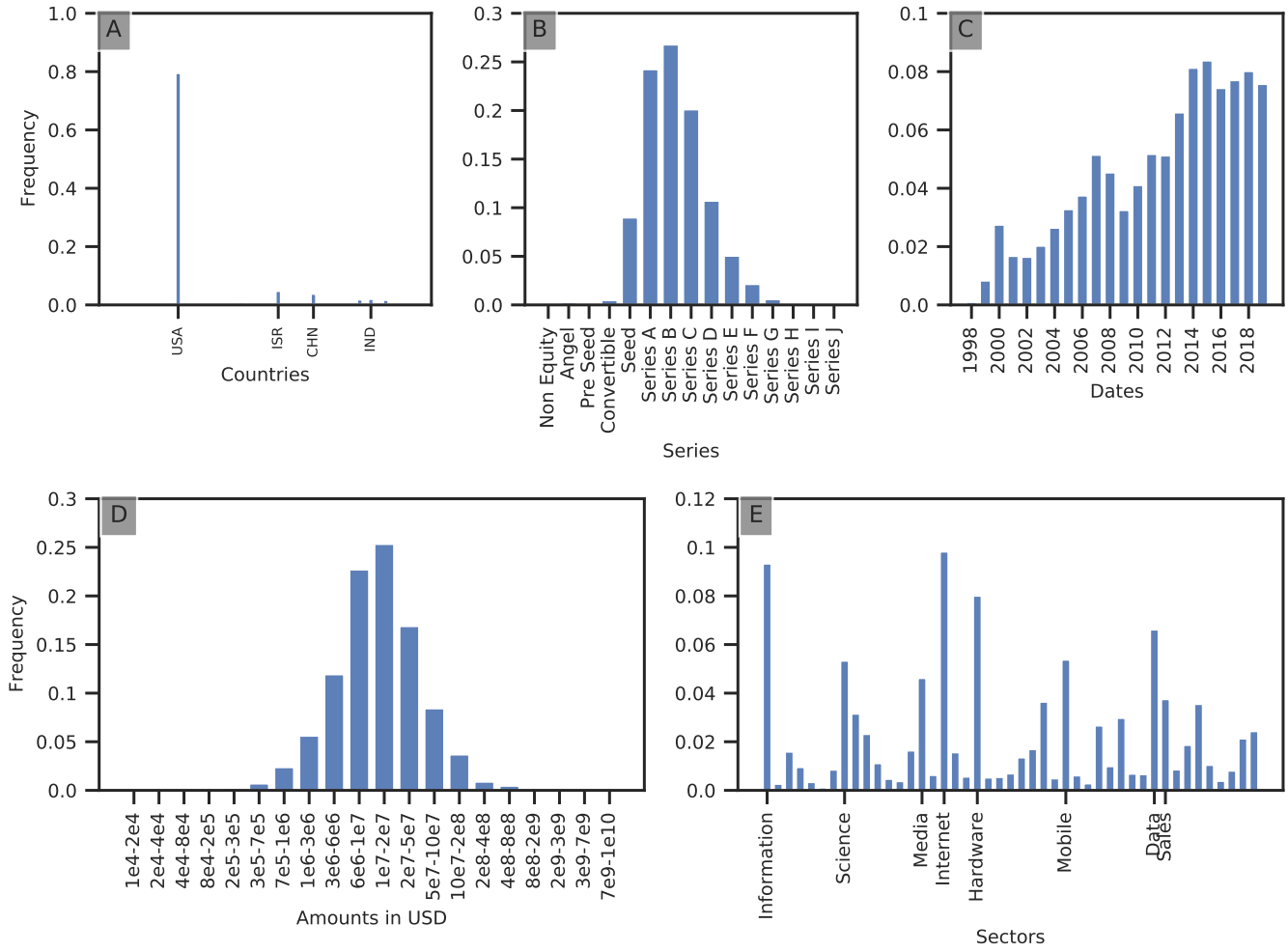

Figure 2.15: **Representative investor of community A0.**

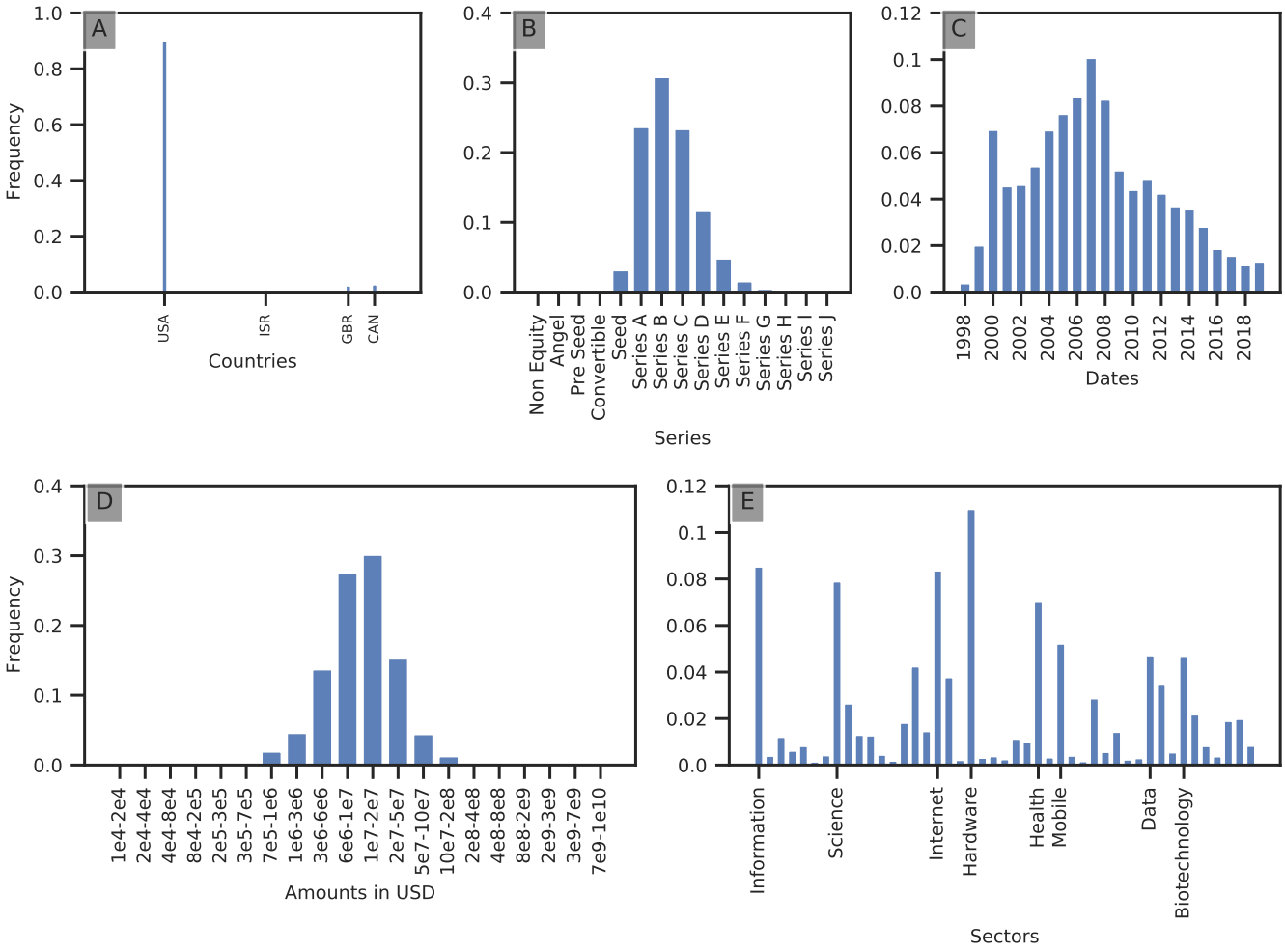

Figure 2.16: **Representative investor of community A1.**

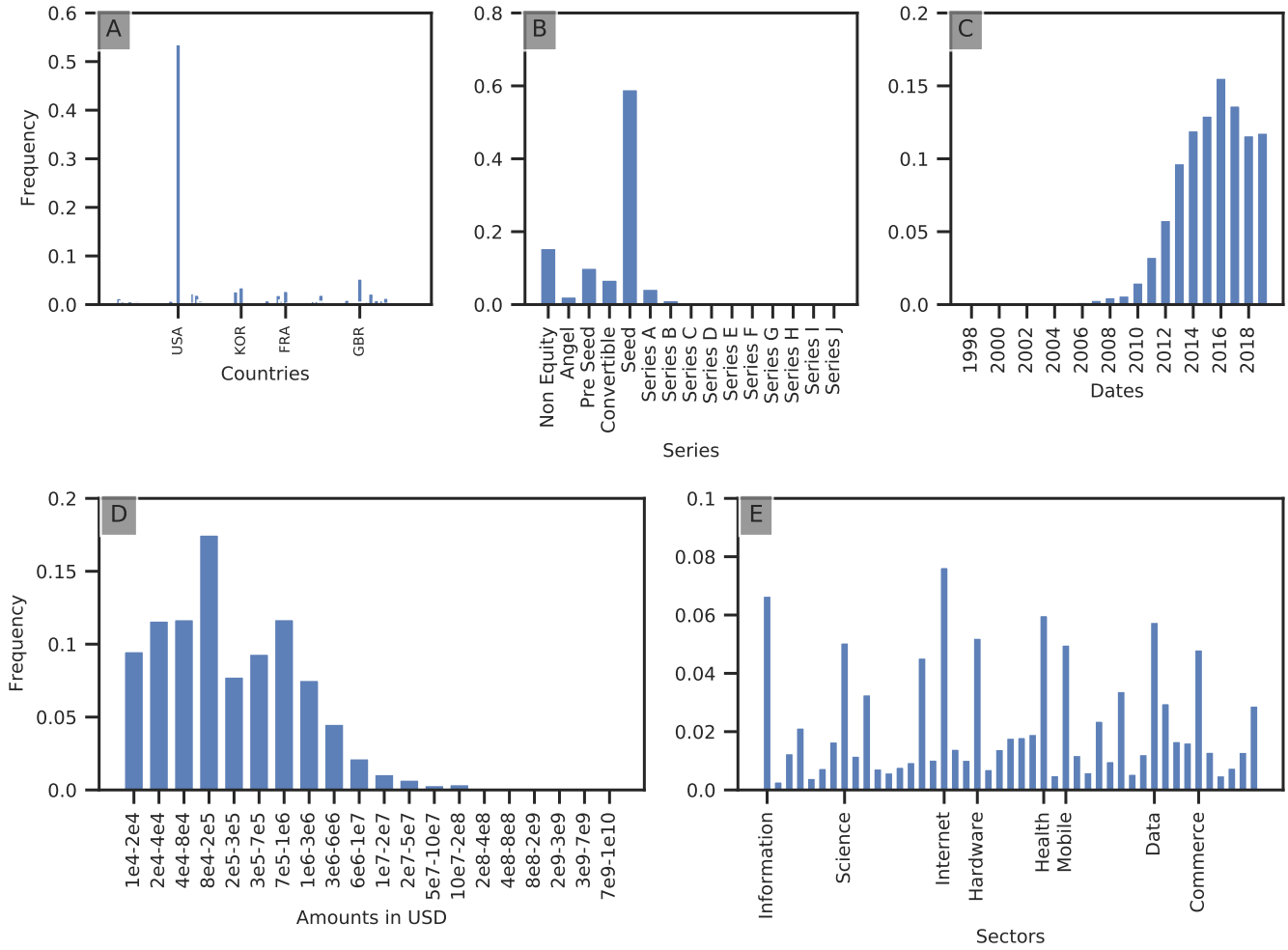

Figure 2.17: **Representative investor of community A2.**

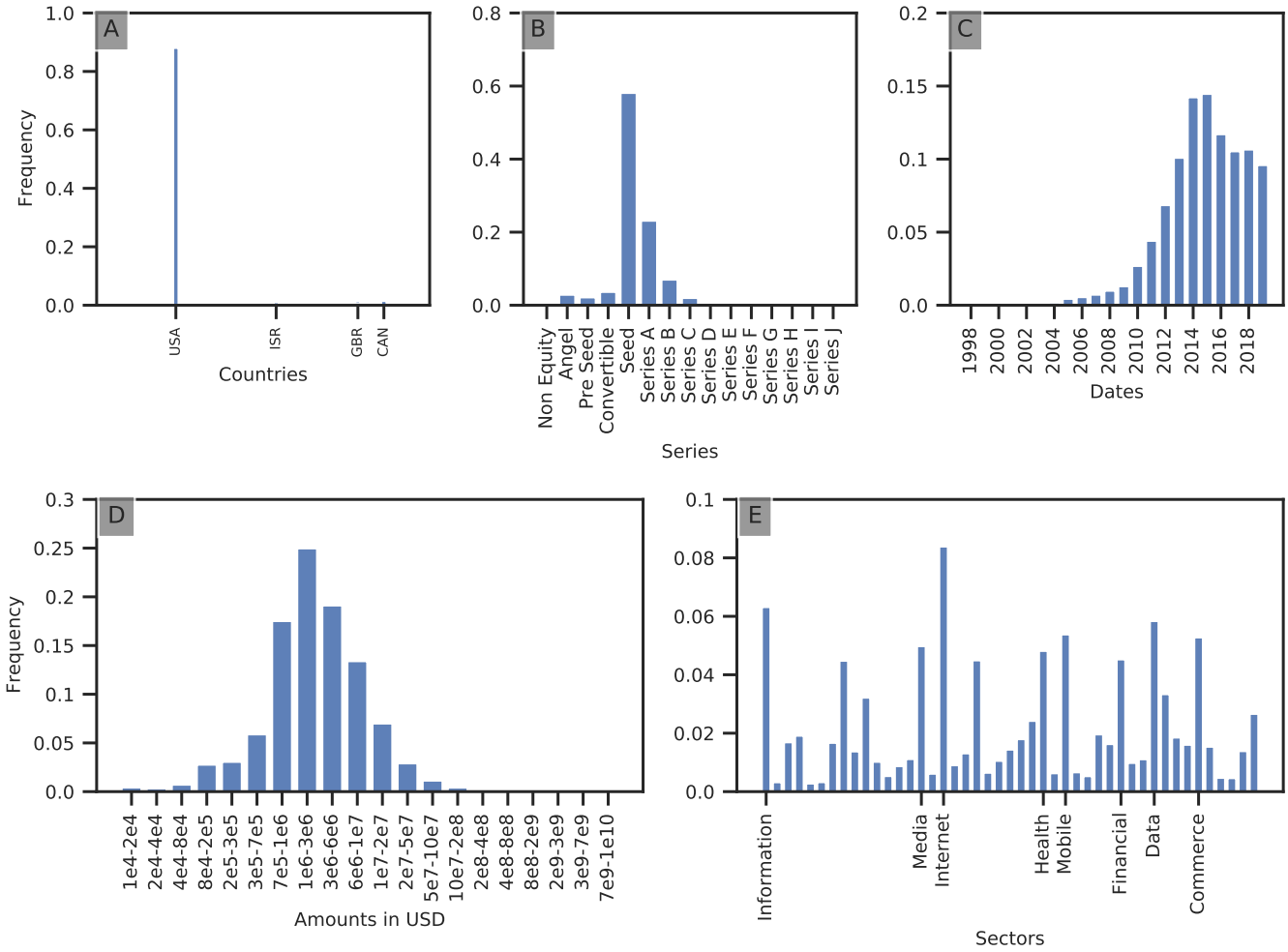

Figure 2.18: **Representative investor of community A3.**

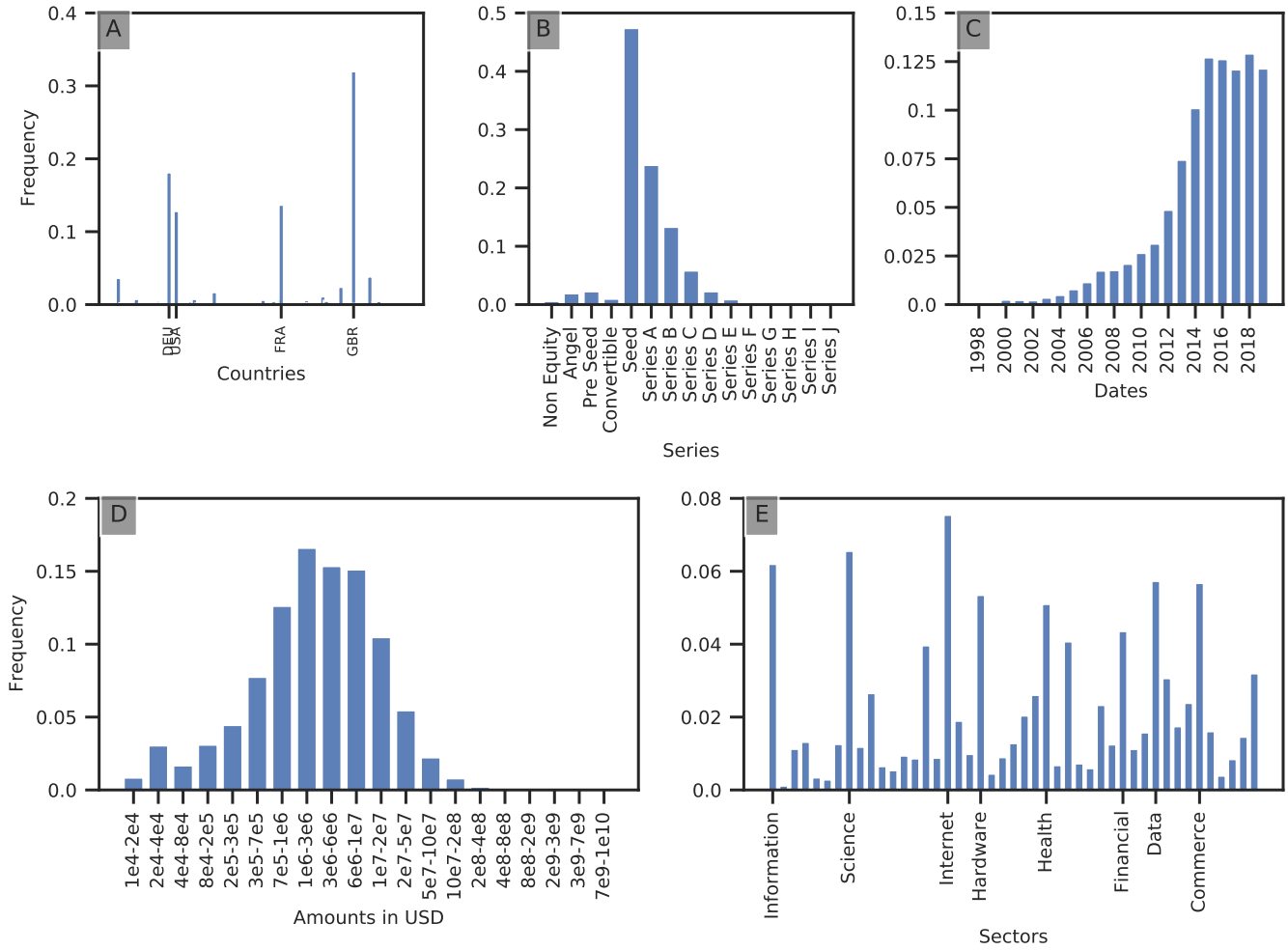

Figure 2.19: **Representative investor of community A4.**

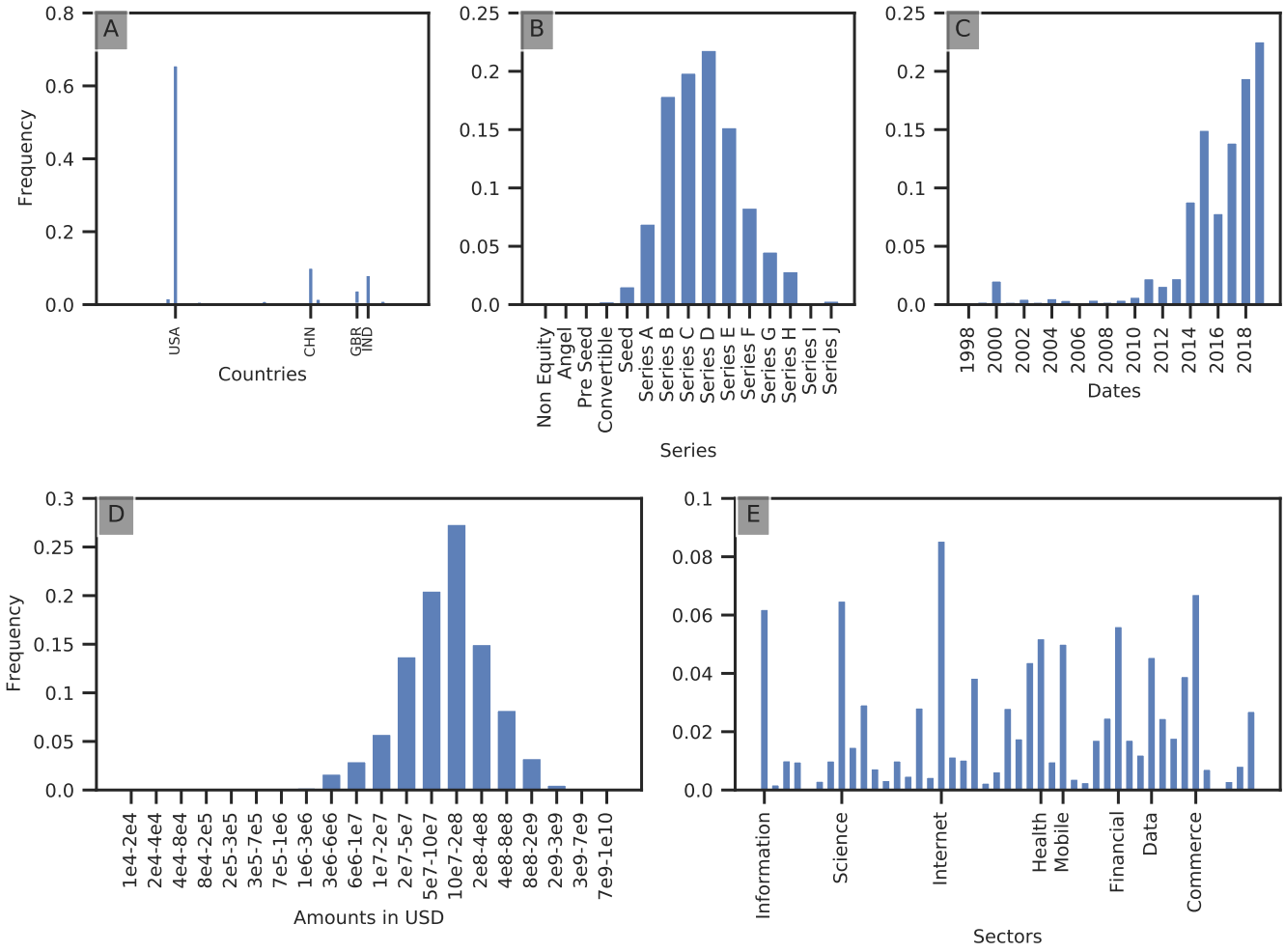

Figure 2.20: **Representative investor of community A5.**

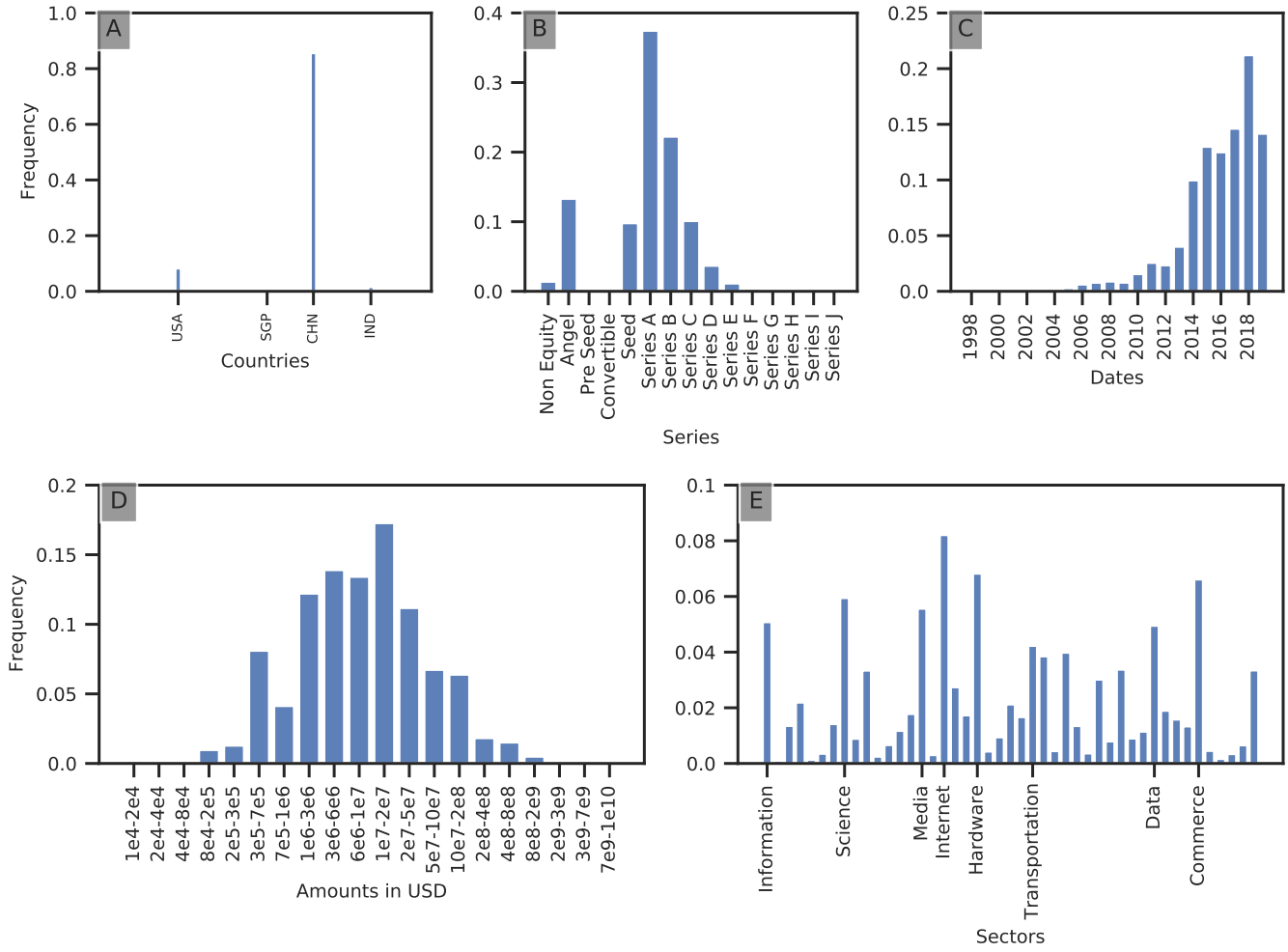

Figure 2.21: **Representative investor of community A6.**

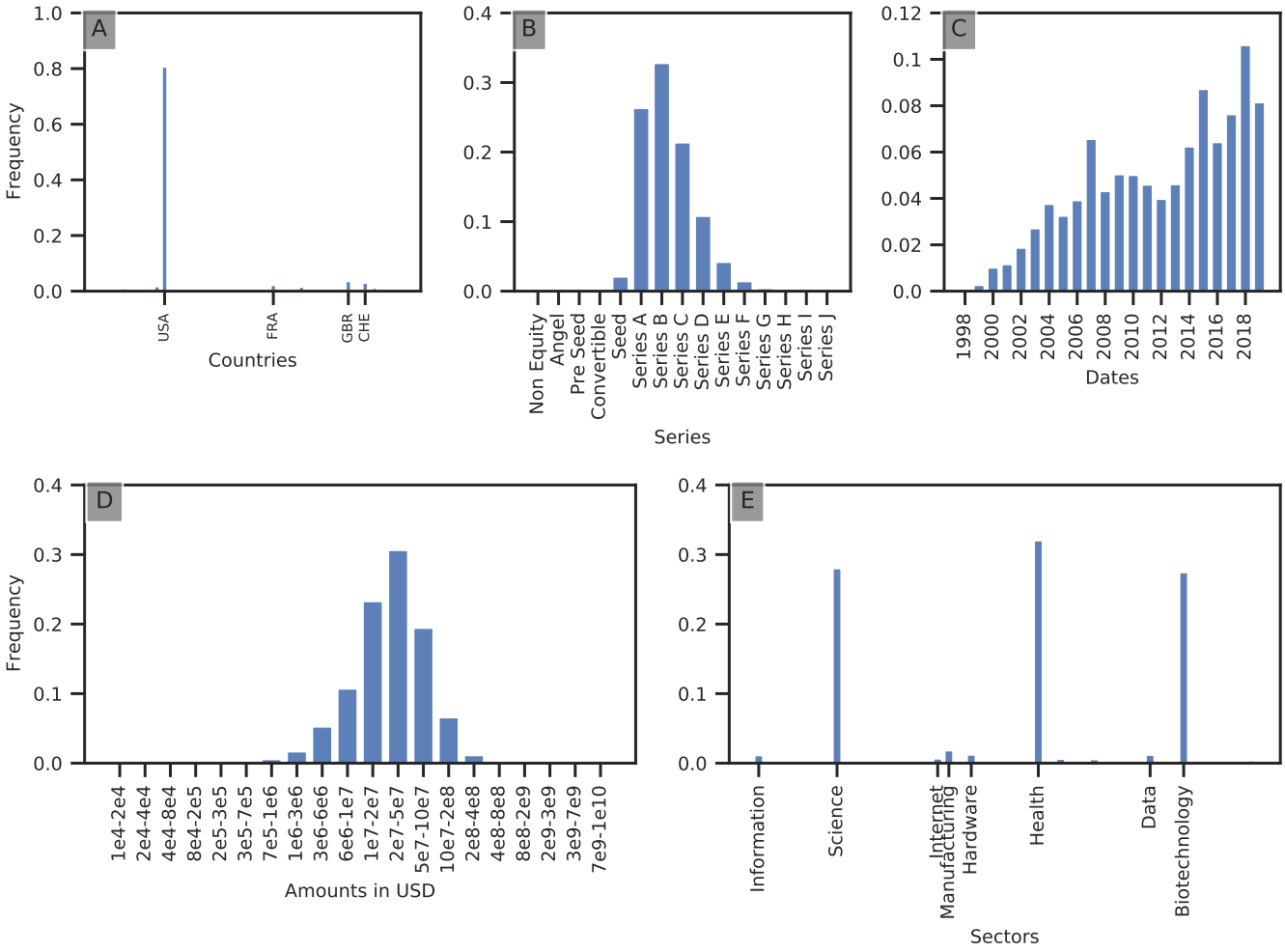

Figure 2.22: **Representative investor of community A7.**

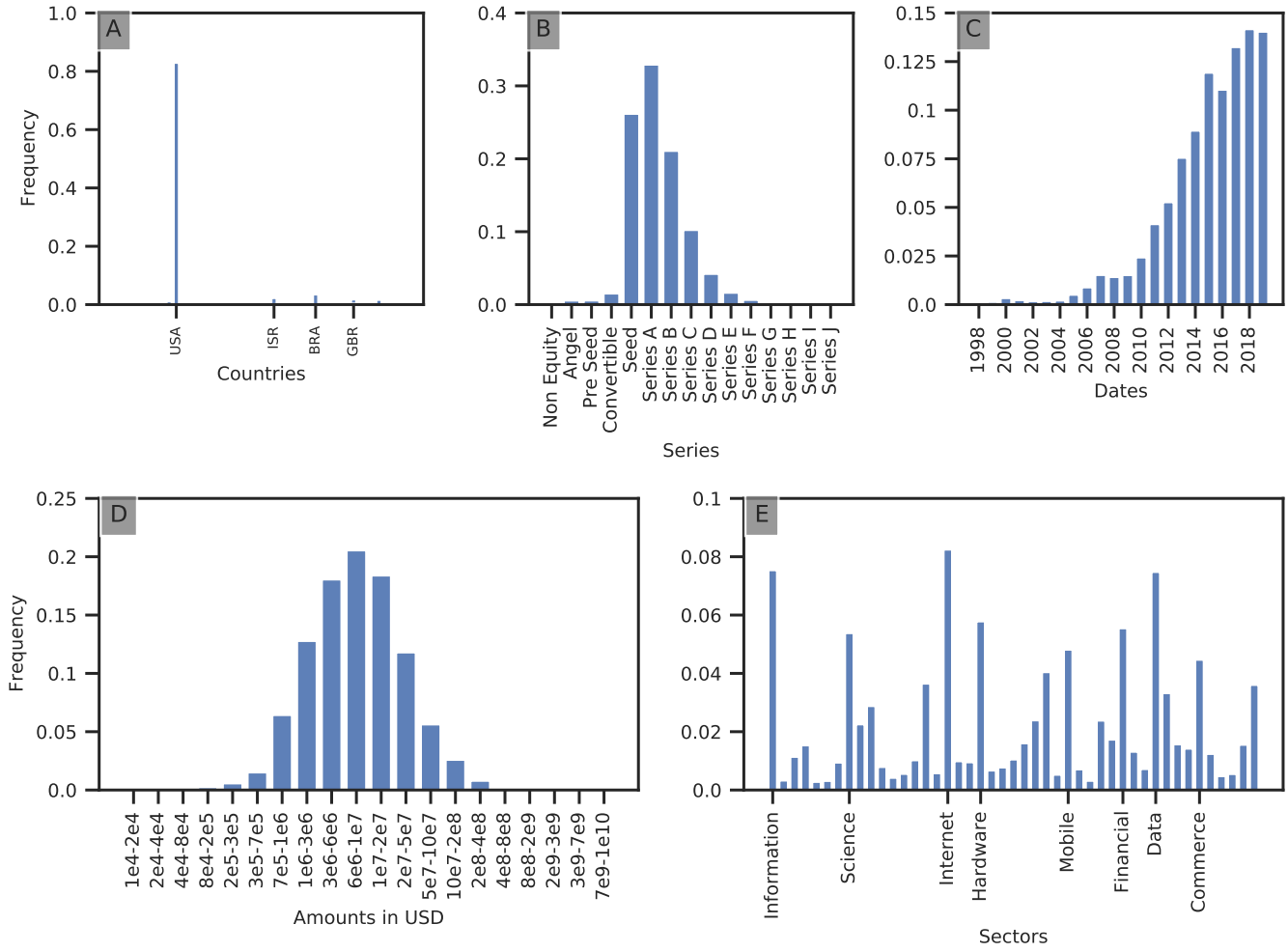

Figure 2.23: **Representative investor of community A8.**

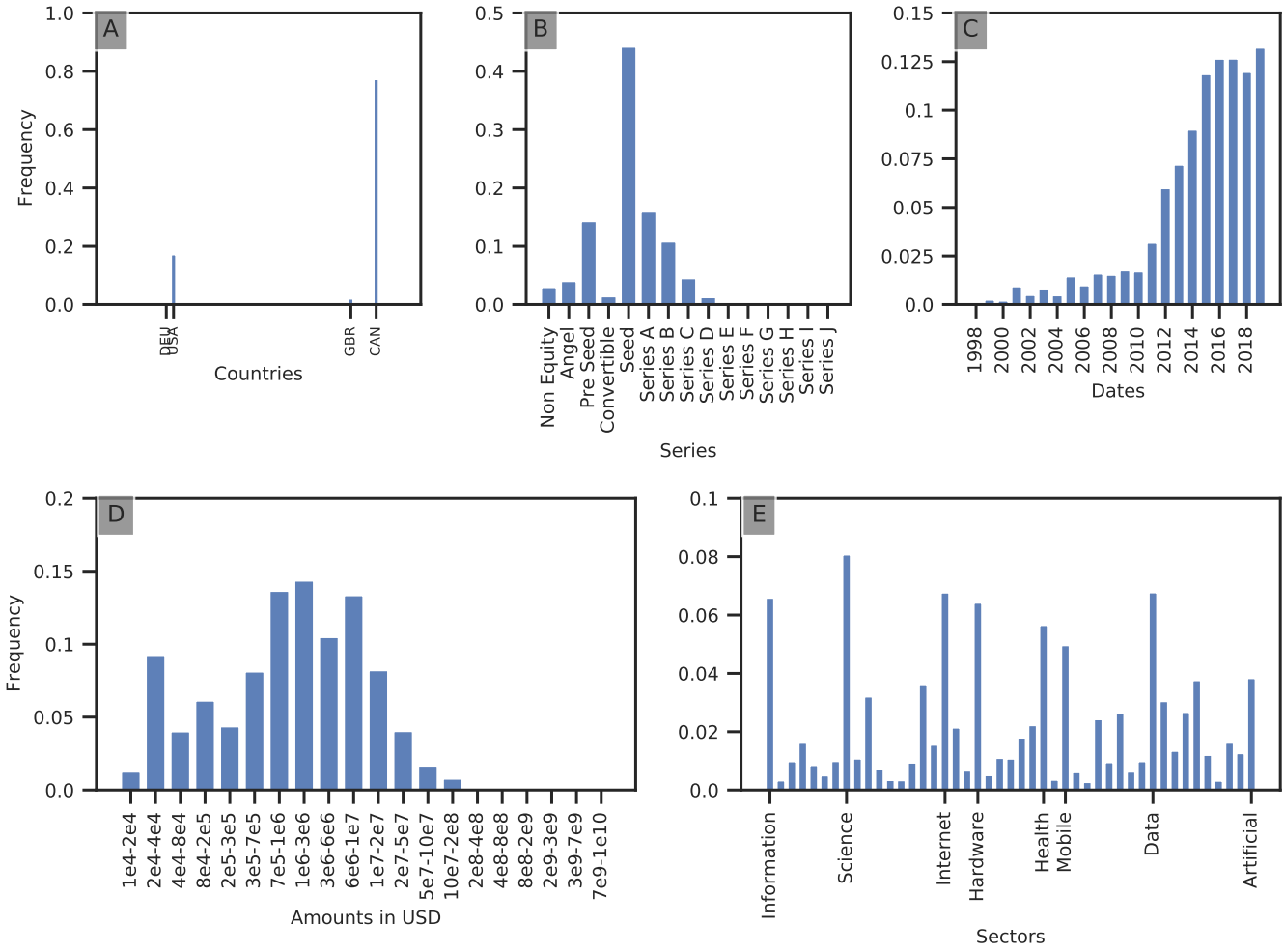

Figure 2.24: **Representative investor of community A9.**

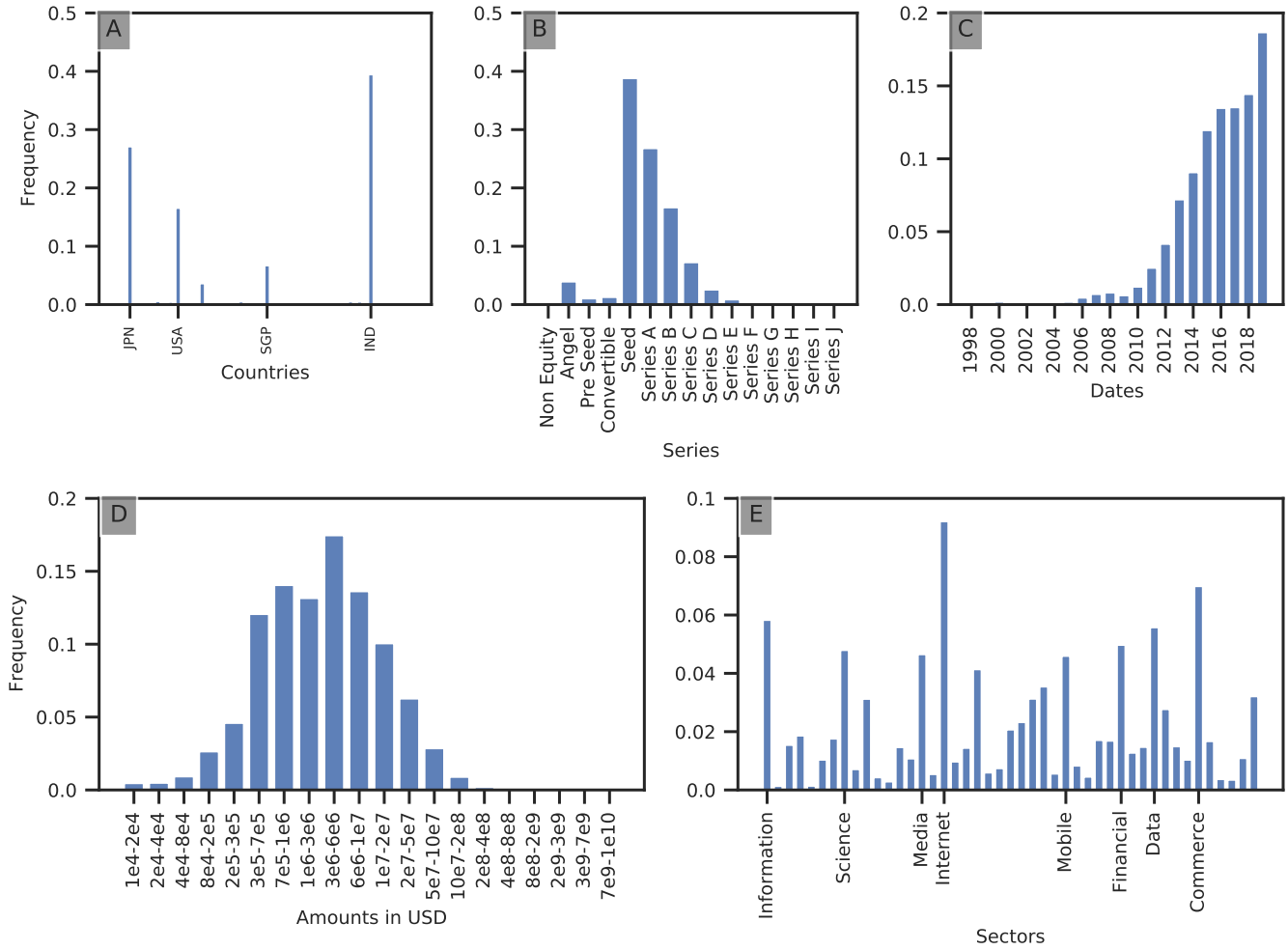

Figure 2.25: **Representative investor of community A10.**

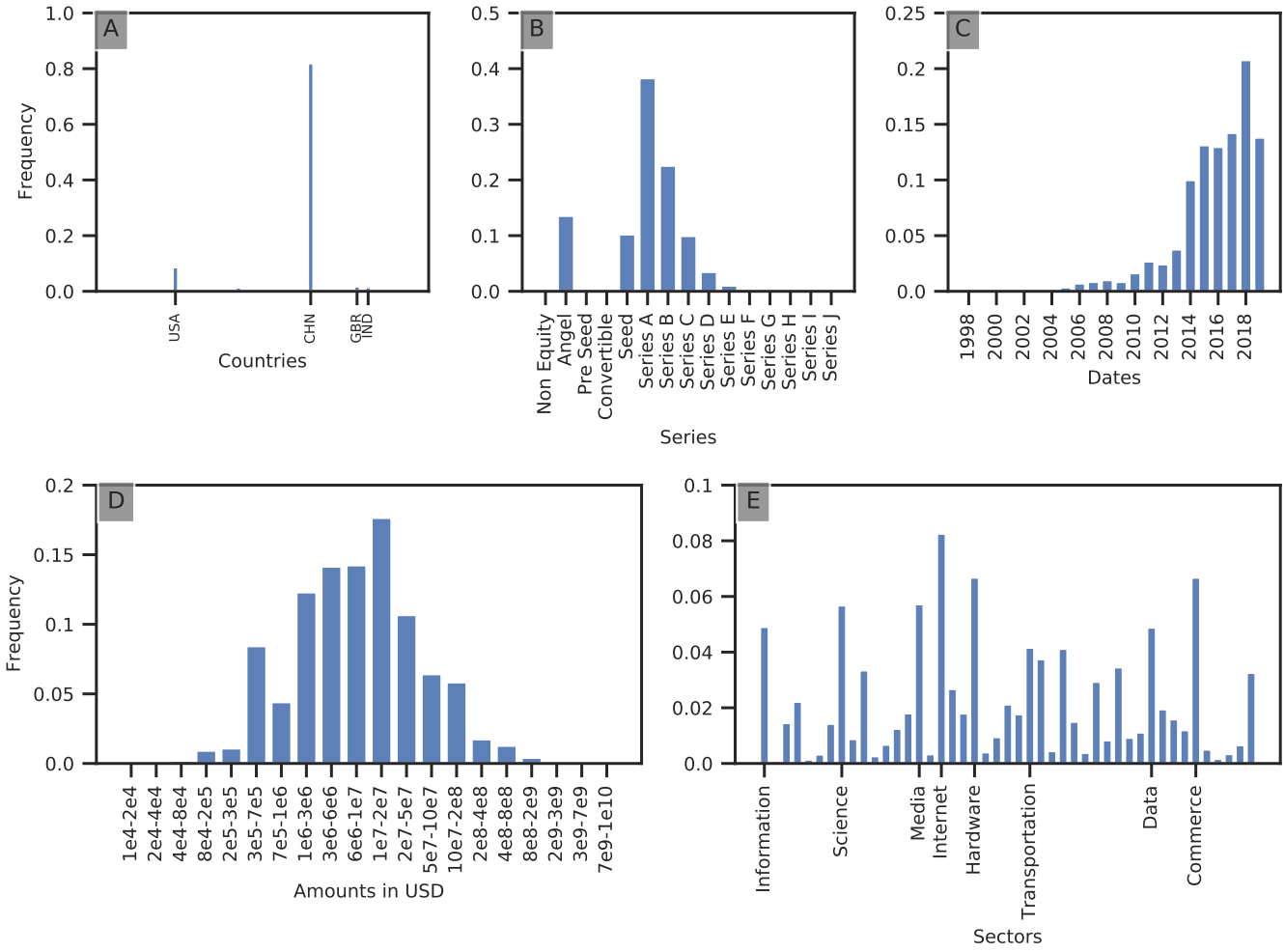

Figure 2.26: **Representative investor of community B6.**

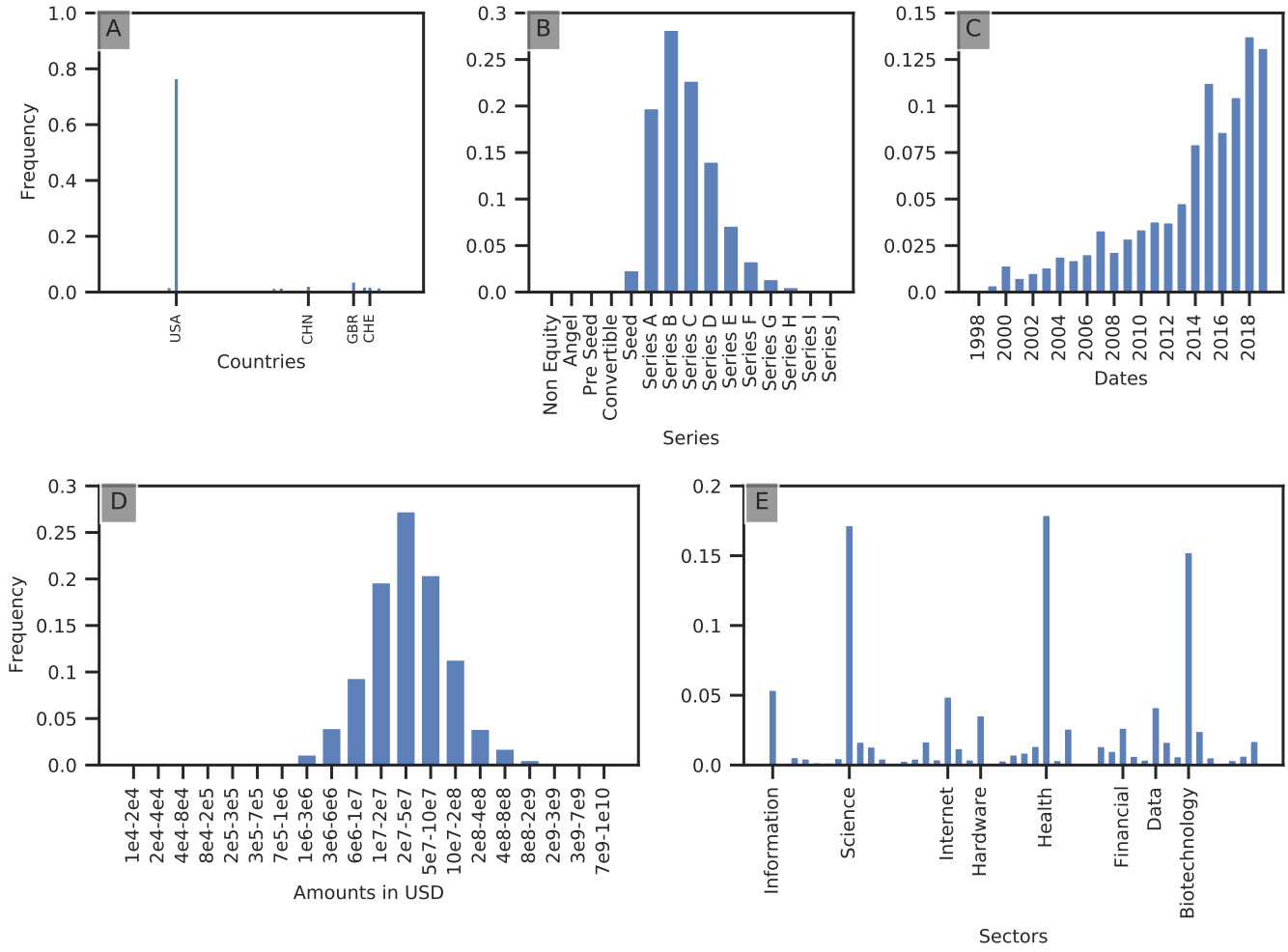

Figure 2.27: **Representative investor of community C7.**

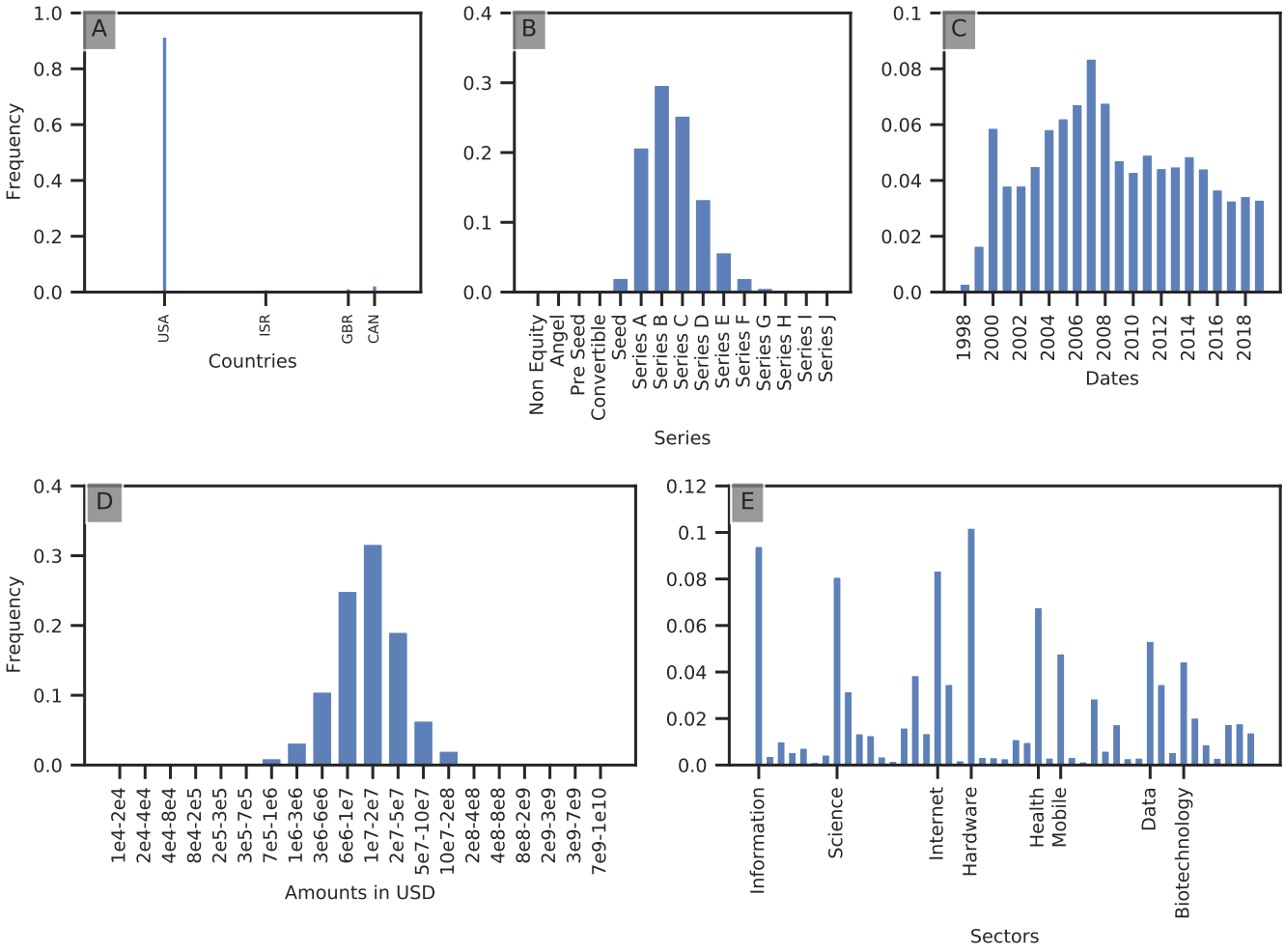

Figure 2.28: **Representative investor of community D7.**

Chapter 3

Automatic text-based grouping of thematically similar documents

Section 3.2 of this chapter is based on Carniel, T., Cazenille, L., Dalle, J. M., & Halloy, J., 2022 : *Using natural language processing to find research topics in Living Machines conferences and their intersections with Bioinspiration & Biomimetics publications* (Bioinspiration & Biomimetics, 17(6), 065008).

In order to discover clusters of thematically-similar patents based on individual patent information to study the investor-technology graph at the "species"-level, a specific methodology is required. Indeed, the ever-increasing production of scientific and technical materials, resulting in hundreds of thousands or millions of documents relevant to a given question, comes at a price : there is too much to read. Flexible automated methods to parse and regroup similar documents in order to help humans analyze large corpora are thus required. Patent specifications present a formalized structure, with a number of parts that describe the patented invention and its context through descriptive writing (such as the title and abstract). It is then submitted to a patent office that will oversee the issuing of patent rights to the inventors and assign the patent to a number of IPC (International Patent Classification) classes it judges suitable. The IPC, however, is unfortunately not easily amenable to network analysis such as ours for several reasons [284, 306, 25]. First, it is a fixed classification that is infrequently updated even though patents are being published increasingly fast, creating a grey area where new technologies are not accurately categorized due to the lack of a sufficiently suitable class amongst the existing ones and making it unable to anticipate the birth of new fields [25]. Second, a patent is a complex document that contains different aspects such as technological descriptions or fields of application. Some patents thus exist at the intersection of several classifications (a patent describing a recommender system for targeted advertising, for instance, exists in both the recommender system space and the targeted advertising space), which the IPC resolves by allowing the patent to have multiple classes. This is, however, problematic when trying to quantitatively study patent data as multi-class analysis is markedly more complex than binary analysis. Third, the patent class taxonomy is massive with over 70 000 subgroups; a method to reduce the number of categories is necessary in order to obtain a smaller graph so that relevant and easily inter-

pretable analyses can be performed. Doing so by cutting the taxonomy closer to the root would be a potential way to proceed, but this would lose information as a patent can exist in different branches of the taxonomy. Finally, there is some inconsistency in the classification between different patent offices and countries [32, 62], where a patent will not necessarily have the same classification between the different offices and examiners.

In this context, we endeavor to develop a methodology that, simply based on the textual contents of the patents, regroups them following their overarching technologies. To do so, we draw methods from the topic modeling literature [1], a subfield in the vibrant field of natural language processing. In this chapter, we will first present in section 3.1 the various algorithmic bricks of the state-of-the-art topic modeling pipeline that is used and, in section 3.2, we will validate its results when applied to a *test corpus* small enough to be manually validated before applying this methodology to the much larger patent dataset in chapter 4. The *test corpus* is comprised of articles and conference proceedings from the field of bioinspiration and biomimetics, a highly interdisciplinary subset of the scientific literature. Indeed, scientific production represents a good testbed before we apply this methodology to patents : the documents show similar structure (title and abstract describing the contents of the document), are relatively similar in terms of technicality of the writing and can be retrieved easily (titles and abstracts are freely available online). The specific subfield of bioinspiration and biomimetics, in turn, was chosen for two reasons. First, its interdisciplinarity means that it broaches a number of varied technical topics, similarly to our patent database. Second, we possess some knowledge of this subfield allowing us to assess the quality of the resulting thematic clusters, which is necessary to validate this methodology before applying it to a larger corpus.

3.1 Topic Modeling

The analysis of a large corpus of text documents is challenging, often requiring a large amount of human resources. Generating a simplified representation of the corpus, however, can often be useful in order to facilitate information extraction, comprehension and analysis of the corpus. This is what topic modeling, an unsupervised machine learning technique that finds the overarching themes of the documents in the corpus and creates groups of thematically similar documents (*clusters*), aims to achieve. A number of techniques exist to perform topic modeling [1], but state-of-the-art performance has been achieved on topic modeling tasks using clustering algorithm on embedded documents [125]. Section 3.1.1 will describe the Transformer architecture, the deep learning model used to create high-dimensional document embeddings. Section 3.1.2 will describe the UMAP algorithm used to perform dimensionality reduction on these embeddings, *i.e.* create a low-dimensional representation of the document embedding vectors. Section 3.1.3 will describe the HDBSCAN algorithm used to extract document cluster memberships from these low-dimensional embeddings. Note that the dimensionality reduction step could be omitted since the clustering algorithm can work in an arbitrarily large space, but is performed as density-based clustering algorithms (such as HDBSCAN used here) tend to work better on low-dimensional datasets.

As the field of machine learning is constantly evolving at a rapid pace, with for instance

new model architectures or new algorithms (such as the Pairwise Controlled Manifold Approximation Projection [315] that aims to improve on the ability of UMAP to preserve both the global and local structure of the data), the individual steps of the pipeline described here are bound to improve, but the fundamental concepts involved in the process remain the same.

3.1.1 Transformers

Natural Language Processing (NLP) tasks have been revolutionized following the introduction of vectorization techniques such as Word2vec [217] in 2013. Vectorization techniques transform words and sentences into dense numerical representations called vector embeddings, allowing for the application of machine learning algorithms that take numerical representations as inputs.

The most popular and best-performing model architecture in dealing with a majority of NLP tasks is the Transformer architecture [304] of which variations gave birth to a number of pre-trained models such as OpenAI's Generative Pre-trained Transformers [242] (GPT) and Google's Bidirectional Encoder Representations from Transformers [89] (BERT).

Here, we will briefly discuss the transformer architecture and its self-attention mechanism that differentiates it from previous NLP deep learning architectures such as recurrent neural networks. The transformer architecture as described in [304] is a fairly straightforward encoder-decoder architecture, with modified encoder and decoder stacks. Throughout the course of this discussion, we will use "word" and "token" interchangeably for simplicity's sake as it is not central to conceptual understanding of the actions performed by the model. To give a brief rundown of the difference, when using transformer models, a word can be subdivided into several tokens(the word *eating* can be split into the its root form *eat* and the suffix *##ing*). This has several added benefits : it allows the model to capture the grammatical form of a word which contains information, and allows the model to deal with unknown words by splitting the word into known units.

Self-attention

The attention mechanism is a mechanism tasked with giving access to all elements of a sequence at each time step, while being selective and determining which sequence elements are most important depending on the context. In the Transformer architecture, the attention mechanism used is a specific type of attention called *self-attention*. Self-attention can be thought of as a mechanism that enhances the information content of an input embedding by including information about the input's context. This mechanism enables the model to weigh the importance of different elements in an input sequence and dynamically adjust their influence on the output. Here, we will describe the self-attention mechanism.

Given an input sequence of length T, each input word in the sequence is first embedded. Then, each of the input words receives three different representations corresponding to the roles it can play : the *query* component is used when a position in the sequence "looks"

at others in order to gather context, the *key* component is used when a position in the sequence is responding to a query's request, and the *value* component is used to modulate the amplitude of the response of the *key* component. This can roughly be put into plain words as "for the *queried* word in the sequence, what is the most related *key* (word) in the sentence to understand what the *queried* word is about, and how much *information* (value) does it contain ?". The self-attention mechanism utilizes three weight matrices $\mathbf{W}^Q$, $\mathbf{W}^K$ and $\mathbf{W}^V$ that are respectively used to project the inputs of the sequence into query, key and value components. These matrices are adjusted as model parameters via backpropagation during model training.

The query, key and value components are obtained via matrix multiplication between the $\mathbf{W}$ matrices and the embedded inputs $\mathbf{x}$.

$$\begin{aligned}\mathbf{q}^{(i)} &= \mathbf{W}^Q\mathbf{x}^{(i)} \text{ for } i \in [1, T] \\ \mathbf{k}^{(i)} &= \mathbf{W}^K\mathbf{x}^{(i)} \text{ for } i \in [1, T] \\ \mathbf{v}^{(i)} &= \mathbf{W}^V\mathbf{x}^{(i)} \text{ for } i \in [1, T]\end{aligned} \tag{3.1}$$

with $\mathbf{W}^Q$ and $\mathbf{W}^K_{d_k \times d}$ and $\mathbf{W}^V_{d_v \times d}$ where d_k is the dimension of vectors $\mathbf{q}^{(i)}$ and $\mathbf{k}^{(i)}$ and d is the size of word vector x. Index i refers to the index position in the input sequence. For instance, the query vector associated with the third input element is given by $\mathbf{q}^{(3)} = \mathbf{W}^Q\mathbf{x}^{(3)}$. Next, the unnormalized attention weights ω are computed using eq. 3.2 :

$$\omega_{i,j} = \mathbf{q}^{(i)^T}\mathbf{k}^{(j)} \tag{3.2}$$

$\omega_{i,j}$ corresponds to the unnormalized attention weight of the j-th input element for the query associated with input element i. The normalized attention scores are then given by eq. 3.3 :

$$\begin{aligned}\alpha_{i,j} &= \text{softmax}\left(\frac{\omega_{i,j}}{\sqrt{d_k}}\right) \\ \text{softmax}(\mathbf{x}_i) &= \frac{e^{\mathbf{x}_i}}{\sum_{j=0}^{j=k} e^{\mathbf{x}_j}}\end{aligned} \tag{3.3}$$

The scaling by $\sqrt{d_k}$ is used in order to ensure model convergence during training. Applying softmax heightens high values and depresses low values, drowning out words irrelevant to the words being attended to.

Finally, the context vector $\mathbf{z}^{(i)}$ (which is the attention-weighted version of the original query $\mathbf{x}^{(i)}$) is computed taking all the other input elements as its context following eq. 3.4.

$$\mathbf{z}^{(i)} = \sum_{j=1}^{T} \alpha_{i,j}\mathbf{v}^{(i)} \tag{3.4}$$

where $\mathbf{v}^{(i)}$ is the value vector computed by multiplying the $\mathbf{W}_v$ matrix with the embedded inputs $\mathbf{x}$ (see eq. 3.1).

Note that in practice, the attention score computation is done using matrix multiplication as shown in eq. 3.5.

$$\begin{aligned} \text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) &= \text{softmax}(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}})\mathbf{V} \\ \mathbf{Q} &= \mathbf{W}^Q\mathbf{x} \\ \mathbf{K} &= \mathbf{W}^K\mathbf{x} \\ \mathbf{V} &= \mathbf{W}^V\mathbf{x} \end{aligned} \tag{3.5}$$

where **Q**, **K** and **V** are respectively the query, key and value matrices of the sequence.

The implementation of this self-attention mechanism is performed through a module called *multi-headed attention*. This involves using h different attention heads, where a single attention head $l \in [1, h]$ is composed of the three matrices $\mathbf{W}^Q_{(l)}$, $\mathbf{W}^K_{(l)}$ and $\mathbf{W}^V_{(l)}$ and performs the computation of context vector $\mathbf{z}^{(l)}$. We perform the computation for each of the h attention heads yielding, for each input i, a total of h context vectors and query, key and value matrices *i.e.* we have context vectors $\mathbf{z}^{(i)}_1$, $\mathbf{z}^{(i)}_2$, ..., $\mathbf{z}^{(i)}_h$, and similarly matrices $\mathbf{W}^Q_{(1)}$, ..., $\mathbf{W}^Q_{(h)}$, $\mathbf{W}^K_{(1)}, ..., \mathbf{W}^K_{(h)}$ and $\mathbf{W}^V_{(1)}, ..., \mathbf{W}^V_{(h)}$. All the resulting context vectors are then concatenated, and multiplied by an additional weights matrix $\mathbf{W}_O$ that was trained jointly with the model in order to get the Z matrix that captures information from all attention heads as shown in eq. 3.6. The intuition behind the multi-head attention mechanism is relatively straightforward : when looking at a word in a sentence, humans pay attention to multiple things. For instance, when attending to a verb of motion, one can pay attention to the direction of the motion (where), the subject of the verb (who), or the means of locomotion (how). A single attention head would have to focus on all these related concepts, becoming silver at all trades and gold at none. Having multiple attention heads allows the model to pay specific attention to each of these concepts. The actions of the different heads have been shown to be interpretable and their importance quantifiable [311], with the main roles of attention heads being positional (several heads are "tasked" with attending to a token's immediate neighbors), syntactic (tracking major syntactic relations in the sentence) and dealing with rare tokens in the input sequence.

$$\begin{aligned} &\text{MultiHead}(Q, K, V) = \text{Concat}(head_1, ..., head_h)\mathbf{W}^O \\ &\text{with } head_l = \text{Attention}(\mathbf{Q}\mathbf{W}^Q_{(l)}, \mathbf{K}\mathbf{W}^K_{(l)}, \mathbf{V}\mathbf{W}^V_{(l)}) \end{aligned} \tag{3.6}$$

where the projections are parameter matrices $\mathbf{W}^Q_{(l)} \in \mathbb{R}^{d_{model} \times d_k}$, $\mathbf{W}^K_{(l)} \in \mathbb{R}^{d_{model} \times d_k}$, $\mathbf{W}^V_{(l)} \in \mathbb{R}^{d_{model} \times d_v}$ and $\mathbf{W}^O \in \mathbb{R}^{hd_v \times d_{model}}$. In the original transformer architecture, parameters are set with $h = 8$ the number of parallel attention layers and $d_k = d_v = d_{model}/h = 64$.

Note that the multi-headed attention mechanism is used in several different ways in the Transformer architecture.

- 1) In the "encoder-decoder attention" layers on the decoder side, queries (**Q**) come from the previous decoder layer, while keys (**K**) and values (**V**) come from the output of the encoder. Every position in the decoder is thus able to attend over all positions of the input sequence.

- 2) The self-attention layers in the encoder stack, where all keys, values and queries come from the output of the previous layer in the encoder. Each position in the encoder can thus attend to all positions in the previous layer of the encoder.
- 3) The self-attention sub-layers in the decoder stack allow each position in the decoder to attend to other positions up to the current word being decoded : in order to prevent leftward flow in the decoder, thus preserving the auto-regressive property (*i.e.* future values are only predicted based on past values), all values corresponding to illegal connections are masked out (*i.e.* "future" positions in the sequence) during the computation of the softmax step in the self-attention calculation (eq. 3.5).

Positional encoding

One of the weaknesses of the model the self-attention mechanism described here is its inability to take into account the order of words in the input sequence as they are all attended to in parallel. To remediate this, a vector is added to each input embedding following a specific pattern which can be either explicitly defined (for instance with a combination of trigonometric functions as is done in [89]) or learned during model training. This pattern helps the model estimate the position of each word, and thus the distance separating two words in the input sequence.

The Transformer architecture

The architecture of the Transformer model is shown in fig. 3.1.

The encoder stack is comprised of $N = 6$ identical encoder layers stacked on top of each other. An encoder layer is composed of two sub-layers (one self-attention layer and one feed-forward neural network) tasked with encoding the input to a high-dimensional continuous representation (usually of dimension 512 or 768) by taking into account the context of each item in the input sequence, thus improving the quality of the dense representation. This also helps the decoder side of the architecture focus on the appropriate items in the input sequence during the decoding process. Each sub-layer has a residual connection around it and is followed by a layer-normalization step. An input fed to an encoder layer goes through the following steps :

- A sequence is received as input.
- Words are vector-embedded (if in the first layer of the encoder) and concatenated with the vector resulting from positional encoding, resulting in positional embeddings *i.e.* embedded words with information about the order of words in the input sequence added.
- Self-Attention is computed on all words of the input sequence using dot-product attention, calculating the contribution of each token to each of the other tokens.
- The attention scores of each word in the input sequence are computed, and passed through the residual connection and normalization layers.

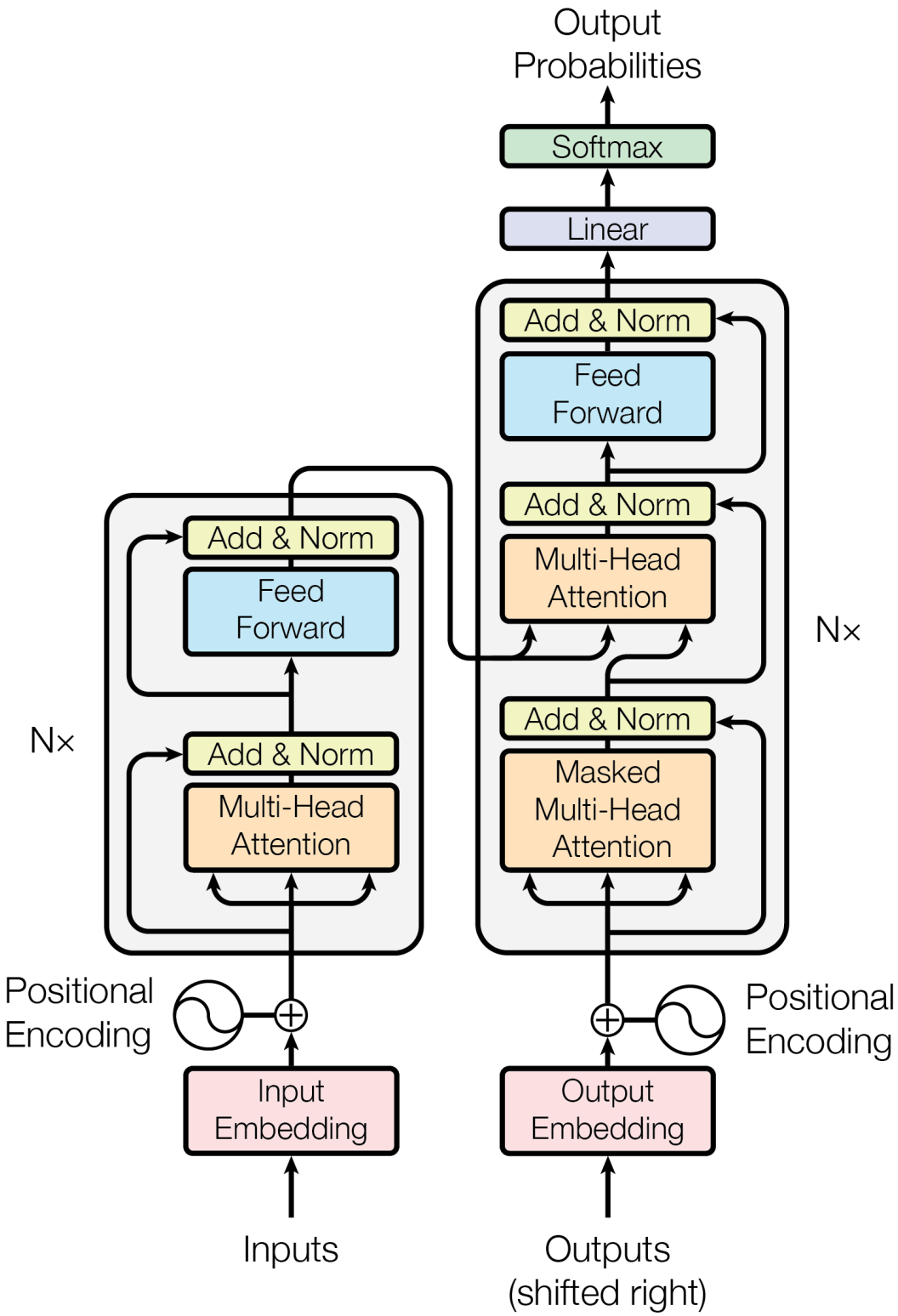

Figure 3.1: **The architecture of an encoder (left) and decoder (right) layer in the Transformer model.** Image taken from [304].

- The attention matrix is then passed through a feed-forward neural network composed of two linear transformations with a ReLU activation function in between as shown in eq. 3.7, and the result is once more fed through a residual connection and normalization layer.

$$\text{FFN}(x) = \max(0, xP_1 + b_1)P_2 + b_2 \tag{3.7}$$

The decoder stack is composed of $N = 6$ identical decoder layers. The decoder is tasked with generating text sequences for tasks such as sequence translation, query answering or text completion. The decoder takes as input the outputs of the previous decoder layer, as well as the outputs of the last encoder layer that contains the attention information. An input fed through a decoder layer goes through the following steps :

- The input goes through an embedding layer and position encoding layer in order to get positional embeddings.
- The positional embeddings are fed into the first masked multi-head attention layer to compute the attention scores for the decoder's input. They are then fed, as in the encoder, to a normalization and residual connection layer.
- The decoder layer has a second multi-head attention called the *encoder-decoder attention* that uses the **encoder stack**'s outputs as the keys and values and the queries from the previous layer in the decoder. This process matches the encoder's input to the decoder's input, giving the decoder access to context on the encoder's input. The output of this multi-headed attention is then passed through the residual connection and normalization layer.
- The result is then passed through a feed-forward neural network, another residual connection and normalization layer.

Finally, the output of the last decoder layer is passed through a linear layer that acts as a classifier, transforming it into a numerical vector that is the size of the vocabulary of the model. This vector is then fed into a softmax layer, turning it into a probability vector where each class has probability between 0 and 1. The index of the highest probability score corresponds to the predicted word, i.e. the next word in the generated text. The output is then added to the list of decoder inputs, and the process is repeated until the end of the generation is reached (for instance the number of desired tokens has been generated).

From Transformers to BERT

Two families of pre-trained Transformer models on large text datasets are commonly used. One is the BERT encoder-only architecture [89], and the other is the GPT decoder-only architecture [242]. In our use cases, we favored the BERT family of models : we want the best possible embedding of our documents to perform topic modeling, a task for which BERT models, being encoder-only, are better suited. The pre-training for BERT models is

performed differently compared to the usual sequence-to-sequence tasks through two unsupervised learning tasks. The first one is the "Masked Language Model", where a random input token is masked and the model has to predict the missing input, and the second one is the "Next Sentence Prediction", where the model is fed pairs of sentences and has to determine if the two sentences follow each other or not. The second pre-training task, in spite of its simplicity, is particularly beneficial when performing Question Answering or Natural Language Inference tasks, where understanding the relationship between two sentences is of particular importance.

The major finding and rationale behind BERT models is that it is possible to create state-of-the-art models for a variety of tasks by simply adding one additional output layer on top of the vector embeddings generated by the pre-trained encoder-only model, thus requiring comparatively low computational resources and time as the pre-trained model already has a general "understanding" of language instead of training new models from scratch for each task. One then simply only needs to fine-tune BERT models on the specific task at hand, rather than performing the training from scratch on every task.

3.1.2 UMAP

The UMAP algorithm (Uniform Manifold Approximation and Projection) algorithm [211] is a dimensionality reduction technique used to reduce the dimensionality of a dataset. Dimensionality reduction is a machine learning technique that is used to project data from an n-dimensional space to an m-dimensional space, with $m \leq n$ while retaining as much of the high-dimensional structure as possible in the low-dimensional space. Dimension reduction is used in multiple contexts, notably for data visualization (reducing high-dimensional datasets to 2 or 3-dimensional datasets for graphical representation) and for pre-processing in machine learning pipelines where the end algorithms (such as distance-based classification) work better on low-dimensional spaces due to the curse of dimensionality [154]. The curse of dimensionality is used to refer to a number of problems arising when working with high-dimensional data, and can be partly mitigated by performing dimensionality reduction on the data. As information is necessarily lost by reducing the dimensionality of the problem, choosing a suitable dimensionality reduction algorithm is of the utmost importance. A number of candidate algorithms exist, each with their own pros and cons, with two of the most famous being Principal Component Analysis (PCA) [323] and t-distributed Stochastic Neighbor Embedding (t-SNE) [193]. PCA, for instance, is non-parametric and tends to capture the global structure at the cost of local similarities, but is strictly linear, which is a strong limitation when dealing with complex data. t-SNE, while non-linear and able to capture local similarities in the data, tends to do so at the cost of the global structure, is parametric, and does not scale well with very large datasets.

UMAP endeavors to improve upon these methods by offering a non-linear, highly scalable algorithm able to model local structure while preserving more global structure than t-SNE. Here, following [211], we will briefly describe how UMAP achieves this.

UMAP hinges on the assumption that the data is uniformly distributed on the manifold. This is, of course, not the case in reality, and so the solution proposed is to assume that the

notion of distance varies across the manifold, giving each point its local notion of distance using Riemaniann geometry. This is done using the k-nearest neighbors of each point to estimate its local distance function. For small k values, the fine detail structures and variations of the Riemannian metric are thus more closely captured; for large k values, larger regions are taken into account when computing the metrics, making them more accurate across the manifold. UMAP thus falls in the class of k-neighbour based graph algorithms, and works in two phases. First, a weighted k-neighbour graph is computed in order to capture the local structure of the high-dimensional data. Then, a low-dimensional layout of this k-neighbour graph is computed and optimized to be as structurally similar to the high-dimensional graph as possible.

Computing the weighted k-neighbour graph

Due to the assumption that each point i in the dataset has its own local metric, we can meaningfully measure distance, allowing us to work in a fuzzy topology where, rather than binarily determining whether a point is in the k-nearest neighborhood of i, we can determine *how far* it is from i. The pairwise similarity $v_{j|i}$ of point i with each of its k approximate nearest neighbors is given in eq. 3.8 :

$$v_{j|i} = \begin{cases} \exp\{-\max(0, d(x_i,\ x_j) - \rho_i)/\sigma_i\} & \text{if } 1 \leq j \leq k \\ 0 & \text{otherwise} \end{cases} \tag{3.8}$$

with ρ and σ defined as follows :

$$\begin{aligned} \rho_i &= \min\{d(x_i, x_j) \mid 1 \leq j \leq k,\ d(x_i, x_j) > 0\} \\ \sum_{j=1}^{k} \exp\left(\frac{-\max(0,\ d(x_i, x_j) - \rho_i)}{\sigma_i}\right) &= \log_2(k) \end{aligned} \tag{3.9}$$

$d(x_i, x_j)$ is the distance between points i and j (not restricted to Euclidean distance, it can take any form), ρ_i is the distance to the nearest neighbor of i, and σ_i is a normalizing factor based on the neighborhood of i. An additional constraint is applied to ensure that each point has at least 1 connection to another point in the dataset (introduced by the term ρ_i in eq. 3.9); the fuzzy confidence thus decays in terms of distance *beyond* the first nearest neighbor. This results in a directed weighted graph, where the edge weight from point i to point j is given by $v_{j|i}$ and the edge weight from point j to point i is given by $v_{i|j}$. Since the nearest-neighborhoods of i and j (and thus the associated values of $v_{j|i}$ and $v_{i|j}$) are different *i.e.* the local distance metrics of each point are not guaranteed to be compatible, symmetrization is performed following eq. 3.10.

$$v_{ij} = (v_{j|i} + v_{i|j}) - v_{j|i}v_{i|j} \tag{3.10}$$

This merges the conflicting weights together, resulting in an undirected weighted graph where the weight of the edge between points i and j is given by v_{ij}, which can be interpreted as a Bernoulli variable where v_{ij} denotes the probability of existence of the edge.

Lowering the dimension

The question, then, simply becomes how to find a lower-dimensional representation of this weighted graph with a fuzzy topological structure as similar to the one previously determined as possible. To do so, the approach used is similar to the one described above, with an important distinction : as the manifold we are trying to embed the data into is simply a low-dimensional euclidean space, we can simply use the euclidean norm to measure the distance w_{ij} between two points on this new graph, as described in eq. 3.11.

$$w_{ij} = \left(1 + a\|y_i - y_j\|_2^{2b}\right)^{-1} \tag{3.11}$$

where $\|.\|_2$ is the euclidean norm, and a and b are hyper-parameters.

Then, once a candidate low-dimensional representation is computed, we need to be able to compare it to the fuzzy topological structure of the high-dimensional representation (the low-dimensional representation can be randomly initialized but, in practice, spectral embedding techniques are used to initialize it into a good state, allowing for both faster convergence and for greater stability when searching for the optimal solution). Given such a measure, finding an optimal low-dimensional representation becomes a fairly straightforward optimization problem and, since both v_{ij} and w_{ij} can be thought of as Bernoulli variables, the cross-entropy loss function shown in eq. 3.12 is a suitable objective function for this optimization problem.

$$C_{UMAP} = \sum_{i \neq j} v_{ij} \log\left(\frac{v_{ij}}{w_{ij}}\right) + (1 - v_{ij}) \log\left(\frac{1 - v_{ij}}{1 - wij}\right) \tag{3.12}$$

The minimization of this function can be seen as a kind of force-directed graph layout algorithm, with the first term $v_{ij} \log\left(v_{ij}/w_{ij}\right)$ providing an attractive force between points i and j when v_{ij} is dominant and the second term $(1 - v_{ij}) \log \frac{1-v_{ij}}{1-wij}$ providing a repulsive force between points i and j when v_{ij} is small. This optimization is performed using a stochastic gradient descent algorithm in order to find the low-dimensional representation minimizing the cross-entropy loss function, *i.e.* the low-dimensional graph whose fuzzy topological representation matches the closest to that of the original high-dimensional graph.

3.1.3 HDBSCAN

Hierarchical Density Based Clustering of Applications with Noise (HDBSCAN) [56] is a density-based clustering algorithm. Clustering algorithms are unsupervised learning algorithms tasked with finding distinct groups in data (clusters). Density-based clustering algorithms are a subclass of clustering algorithms that identify distinct clusters in spatialized data, based on the assumption that similar data points are in high-density areas of the space (the clusters), and that the various clusters are separated by low-density regions. HDBSCAN generates a density-based clustering hierarchy and flattens it, extracting only

the most significant clusters. This is done, once the clustering hierarchy has been computed, by considering the task of extracting the set of significant clusters as an optimization problem with the objective of maximizing the overall stability of the composing clusters.

In order to do so, the HDBSCAN algorithm works in several phases as described in [210].

Building the hierarchy of connected components

First, the minimum spanning tree of the distance-weighted graph needs to be built. The minimum spanning tree is defined as the subset of edges of the graph that connects all vertices together without allowing for any cycles and minimizing the total edge weight. To begin with, points with low density are spread apart, in order to maximize the efficiency of the single linkage algorithm used to build the minimum spanning tree. To do so, two distances are defined : the core distance defined for a point x as the distance to its k-th nearest neighbor and a mutual reachability distance between points a and b defined in eq.3.13.

$$d_{mreach-k}(a,b) = \max\left(core_k(a),\ core_k(b),\ d(a,b)\right) \tag{3.13}$$

with $d(a,b)$ the metric distance between points a and b.

Dense points (*i.e.* with low core distance) will remain the same distance from each other, but points with high core distance are isolated to be at least their core distance from any other point.

A weighted graph is then built by using data point as vertices and the mutual reachability distance of all pairs of points as edge weights. A threshold value t on the edge weights is then set and gradually lowered, and edges with weight above that threshold are removed from the graph. As the threshold is lowered, the graph will break into more and more connected components, yielding a hierarchy of connected components (clusters). This is computationally expensive for large datasets in the graph as the number of edges grows as n^2, with n being the number of vertices in the graph. To remediate that, using Prim's algorithm, the minimum spanning tree is built from the ground-up : the tree is built one edge at a time by adding the lowest weight edge that connects the current tree to an isolated vertex. Once this minimum spanning tree is built, the hierarchy of connected components is then created by sorting all edges of the tree by distance in increasing order and then iterating through, creating a new merged cluster at each step.

Condensing the hierarchy of connected components

This cluster hierarchy, however, is difficult to analyze since there are as many splits as there are edges, and so we need to find a suitable way to obtain a set of flat clusters from this hierarchy of clusters with variable densities. Using a **minimum cluster size** hyper-parameter, the hierarchy is then traveled. For each split in the hierarchy, both sides of the split are considered as true clusters if they are above the minimum cluster size and, if one side of the split is smaller than the minimum cluster size, only the larger cluster is considered a true

cluster. This thus reduces the complex hierarchy to a much more manageable one that only contains true clusters and, at the same time, adds information on cluster persistence as a function of distance (due to splits happening as the threshold on mutual reachability distance is lowered).

Selecting stable clusters

This persistence information allows for the quantitative selection of the clusters that persist over large distance spans. Indeed, short-lived clusters tend to be artifacts of the process leading to the construction of the minimum spanning tree, and are thus more likely to be noise. To perform this distinction, a new measure $\lambda = \frac{1}{t}$ is defined to consider the persistence of clusters as a function of threshold t. Each cluster has associated values λ_{birth} and λ_{death}, respectively corresponding to the λ value when the cluster split off to become its own cluster and to the λ value when the cluster split into smaller clusters (if applicable). For a given cluster, each of its point p also has a λ_p value that corresponds to the λ value at which the point splits from the cluster, with $\lambda_{birth} \leq \lambda_p \leq \lambda_{death}$ as the point either splits from the cluster or leaves the cluster when it is split into two smaller clusters. The cluster stability $s(c)$ is then computed for each cluster following eq. 3.14.

$$\begin{aligned} s(c) &= \sum_{p \in c} (\lambda_p - \lambda_{birth}) \\ s_d(c) &= \sum_{b \in \text{children}(c)} s(b) \end{aligned} \tag{3.14}$$

Finally, all leaf nodes (*i.e.* the true clusters at the lowest levels of the hierarchy) are declared to be selected clusters. By working up through the hierarchy (the reverse topological sort order of the graph), we compare the stability of each cluster $s(c)$ with the stability of its descendants ($s_d(c)$ in eq. 3.14). If $s(c) < s_d(c)$, $s(c) \leftarrow s_d(c)$ and the process continues; if $s(c) > s_d(c)$, the cluster is selected and all its descendants are unselected. This process is repeated until the root node is reached, and the flat clustering is formed by all selected clusters after the root node has been reached. Performing cluster allocation is then straightforward, as all points in selected clusters are allocated to the cluster in question and all points in non-selected clusters are considered noise points and are unlabeled.

3.2 Testing the topic modeling pipeline

The Living Machines conference is targeted at the intersection of research on new technologies inspired by the scientific investigation of biological systems (biomimetics) and research that seeks to interface biological and artificial systems (biohybrid systems). We seek to highlight the most exciting international research in both of these fields united by the theme of “Living Machines". The most recent conference was dedicated to reflecting on how the field of Living Machines has evolved over the last 10 years and how it will

progress in the next 10 years. Leaders in the field presented their perspectives on the last 10 years of bio-inspired locomotion, bio-inspired & biomimetic soft robotics, bio-hybrid robotics, invertebrate robotics, plant robotics and neuro-robotics. They also highlighted the current challenges, unanswered questions and provided their predictions for the future of these research fields. This Bioinspiration and Biomimetics special issue seeks to compile these perspectives and provide a snapshot of the current state of the field, highlight open questions and look ahead to the future.

The early detection of emerging scientific and technological trends has been an important topic of study [3] that is becoming more and more relevant with the exponential growth of scientific and technological production. As the financial means and workforce are limited, choices must be made in allocating resources towards the development of some technologies and scientific domains at the unfortunate expense of others. The decisions in these two domains, however, are not independent: indeed, the interplay between industrial and academic research is a well-identified and desirable phenomenon [305]. This interaction is seen as mutually beneficial as universities can gain access to greater financial resources compared to the more traditional government grants and companies can get access to scientific expertise in select domains. Both sides of the interaction, however, still elude quantitative, large-scale characterization. This characterization is necessary in order to make full use of the links between academia and industry, in particular through the early detection of the rising trends in scientific and technological production as they start to form and, conversely, the detection of other fields where such growth is not present but could be thought of as desirable. So far, most endeavors in that direction were made possible using expert opinions of select individuals. This has now become - and has in reality been for some time - impossible. Indeed, current research in science and technology produces a very large amount of scientific articles and patents. This booming production, however, comes at a cost: there is too much to read. This is particularly true in an age where interdisciplinary research is becoming increasingly important and requires researchers to be aware of the state-of-the-art of various fields. Keeping track of this massive ebb and flow is currently hardly feasible as it requires too much expertise and time. A large part of the matter, then, becomes one of scientific surveillance: researchers need to be helped with staying afloat in an accelerating ever-rising sea of scientific and technological production by having access to instruments aiding in the continuous acquisition of the knowledge crucial to their work [91]. As the challenges that we face are getting more and more complex, the various tools used to tackle them also need to grow in sophistication. Modern problems require modern solutions [167].

Here we propose to develop artificial intelligence techniques to automate scientific and technological surveillance. This method aims to provide valuable assistance to researchers and scientific leaders to map and analyze the state of the art and to detect emerging and underdeveloped topics. We propose a step towards developing a method for automatic scientific surveillance in order to automate the early detection of scientific and technical trends.

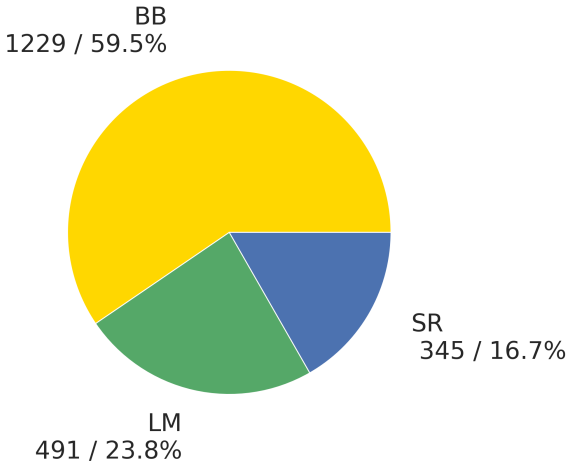

Figure 3.2: **Composition and number of articles (with outliers filtered out) in the merged corpus (ME).** LM = “Conference on Biomimetic and Biohybrid Systems" often referred to as Living Machines, BB = “Bioinspiration & Biomimetics", SR = “Soft Robotics".

Goals

This chapter proposes a method to provide more than a snapshot of the current state of the field : new technologies inspired by the scientific investigation of biological systems (biomimetics) and research that seeks to interface biological and artificial systems (biohybrid systems). Here, we focus on analyzing the research themes of the Living Machines conferences (LM) by taking into account all the articles published so far. As the special issue is published in this journal we then analyze the research themes published in this journal also taking into account all published articles to date. We compare both research themes sets and analyze the intersection between the Living Machines conferences (LM) and the journal Bioinspiration & Biomimetics (BB). Finally, since soft robotics is a rather new and emerging trend in this field we add to our study all the articles published in the new journal Soft Robotics (SR). We again analyze the intersection between these three corpora and we give a global view of the merged corpus (ME) including all the 2099 articles published to date in the Living Machines (LM) conferences and the journals Soft Robotics (SR) and Bioinspiration & Biomimetics (BB).

We then highlight open questions and look to the future or underrepresented topics.

3.2.1 Methodology

The method used to automatically read scientific corpora and cluster similar articles together based on their topic was presented in section 3.1. Four corpora on which analyses will be performed are presented in this chapter :

- the Living Machines conference proceedings (LM) between 2012 and 2020 (494 articles)

- the Bioinspiration & Biomimetics journal issues (BB) (1260 articles)
- the Soft Robotics journal issues (SR) (345 articles)
- the Merged corpus (ME) that is built by concatenating the 3 other corpora presented above i.e. 2099 articles (Fig. 3.2).

Figure 3.2 shows the proportion of each corpus in their merged corpus.

We will briefly summarize the topic modeling pipeline applied in the context of our academic corpora. The algorithm reads the title and abstract of each article to create a latent representation of the topics studied in the article. Articles discussing similar topics are located closer in the latent space. The embedding of each article is performed using SciBERT [24], a BERT model [89] specifically trained on scientific literature.

Each token is embedded in a 768-dimensional embedding space through the application of this model. We reduce the dimension of this space to a projected latent space of low dimension using the UMAP algorithm [211]; the selected dimension varies following the data-set used and is selected by manually choosing the dimension which maximizes the human interpretability of the results and minimizes the number of unlabeled articles by the clustering algorithm presented below. This dimensionality reduction step condenses the information and increases the performance of the HDBSCAN community detection algorithm [210]. This dimensionality reduction also provides the added benefits of significantly reducing memory requirements and subsequent computation time, which turn out to be particularly important when working with much larger amounts of data such as complete corpora of scientific articles. We note that the results presented are qualitatively valid for higher dimensional UMAP projections.

The projected latent representations of all articles is used in order to filter outliers in the corpus, i.e. articles that contain no valuable information (for instance empty abstracts and/or titles, errata and corrigenda). To do so, the radius of the complete corpus in latent space is computed with 2 different methods depending on the dimensionality of the projected latent space. Indeed, as convex hull algorithms run in roughly $\mathcal{O}(n^{d/2})$ time [66] with d the dimension of the latent space, it is suitable for lower dimensions but quickly becomes unsuitable for higher dimensions. For dimensions 5 and lower, we compute the convex hull of all points and measure the pairwise euclidean distance of all points on the hull which corresponds to the radius of the corpus. For dimensions greater than 5, we simply compute the pairwise euclidean distance of all points in the corpus and take the maximum pairwise distance which works for relatively small corpora but is not suitable for larger corpora. Using this radius, we compute for each article its mean distance to the 5% closest articles in the corpus using a KNN model and consider the article to be an outlier if that mean distance is greater than a third of the radius; it is then removed from the corpus. As some groups of legitimate articles can be localized relatively far away from the centroid of the corpus in the latent space, this methodology allows to discriminate between these articles (specific but relevant to our analysis) and true outliers.

Using this latent space projection, the HDBSCAN [210] algorithm is applied to our low-dimensional vectorial representations to perform unsupervised clustering in latent space,

automatically grouping items based on their main themes. HDBSCAN can consider points of the dataset as *noise*; these points will be referred to as *unlabeled*.

To find the vocabulary specific to each cluster, we use a model called class-based TF-IDF (c-TF-IDF) as defined in Eq. 3.15. This model consists in applying the standard TF-IDF method [243] treating each cluster as a single document to retrieve the top n-grams with the highest values as computed in eq. 3.15, where our n-grams here are defined as sequences of *n* adjacent words from a given text. As TF-IDF reflects how important an n-gram is to a document in a corpus, c-TF-IDF is well-suited to our use case as it reflects how important an n-gram is to a cluster in a collection of clusters. It is calculated with the following equation:

$$x_{w,c} = \frac{w_c}{A_c} \times \log \frac{m}{\sum_{j=0}^{n} t_j} \tag{3.15}$$

where $x_{w,c}$ represents the importance of n-gram w within class c, w_c the number of occurrences of ngram w in class c, A_c is the total number of n-grams in class c, m is the number of documents in the sample, n is the number of different classes and t_j is the frequency of n-gram t across all classes.

We then label each of the resulting groupings by their research theme using both this cluster-specific vocabulary and by looking at the individual articles in the various clusters.

The preprocessing and analysis scripts are written in Python 3.8, the BERT models are provided by the Hugging Face framework [324] and their implementation is realized using PyTorch [233].

The average similarity between clusters is computed following equations 3.16 and 3.17:

$$s(p_k, p_l) = 1 - \frac{||p_k - p_l||_2}{\max_{k,l}(||p_k - p_l||_2)} \tag{3.16}$$

$$S(c_i, c_j) = \frac{\sum_{k=0}^{n_i} \sum_{l=0}^{n_j} s(p_k, p_l)}{n_i n_j} \tag{3.17}$$

where $S(c_i, c_i)$ denotes the final similarity between clusters i and j, n_i the number of articles in cluster i, p_k the latent space projection of the $k-th$ article in cluster i with $k \in [0, n_i]$ and $s(p_k, p_l)$ the euclidean similarity between articles p_k and p_l. Since we divide each row by of the matrix by its maximum value, this similarity measure is asymmetrical (i.e. similarity between cluster 1 and 2 is not the same as similarity between cluster 2 and 1). The euclidean distance is suitable to measure similarity between two articles as we are working in a low-dimensional version of the latent space.

$$MI(U, V) = \sum_{i=1}^{|U|} \sum_{j=1}^{|V|} \frac{|U_i \cap V_j|}{N} \log \frac{N|U_i \cap V_j|}{|U_i||V_j|} \tag{3.18}$$

$$AMI(U, V) = \frac{MI(U, V) - E(MI(U, V))}{\langle H(U), H(V) \rangle - E(MI(U, V))} \tag{3.19}$$

Dataset	UMAP		HDBSCAN		
	n_components	n_neighbors	min_samples	cluster_epsilon	min_cluster_size
LM	5	3	1	0.3	25
BB	10	5	1	0.45	40
SR	5	3	1	0.7	15
ME	10	3	1	0.45	40

Table 3.1: **Hyperparameters used for the UMAP and HDBSCAN algorithms for the various datasets.** Other hyperparameters of the algorithms that are not shown here use the default values of the UMAP and HDBSCAN python packages.

where U (resp. V) is a label assignment of N items with i (resp. j) separate classes, H their entropy, E the expected value of the mutual information between the two partitions as defined in [309].

The hyperparameters used to get the results in this chapter are described in Table 3.1.

3.2.2 Results

We use our NLP-based method to automatically extract research themes and their trends in all articles published in LM, BB and SR. Doing this detection and classification manually would involve reading the title and abstract of the 2099 articles and then classifying them. Then, when the themes are detected and classified we perform a comparison between the 3 corpora (LM, BB and SR).

Thematic clusters extraction and analysis

Figure 3.3 shows, for each corpus, individual articles in a 2-dimensional projection of the latent space. The number of clusters varies between 7 clusters in the SR and BB corpora and 8 clusters in the LM and ME corpora (without counting the unlabeled clusters). This low-dimensional projection retains some local structure so that two articles close in this latent space are similar in topic; some information is however inevitably lost as we go from a 768-dimensional to a 2-dimensional vectoral representation, thus resulting in some articles being visually far from the rest of their clusters in the 2-dimensional representations of our corpora. Each cluster (see Fig.3.3) is then named manually by inspecting its top 5 n-grams as computed following equation 3.15 (see Table 3.3 for the list of 1-gram and 2-grams for each cluster) and by checking a few articles belonging to the cluster.

Words written in *italics* from here on will denote specific cluster topics, prefixes will denote clusters in a given corpus and words written in **bold** will denote specific n-grams (i.e. SR.5 designates cluster 3 in the Soft Robotics corpus, its articles are about *Materials* and **fabrication** and **soft material** are two of its representative n-grams).

Figure 3.4 shows the proportion of each cluster in their corresponding corpus and the

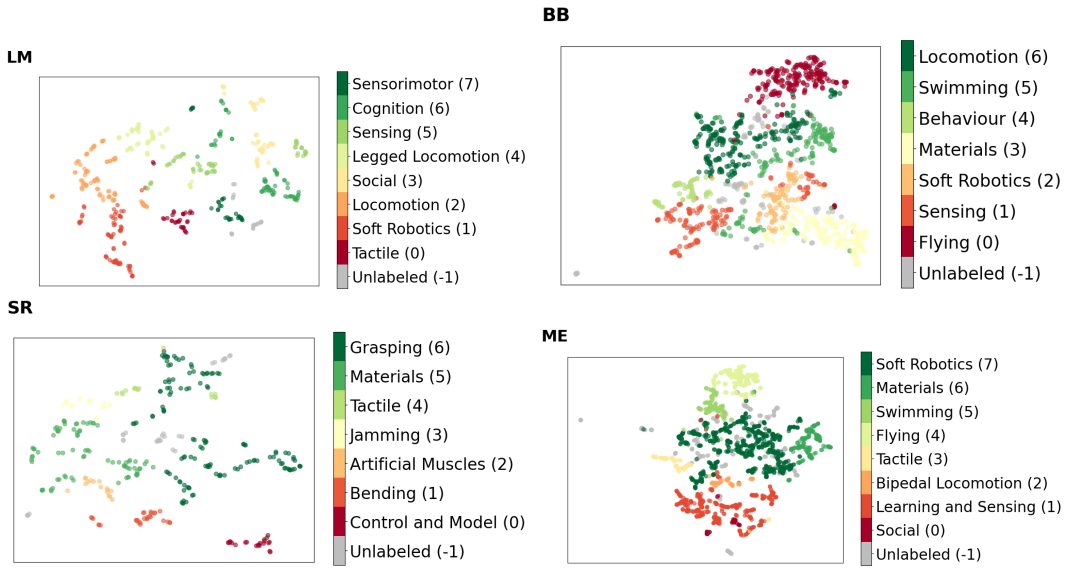

Figure 3.3: **Clustering of research themes by natural language processing (NLP).** The latent representations of articles are computed using the SciBERT model and projected in a two-dimensional space for the various corpora. The topic labels and their associated colors are assigned after applying the HDBSCAN clustering algorithm on a 5 or 10-dimensional projection (depending on the corpus) of the SciBERT latent representations. Dimensionality reductions are performed with UMAP. Naming the clusters is done manually by inspecting the n-grams and checking some articles belonging to the cluster.

	Cluster -1	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7
LM	Unlabeled	Tactile	Soft Robotics	Locomotion	Social	Legged Locomotion	Sensing	Cognition	Sensorimotor
BB	Unlabeled	Flying	Sensing	Soft Robotics	Materials	Behavior	Swimming	Locomotion	N/A
SR	Unlabeled	Control and Model	Bending	Artificial Muscles	Jamming	Tactile	Materials	Grasping	N/A
ME	Unlabeled	Social	Learning and Sensing	Bipedal Locomotion	Tactile	Flying	Swimming	Materials	Soft Robotics

Table 3.2: **Cluster labels for each corpus.** Cluster -1 for each corpus corresponds to the articles that were unable to be labeled by the algorithm.

associated names. We see that the *Social* and *Cognition* clusters represent roughly a quarter of the LM corpus, and are topics that are not as prevalent in the other 2 corpora, with BB having a small *Behavior* cluster (only 5% of the corpus). *Locomotion*-related clusters represent roughly 28% of the LM corpus (*Locomotion* and *Legged Locomotion*) and 49% of the BB corpus (*Flying*, *Swimming* and *Locomotion*), whereas there is no clearly defined *Locomotion*-related cluster in the SR corpus. The LM corpus has three clusters dedicated to *Perception*-related topics (*Tactile*, *Sensing* and *Sensorimotor*) which combined represent roughly 26% of the corpus, whereas this topic only represents 12% of the BB corpus (*Sensing*) and is weakly representend in the SR corpus (5% with the *Tactile* cluster). *Soft Robotics* articles represent 13% of the LM corpus. *Manufacturing* topics represent 17% of the BB corpus (*Materials*) and 24% of the SR corpus (*Materials*). A general *Soft Robotics* cluster represents 15% of articles in the LM corpus and roughly 8% of articles in the BB corpus, but is not present in the SR corpus. Indeed, as SR explicitly contains Soft Robotics-related articles, it is more homogeneous that the LM and BB corpora and its clustering is thus more finely-grained; this is why the clustering of the SR corpus yields specific topics such as *Artificial Muscles*,

Bending, *Jamming* and *Grasping*.

This behavior can be expected as the granularity of the clusters given by our algorithm depends on the heterogeneity of the base corpus, with homogeneous corpora yielding more specific clusters and heterogeneous corpora yielding more general clusters (i.e. clustering a corpus of general scientific articles will cluster articles following large domains such as physics, social sciences, economy, etc whereas clustering a more specific domain such as physics would yield subfields of physics i.e. condensed matter, astrophysics, biophysics, etc). The clustering of the SR corpus, being a subfield of bioinspiration, thus yields subdomains of the field of Soft Robotics rather than a more general *Soft Robotics* cluster as is for instance the case with the LM corpus that deals with more varied themes.

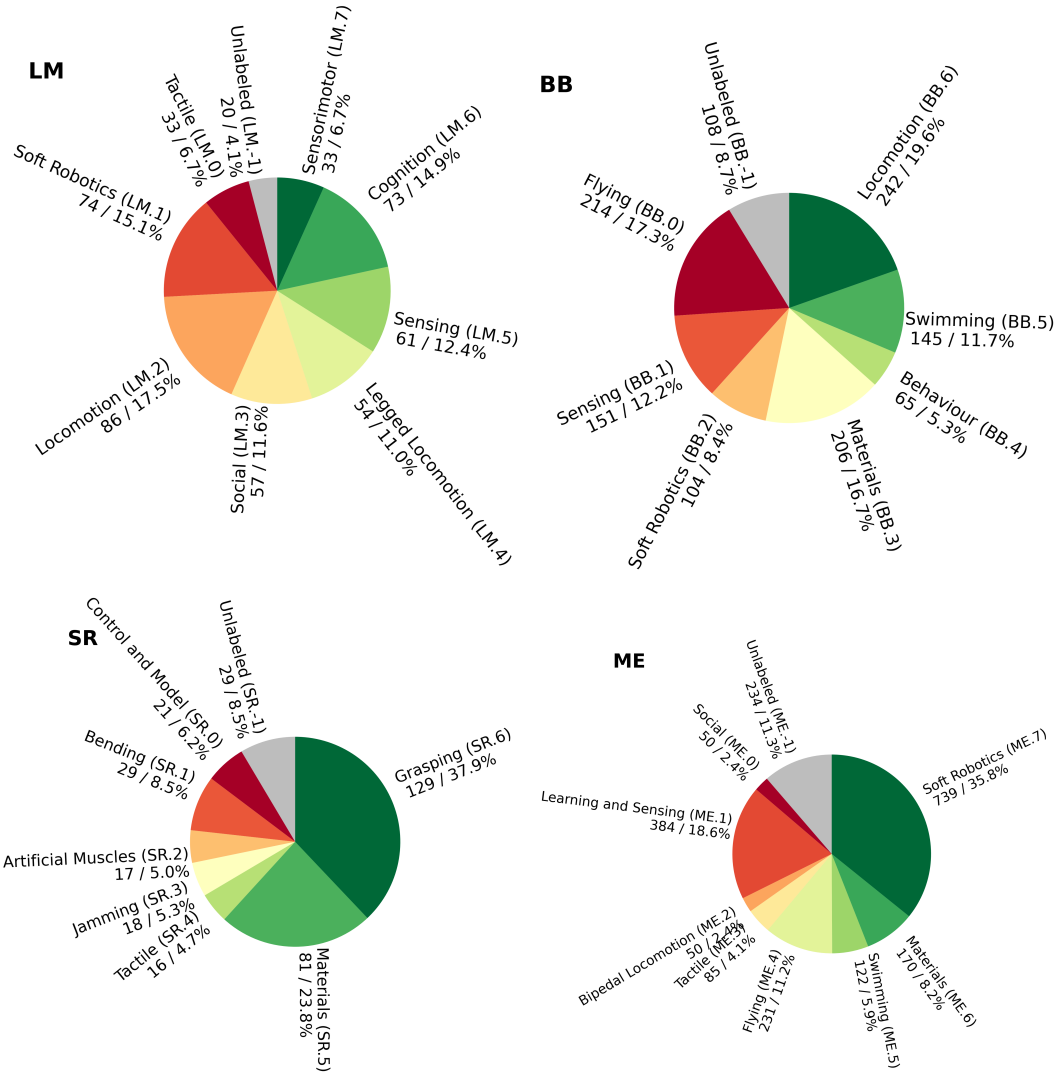

Figure 3.4: **Number and proportion of articles in each cluster for the various corpora.** Cluster labels are given by the authors by looking at the top words in Table 3.3 for each cluster and then manually checking articles in each cluster.

Figure 3.5 shows a flowchart that groups and classifies the topics found according to the basic principles for designing and building a robot. The LM and SR corpora contain mainly

robotics while the BB corpus contains more general themes in biomimicry. It should be noted that some robotics themes join general biomimicry issues such as the question of materials.

Most of the SR clusters concern various research related to the actuation of robots. As the field of soft robotics is new and based on new soft materials the question of actuation of these new types of robots is essential. It should be noted that the majority of clusters (4 from SR, 3 from BB, 2 from LM) focus on robot actuation which remains a central theme in robotics.

The LM corpus covers the widest range of topics concerning robot design, namely control, activation, sensing and social robotics which includes collective robotics, social interactions with animals and humans.

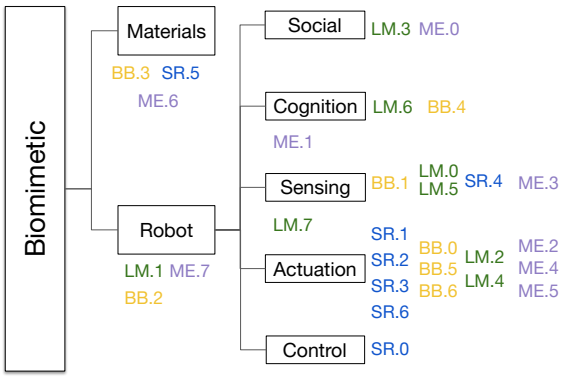

Figure 3.5: **Architecture of all clusters ordered like a robot design.** By grouping the cluster, it shows which general theme gather the research efforts. LM.1, BB.2 and ME.7 in the “Robot" category are gathering all soft robotics in one cluster. The “Social" category contains both human-robots interaction and collective robotics. Most clusters pertain to the study of bio-inspired actuators. ME.1 is both related to “Cognition" and “Sensing". LM.7 is both related to “Sensing" and “Actuation".

Research backgrounds

In the LM corpus we find eight meaningful clusters.

LM.0 (*Tactile*) contains articles dealing with the design of elements of the tactile system such as [80] describing the design and higher performance of a biomimetic fingertip (fingerprint) compared with a smooth fingertip, [14] presenting the anthropomorphic design and gripping performance of a robotic finger and [219] introducing the design and performance of a multi-element sensory array based on the mammalian whisker sensory system.

LM.1 (*Soft Robotics*) contains articles in the field of Soft Robotics. Webster *et al.* [318] present *Aplysia Californica* (a species of sea slug) and its potential as a source of actuator and scaffold material for biohybrid robots. Follador *et al.*[107] present a biological investigation of octopus suckers in order to determine specifications for the design of artificial suction

cups. Tonazzini *et al.*[296] present soil penetration strategies of plant roots in order to inspire the design of explorer robots.

LM.2 (*Locomotion*) contains articles dealing with general locomotion, such as [54] with its six-limb octopus-inspired robot achieving effective locomotion through the coordination of its various limbs, [258] presenting a mathematical model of the crawling mechanism in larval *Drosophila* accounting for key anatomical features of the larva and [286] describing the spontaneous transitions between various types of gaits using a quadruped robot model with a head segment and a postural reflex mechanism.

LM.3 (*Social*) contains articles presenting social interaction involving robots. Mazzei *et al.*[206] present the design and implementation of an hybrid cognitive architecture controlling the reaction and facial expressions of a social humanoid robot during basic social interaction tasks. Fernando *et al.*[102] present the Expressive Agents for Symbiotic Education and Learning (EASEL) project, which is tasked with exploring and understanding human-robot symbiotic interactions. Lazzeri *et al.*[178] describe the authors' research on the impact of appearance and behavior on the design of a believable social robot.

LM.4 (*Legged Locomotion*) contains articles presenting locomotion and control architectures that specifically contain legs. Steele *et al.*[280] present the design and development of a bio-inspired knee joint mechanism, Schneider *et al.*[264] present the design and test results of the hexapod robot HECTOR built using embedded, custom designed and compliant joint drives and Szczencinski *et al.*[287] analyzes how central pattern generators can entrain joints of the MantisBot to positive velocity feedback to successfully implement active reaction during walking.

LM.5 (*Sensing*) contains articles dealing with the design of novel sensors and the integration of sensory data, such as [320] presenting a model for the integration of sensory data in the design of a neural network tasked with controlling flight in a robot simulating honeybee foraging, [41] presenting the design of a novel sensor inspired from the electric fish in order to tackle underwater exploration of objects and [288] that describes the design and the comparison with experimental data of a dynamical model describing the adaptive response of sensory organs on insect legs tasked with detecting cuticular strain.

LM.6 (*Cognition*) contains articles describing the design of various control architectures controlling robot behavior. Renaudo *et al.*[247] presents the design and results of a hybrid control architecture tasked with a simple habit learning task based on the coordination of model-based and model-free reinforcement learning. Terekhov *et al.*[289] presents the design and performance of a block-modular neural network architecture allowing parts of the existing network to be re-used to solve novel tasks while retaining performance on the original task. Ognibene *et al.*[230] presents theoretical insights into the unveiling of hidden information through epistemic actions and the experimental benefits of using this actively-gathered information in order to efficiently accomplish a seek-and-reach task.

LM.7 (*Sensorimotor*) contains articles dealing with the integration of sensory information into robotic behavior, such as [147] presenting a closed-loop control architecture using visual information to control the movement of a robot which in turn generates optic flow, [303] presenting a complete model based on the coordination of two main classes of reflexes in order to stabilize the human gaze and its results in performing various locomotion tasks

and [221] presenting a control algorithm for the integration of different biological neural models of eye movements in order to design a biologically plausible ocular neurocontroller.

In the BB corpus we find seven meaningful clusters.

BB.0 (*Flying*) contains articles dealing with the aerodynamics and flight of artificial flyers. In [223], the authors present a humming-bird inspired micro air vehicle and the study of its flexible wing aerodynamics. In [267], the authors present a novel fabrication process to create highly complex centimeter-scale wings. In [237], the authors present the design and controlled flight of an insect-like tailless micro-air vehicle.

BB.1 (*Sensing*) contains articles dealing with the design and modeling of sensory organs, such as [246] presenting a model of the lateral line of fish to investigate their behavior when affected by external flow fields, [94] presenting novel optical design methods and characterizations in order to study various compound eye concepts fabricated by micro-optics technology and [216] presenting experimental methods and models in order to study the physics of pressure difference receiving ears.

BB.2 (*Soft Robotics*) contains articles in the field of Soft Robotics. In [127], authors present a survey describing the working principles, the various uses and the future challenges and opportunities of dielectric elastomer actuators as soft actuators used in the design of soft robots. In [55], the authors present the design of a novel octopus-inspired multifunction silicon arm able to perform diverse tasks with minimum control. In [157], the authors present a soft actuator-based annelid robot and its effectiveness in performing effective locomotion in a large variety of unstructured environments.

BB.3 (*Materials*) contains articles dealing with the design and fabrication of new materials, such as [301] presenting a novel fabrication strategy to manufacture bio-inspired materials and their associated mechanisms with sufficient microstructural organization and mechanical performance, [123] studying the performance of laser-created bio-inspired textures with specific morphologies and [297] presenting the various technologies leveraged in order to build automatic self-healing materials.

BB.4 (*Behavior*) contains articles dealing with the study and design of control architectures related to animal behaviors. In [236], authors present a robotic platform-controlled fish replica simulating the courtship behavior observed in male fish and its impact on the positional preferences of female fish. In [310], authors present an abstract mathematical model and two decentralized control algorithms controlling autonomous flying robots in a swarm and the experimental results on their applicability on a group of autonomous quadcopters. In [4], authors present an overview of our knowledge of the soaring flight and strategy of birds and the associated control strategies that have been developed for soaring unmanned aerial vehicles in simulations and applications on real platforms, with an additional control strategy for exploiting thermal updrafts.

BB.5 (*Swimming*) contains articles dealing with the study and design of robots for underwater locomotion, such as [308] presenting the design, fabrication and performance of a jellyfish-inspired underwater vehicle, [101] presenting a mechanically-actuated foil model used in order to study the impact of body shape on various aspects of the swimming performance and [83] presenting the undulatory swimming properties of a knifefish-inspired

robot.

BB.6 (*Locomotion*) contains articles dealing with the study of various forms of locomotion. In [10], authors present two different jumping robots as a means of performing locomotion across rough terrains. In [189], authors present the various modes of locomotion adopted by different genus groups in multiple medias in order to lay the foundation required to design vehicles capable of multi-modal locomotion. In [35], authors present a model of the dynamics of human running to study running stability without explicit stabilization control strategies.

In the SR corpus we find seven meaningful clusters.

SR.0 (*Control and Model*) contains articles dealing with theoretical modeling and control algorithms of robots, such as [222] presenting a novel model for a soft continuum manipulator based on a material model, [115] presenting a machine-learning based approach for kinematic control of continuum manipulators capable of exhibiting adaptive behavior and [192] presenting a theoretical approach for the modeling of a pressure-operated soft snake robot. SR.1 (*Bending*) contains articles dealing with the bending subset of actuation. In [68], authors present a method for fabrication and dynamical modeling of a novel bidirectional bending soft pneumatic actuators that embeds a curvature proprioceptive sensor. In [6], authors present an analytical model and to estimate the bending displacement of a given of pneumatic soft actuators and its experimental validation. In [269], authors present a novel shape memory alloy strip-based bending actuator for a soft robotic hand with an analysis and design model of the shape memory alloy strip and its experimental validation.

SR.2 (*Artificial Muscles*) contains articles dealing with the design and modeling of artificial muscles, such as [131] presenting the design of a high-contraction ratio pneumatic artificial muscle using a novel actuation concept, [172] presenting an improvement on the design of thin McKibben muscles and [5] presenting the design of a novel extensor-contractor pneumatic artificial muscle.

SR.3 (*Jamming*) contains articles that deal with jamming in soft robotics. In [282], authors present a novel replacement for the traditionally rigid linkage between robot joints by adding an additional capability of stiffness controllability to the links. In [331], authors present the design and implementation of a novel approach to variable-stiffness tensegrity structures relying on the use of variable-stiffness cables. In [69], authors present the technical and market feasibility of a prosthetic jamming terminal device prototype in a pilot study with two upper-limb amputees and its performance compared to an existing commercial device.

SR.4 (*Tactile*) contains articles dealing with the design of elements linked with the tactile system such as [177] presenting a supervised machine learning approach to interpreting human touch in soft interfaces, [231] presenting the design of a lightweight soft robotic arm-wrist-hand system and [151] presenting the design and manufacturing of a 3-D printed tactile robot hand housing a soft biomimetic tactile sensor.

SR.5 (*Materials*) contains articles with the design and fabrication of novel materials. In [198], authors present approaches to designing and fabricating soft fluidic elastomer robots by studying three viable actuator morphologies. In [13], authors review the proper-

ties and characteristics of soft ionic polymer-metal nanocomposites and their applications in the field of soft robotics. In [252], authors present the different research efforts to develop actuators and robots for different types of structures containing shape memory alloys along with their respective strengths and weaknesses.

SR.6 (*Grasping*) contains articles dealing with the ability of robots to grasp items in their environment, such as [319] presenting the design, fabrication and experimental performance of a novel variable stiffness robotic gripper using soft actuating and particle jamming, [162] presenting the development of an autonomous motion planning algorithm for a soft planar grasping manipulator and [329] presenting a the design principle and experimental results of a novel bioinspired robotic finger in order to address two major challenges in soft pneumatic grippers.

Similarity and intersections between thematic clusters

Figure 3.6 shows the proportion of the origin corpus (LM, SR, BB) of individual articles in each cluster in the ME corpus. We see that corpora generally deal with different topics. Even though the ME corpus is dominated by BB in terms of number of articles, we see that clusters ME.0 and ME.1 are mostly composed of articles from LM. ME.0 and ME.1's articles deal with *Social* and *Learning and Sensing*, which are themes closely related to the Living Machines conferences. Most SR articles are found in ME.7 which is to be expected as this cluster deals with Soft Robotics. ME.4 (*Flying*), ME.5 (*Swimming*) and ME.6 (*Materials*) are almost exclusively comprised of articles from BB. ME.2 (*Bipedal Locomotion*) and ME.3 (*Tactile*) are smaller clusters comprised of articles from all 3 corpora. Most research areas present in the individual corpora are thus still individually represented in the ME corpus except for the very specific clusters in the SR corpus.

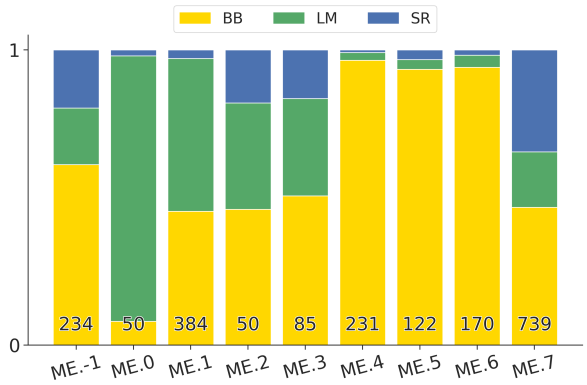

Figure 3.6: **Proportion of articles represented by the corpus of origin (BB, LM, SR) for each cluster of the ME corpus.** The total number of articles in each cluster is given at the bottom of each bar.

Figure 3.7 shows the cross-conference publication matrix where the proportion of authors publishing in multiple corpora is computed. In order to do so, we retrieve for all authors the list of the corpora they published in. As author names are not standardized between the various corpora, we perform homogenization steps. Individual author names

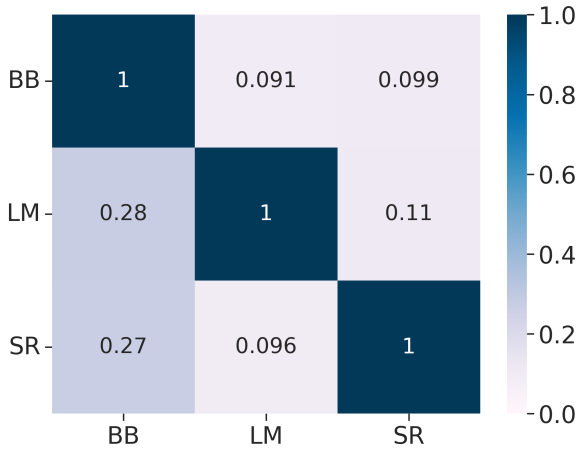

Figure 3.7: **Cross-conference publication matrix.** The number of individual authors publishing in other corpora is measured for each corpus, and the resulting matrix is then row-normalized, resulting in an asymmetrical matrix as the number of authors in each dataset is different. Each row thus represents the percentage of authors for the given row that publishes in the conferences corresponding to the columns.

are extracted and we only retain the first and last name removing middle names if present. We then only keep the capitalized first letter of the first name and the complete last name with the first letter capitalized (i.e. author *Herbert George Wells* would be named *H Wells* after homogenization). This methodology unfortunately is not able to discriminate between authors that have similar names. We then increment the pairwise count of the cross-conference publication matrix for all possible combinations in the list (e.g. if an author has published in LR and BB, the cell corresponding to row LR and column BB will be incremented by 1). Authors that published in a single corpus simply increment by a unit the diagonal cell corresponding to the corpus. We then row-normalize the final count matrix in order to have a proportion of authors that, for a given row, publish in other conferences, resulting in an asymmetrical matrix. If we look at the first row, we see that roughly 9% of authors in BB publish in LM and 10% publish in SR. 28% of LM authors publish in BB, and 11% of them publish in SR. Finally, 27% of SR authors publish in BB and roughly 10% publish in LM. We thus see that the LM and SR communities seem fairly distinct, and that BB equally attracts members from both the other communities. Furthermore, only 69 authors out of 4628 total individual authors published in all 3 corpora, suggesting that the overlap between all 3 communities is small.

Figure 3.8 shows the pairwise similarity matrix in the projected latent space of the ME corpus. Cross-cluster similarity values are computed following Eq. 3.17 for all pairs of clusters and the resulting matrix is then row-normalized, yielding an asymmetrical matrix. High matrix values denote thematically similar clusters, low matrix values denote thematically dissimilar clusters.

As expected, most clusters are maximally similar to themselves and those who are not have very high autosimilarity values (0.98 for LM.1 and BB.2, 0.99 for SR.5 and SR.6), suggesting that the clustering in each of the latent spaces holds on a fundamental level even if measured in a different latent space. We see that the upper-left block corresponding to

similarity between SR clusters is high, with the lowest value being 0.89. This is coherent with the fact that all articles in the SR corpus are about Soft Robotics, which is a sub-space of our ME latent space, meaning that SR clusters are relatively close in the ME latent space.

LM.3 (*Social*) and LM.6 (*Cognition*) have specific patterns, being very similar to each other and globally dissimilar to all other clusters, which can be expected as these topics are virtually nonexistent in either SR and BB. This research topic and the corresponding research community are more specific to LM conferences.

BB.0 (*Flying*) is fairly dissimilar to all other clusters as this topic is a very specific sub-domain of locomotion that is quite different to the rest of the corpus. Apart from itself, it is however most similar to BB.5 (*Swimming*) which can be explained as both clusters still deal with aerodynamics-related topics.

We see that articles dealing with similar topics cross-corpora have high similarity. For instance, LM.1 (*Soft Robotics*) has similarity of 0.96 with BB.2 (*Soft Robotics*) and BB.3 (*Materials*) and is fairly similar to all SR clusters, LM.2 (*Locomotion*) and BB.6 (*Locomotion*) have a similarity of 0.99, LM.0 (*Tactile*) and SR.4 (*Tactile*) have a similarity of 1.0, BB.1 (*Sensing*) has similarity 0.99 with LM.0 (*Tactile*) and similarity 0.97 with LM.5 (*Sensing*).

These results suggest that our methodology thus captures the abstract theme of the articles and clusters them following that theme regardless of the corpus of origin.

Figure 3.9 shows the comparison of thematic intersections of ME clusters in other corpora and the associated Adjusted Mutual Information (AMI) computed following eq. 3.19. For each corpus, we compare the clustering of its articles with the clustering of the same articles in the ME corpus.

In the top left panel of Fig. 3.9, we see that articles in LM.0 (*Tactile*) are mostly found in ME.3 (*Tactile*). Articles in LM.3 (*Social*) are mostly found in ME.0 (*Social*). Articles in LM.1 (*Soft Robotics*) and LM.2 (*Locomotion*) are mostly found in ME.7 (*Soft Robotics*). Articles in LM.5 (*Sensing*), LM.6 (*Cognition*) and LM.7 (*Sensorimotor*) are mostly found in ME.1 (*Learning and Sensing*). Articles in LM.4 (*Legged Locomotion*) seem to be evenly spread between ME.4 and ME.7. These allocations are as exepected on a general level, but we can also notice that some ME clusters such as ME.4 (*Flying*), ME.5 (*Swimming*) and ME.6 (*Materials*) are not represented in the LM corpus. Most unlabeled articles in the LM corpus are allocated to ME.1.

In the middle left panel, we see that articles in BB.0 (*Flying*) are found in ME.4 (*Flying*). Articles in BB.2 (*Soft Robotics*) and BB.6 (*Locomotion*) are mostly found in ME.7 (*Soft Robotics*), as was the case in the LM corpus. Articles in BB.4 (*Behavior*) are mostly found in ME.1 (*Learning and Sensing*), and articles in BB.3 (*Materials*) are mostly found in ME.6 (*Materials*). A number of BB.1 articles (*Sensing*) are unlabeled, with the rest being found in ME.1 and ME.3 (*Tactile*). Articles in BB.5 (*Swimming*) are mostly found in ME.5 (*Swimming*). Once again, these cross-corpora allocations are as expected, with global themes being conserved between individual corpora and the ME corpus. The unlabeled articles are spread relatively evenly throughout the individual ME clusters.

The bottom left panel shows that almost all articles in the SR corpus are found in ME.7 (*Soft Robotics*), which is to be expected as all articles in the SR corpus have a Soft Robotics

component. Most articles in SR.4 (*Tactile*) are found in ME.3 (*Tactile*). As Soft Robotics is a dedicated cluster in the ME corpus, the SR articles are correctly clustered in the ME corpus.

Temporal evolution of research trends

Figure 3.10 shows the temporal evolution of the cluster proportions in their corresponding corpus. For every corpus and for every year, the proportion of articles in each cluster compared to the total number of articles published in the year is computed with the cluster colors for each corpus corresponding to those shown in figures 3.3 and 3.4.

This temporal evolution seems to be dominated by fluctuations, and no clear discernible pattern emerges. We can however see that cluster SR.6 (*Grasping*) replaces SR.5 (*Materials*) in 2016 as the dominant topic in the SR corpus. SR.3 (*Jamming*) also emerges around 2016 and is present in all subsequent years, albeit weakly for some editions. As Soft Robotics is still an emerging domain thus making SR the youngest corpus in our study, its lower maturity might explain the emergence of new clusters representing subtopics of the domain.

ME.4 (*Flying*) was one of the dominant topics in the ME corpus between 2009 and 2011, and has drawn considerably less attention in the later editions. Conversely, the *Learning and Sensing* (ME.1) and *Soft Robotics* (ME.7) clusters have grown in importance from 2012 onward. *Materials* (ME.6) was one of the main topics until 2011 in our merged corpus, and has been weakly represented since then.

Figure 3.11 shows the temporal evolution of the 1-gram and 2-gram proportion in the ME corpus. The n-grams were taken from Table 1 in the Supplementary Information and underwent manual curation in order to remove non-descriptive n-grams and duplicates and to fuse singular and plural forms. The number of articles containing the n-grams is computed for each year and divided by the total number of articles in the given year. We see that *Flying*-related keywords such as **flapping**, **wing**, **flight**, **wing kinematics** or **micro air** are particularly prevalent between 2009 and 2011 but become comparatively less common afterwards. Keywords related to *Soft Robotics* such as **soft actuators**, **variable stiffness**, **pneumatic**, **materials** or **soft robots** are becoming some of the most frequent words from 2017 onward, showing the emergence of the field of Soft Robotics. **Neural networks** and **real time** systems are also seeing increasing interest in recent publications, most likely facilitated by the democratization of Artificial Intelligence. **Locomotion**, on the other hand, has been a prevalent n-gram for the entirety of the period of study as one of the pillar topics of robotics.

Figure 3.12 shows the number of occurrences of selected keywords in the ME corpus.

Future research directions

We investigate which research themes are currently underdeveloped in the studied corpora. Namely, we want to identify which themes could become emerging and active research directions in the future. Previous plenary talks [241, 207, 270] in the LM 2021 conference already proposed several themes which could become emerging topics, mostly related to

the notion of sustainability [28]. This particular topic is becoming increasingly popular in Science due to the effects of climate change, resources depletion and ecological problems that already impact the world today. Another emerging theme of research would be the development of biohybrid systems that include both natural and artificial components. Figure 3.12 shows the number of articles of the ME corpus that include keywords related to specific underdeveloped topics, related either to sustainability or to biohybrid approaches. Out of 2099 articles in the corpus, we find that a lot of topics considered essential for our future (such as sustainability, agriculture or ethics) are only weakly represented with less than 20 articles mentioning sustainability-related keywords and less than 10 articles mentioning either ethics-related or agriculture-related keywords. Efficiency-related keywords, encompassing both material efficiency and energy efficiency, are mentioned in 60 articles whereas energy autonomy-related keywords and self healing-related keywords are mentioned in less than 20 articles n the merged corpus. Biohybrid systems are mentioned in 36 articles and 14 articles mention biorobotics. Other biohybrid-related keywords (*ecosystem-active robots*, *microorganism-robots*, *mixed societies* or *organic control*) are virtually absent from the corpus as they are mentioned six times or less.

3.2.3 Discussion

We developed a methodology to generate a state-of-the-art of the research themes in the Living Machines conferences and associated journals (Bioinspiration & Biomimetics and Soft Robotics). We have extracted the research trends of this field in a context of scientific surveillance. The methodology developed here is a first step in automating the classification of research topics in the scientific literature. We group the research topics into clusters and present new metrics to compare these clusters with each other and with the different corpora considered. Finally, we identify research topics that are currently underrepresented in this community and potential research directions.

The corpus analysis presented here aims to test and develop techniques applicable to larger corpora containing thousands or even millions of articles. However, the scalability remains to be tested. The choice of this limited corpus aims first to locate LM conferences in relation to the community that publishes in Bioinspiration & Biomimetics. Second, the choice is also guided by our knowledge of the field as experts which allows us to quickly determine if the method produces meaningful results.

It is conceivable that text analysis algorithms could be useful for classifying articles, writing summaries for groups of articles and identifying emerging research trends. The results could then contribute to make the state of the art in a field. It could also allow experts in the field to identify research work which is underdeveloped or over-represented.

Our approach could be improved in several ways.

It would also be interesting, as long as the corpora are larger, to do internal analysis of each main cluster. This could reveal the hierarchical structure of themes within a cluster. Our clustering approach already makes use of HDBSCAN, a hierarchical clustering algorithm – however we do not currently take into account the sub-clusters it finds. Indeed, as it is a “hard-clustering method", documents classified by HDBSCAN can only be assigned

to one cluster at a time (as opposed to “soft clustering" methods where documents can be assigned to several clusters). This results to a fragmentation of the dataset into a large ($>$ 50) number of sub-clusters with possibly very redundant information, making them difficult to understand by humans. As such, our method could be improved by making use of hierarchical soft clustering algorithms. This would allow each document to be part of several clusters, which would be useful to accurately classify articles at the intersection of several domains.

The quality and clarity of the clustering results obtained by our current approach heavily depend on carefully chosen hyper-parameters. Indeed, there could exist several concurrent clustering instances of the datasets that could be relevant. However, the relevance of each clustering instance is difficult to quantify automatically, and require the intervention of a human expert to select the most relevant and understandable clustering instance. The selected clustering instance may not be the one with the least amount of unclassified documents. Our approach could be improved to automatically select good hyper-parameters to generate relevant clustering instances, and to reduce the number of unclassified documents as much as possible.

Some part of our analysis are currently done manually and could be automated. In particular, Fig. 3.5 could have a structure generated based on a hierarchical clustering method (cf previous paragraphs). The names of categories in Fig. 3.5 and cluster labels in Table 3.2 could be obtained automatically through a keyword generation method. It is very difficult, even for a human expert, to select labels that accurately describe the particular themes of a given cluster. Currently, our approach already extracts relevant keywords of each cluster, that will be then used by the expert to select labels. Often these labels are not terms that are not directly found by the keyword extraction method, because they describes the cluster content at an higher level of abstraction compared to the terms used directly in the articles. However, most keywords generation approaches [268, 126] are extractive (i.e. they find the most relevant existing words in the studied text) rather than abstractive (i.e. they generate new keywords that are not present in the text).

Alternatively, it may be possible to qualify the content of each cluster not just with labels but also by small summaries of their research themes. This would be achieved through abstractive multi-document summarization techniques [29].

Here, we focus on a limited corpus to answer an analysis between 3 types of publications related to Living Machines. Later it will be interesting to extend this type of analysis by searching in the literature the same themes as those found in our analysis here. To do this type of analysis means to study a corpus composed of several thousands of articles which becomes impossible to do manually.

Here, we only use the title of the articles and their abstract and keywords. It is clear that it would be interesting to extend the “reading" to the body of the articles, which would open new questions. Such process would involve additional technical difficulties: abstracts are already condensed versions of the main text of the article, which would make them easier to understand by the algorithms. However the main text would incorporate more details about their research theme which could be captured by the language model to refine its classification.

This short analysis shows that the LM conferences are more focused on robotics and control architectures. This also shows that the actuation aspects of robotics represent a dominant research theme. The theme of actuation includes the largest number of clusters. The originality of these conferences lies in the diversity of robotics topics covered from a biomimetic point of view. The social aspects and in particular the human-robot relations are well-represented and absent from the wider field covered by Bioinspiration & Biomimetics and Soft Robotics. Similarly the cognitive aspects are also more present than in Bioinspiration & Biomimetics and Soft Robotics. Surprisingly, for a community close to the living world, the themes concerning the sustainability of technologies are still mostly absent while ecological issues are the most pressing [130].

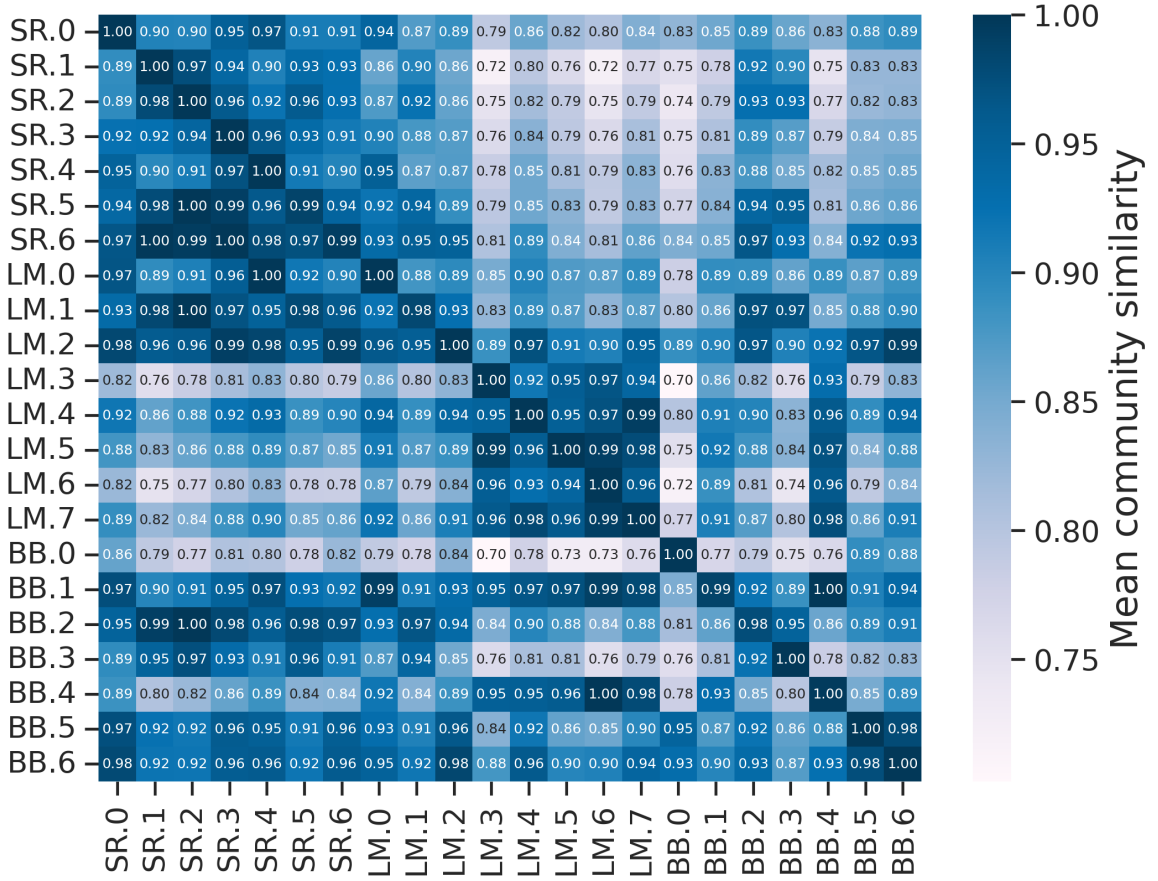

Figure 3.8: **Pairwise cluster similarity matrix in the latent space of the merged corpus.** The pairwise euclidean similarity between all clusters was computed according to equation 3.17; the similarity is then divided by the maximum value for each row of the matrix, making this similarity measure asymmetrical. Higher similarity measures between two clusters correspond to thematically similar clusters. The row corresponding to a given cluster provides information about the ranked similarity of the other clusters. The column corresponding to a given cluster provides global information about how this cluster ranks for each of the other clusters. The unlabeled clusters in each corpus were removed to reduce noise in the figure.

Figure 3.9: **Comparison of thematic intersections of ME clusters with clusters in other corpora.** Each panel represents the proportion of documents in each cluster of a given corpus also present in each ME cluster. The cluster to which the articles belong is indicated by the labels on the x-axis. The colored histogram represents their membership to the other corpora identified by the colored labels above the panels. Articles initially categorized as part of one specific cluster are analyzed to see to which ME cluster they belong to. For instance, looking at the top left panel, we see that articles belonging to cluster 0 in the Living Machines corpus (LM.0) are mostly found in cluster 6 of the Merged corpus (ME.6) with the rest of them either unlabeled or belonging to clusters 2 (ME.2) and 5 (ME.5) in the Merged corpus. The Adjusted Mutual Information (AMI, eq. 3.19) score is also provided to measure the global similarity between the clustering results compared for each panel. Values of the AMI close to 0 correspond to random clusterings, high values of the AMI correspond to similar clusterings. The numbers under each cluster name correspond to the number of articles in the cluster present in both corpora.

Figure 3.10: **Temporal evolution of the proportion of articles in each theme over the editions of the various corpora.** The colors represent individual clusters and correspond to those shown in figures 3.3 and 3.4.

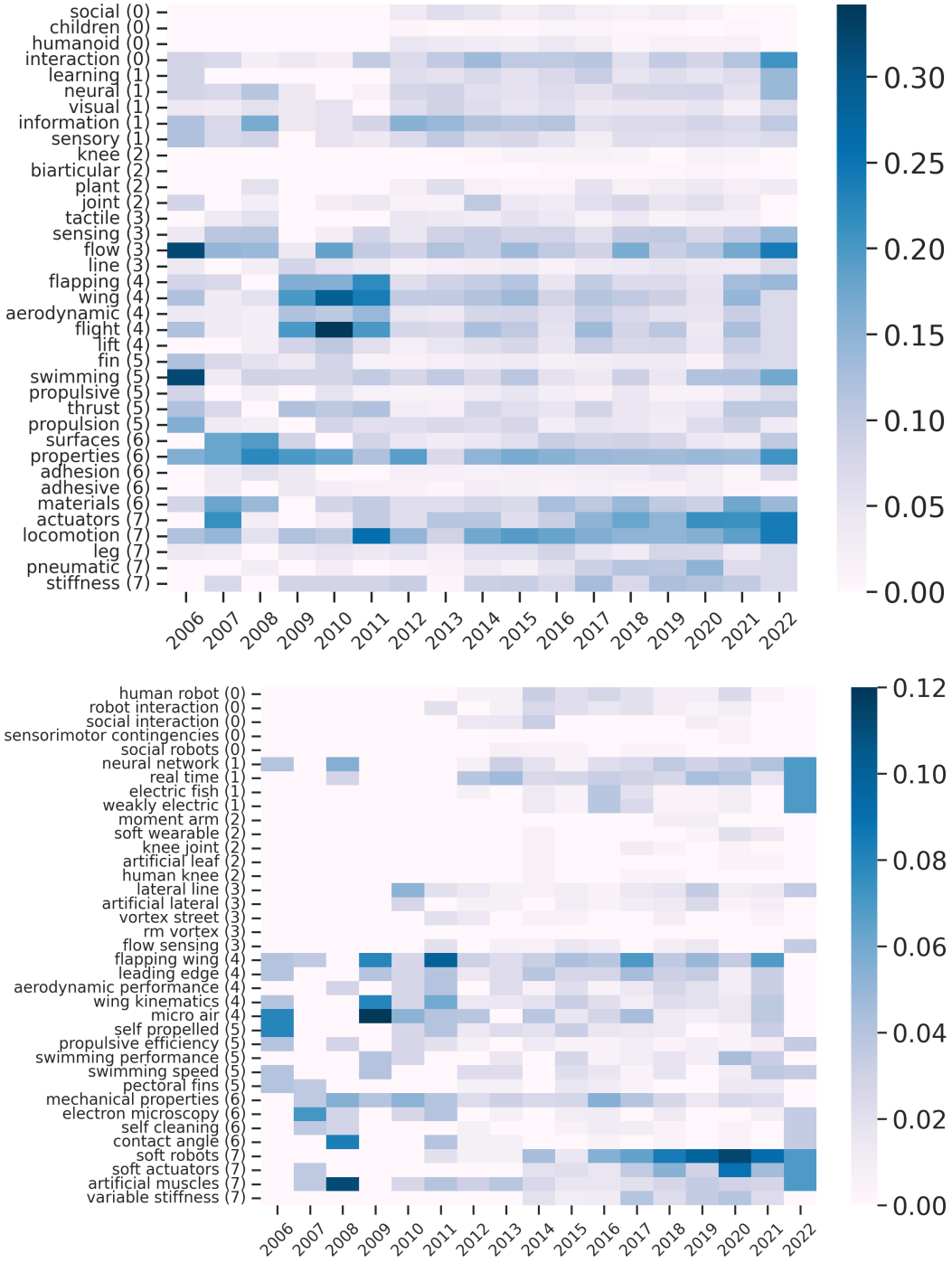

Figure 3.11: **Temporal evolution of the proportion of 1 (top) and 2-grams (bottom) in the ME corpus normalized by the total number of articles for a given year in the corpus.** Manual curation was performed to remove non-descriptive keywords (e.g. **bio-inspired** or **experimental results**). As a reminder, the names of each cluster are as follows : 0 = *Social*, 1 = *Learning and Sensing*, 2 = *Bipedal Locomotion*, 3 = *Tactile*, 4 = *Flying*, 5 = *Swimming*, 6 = *Materials*, 7 = *Soft Robotics*.

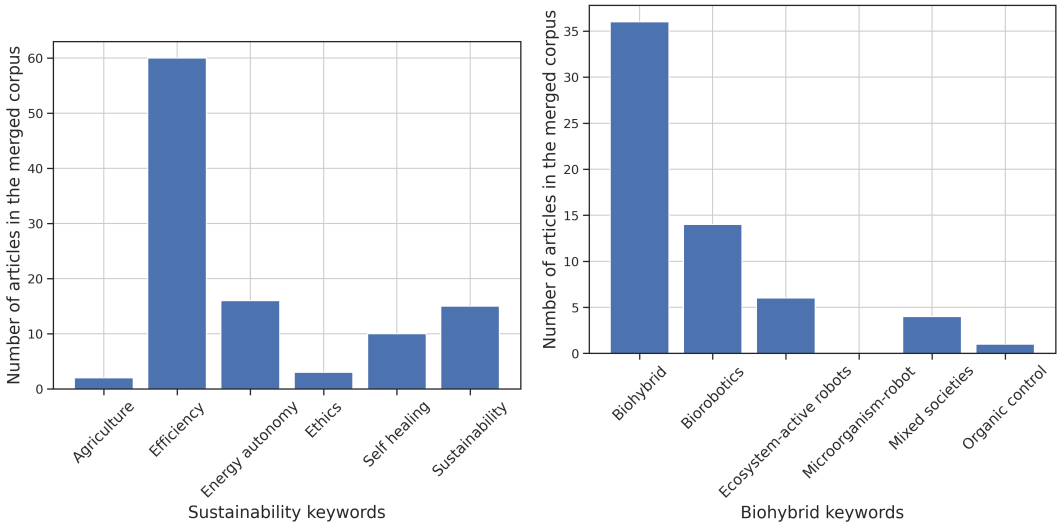

Figure 3.12: **Number of articles in the merged corpus containing keywords relating to major themes.** The themes were selected based on Plenary Talks of the Living Machines 2021 conference [241, 207, 270]. The list of keywords used to find articles relating to each major theme is shown in Table 3.5.

3.3 Appendix

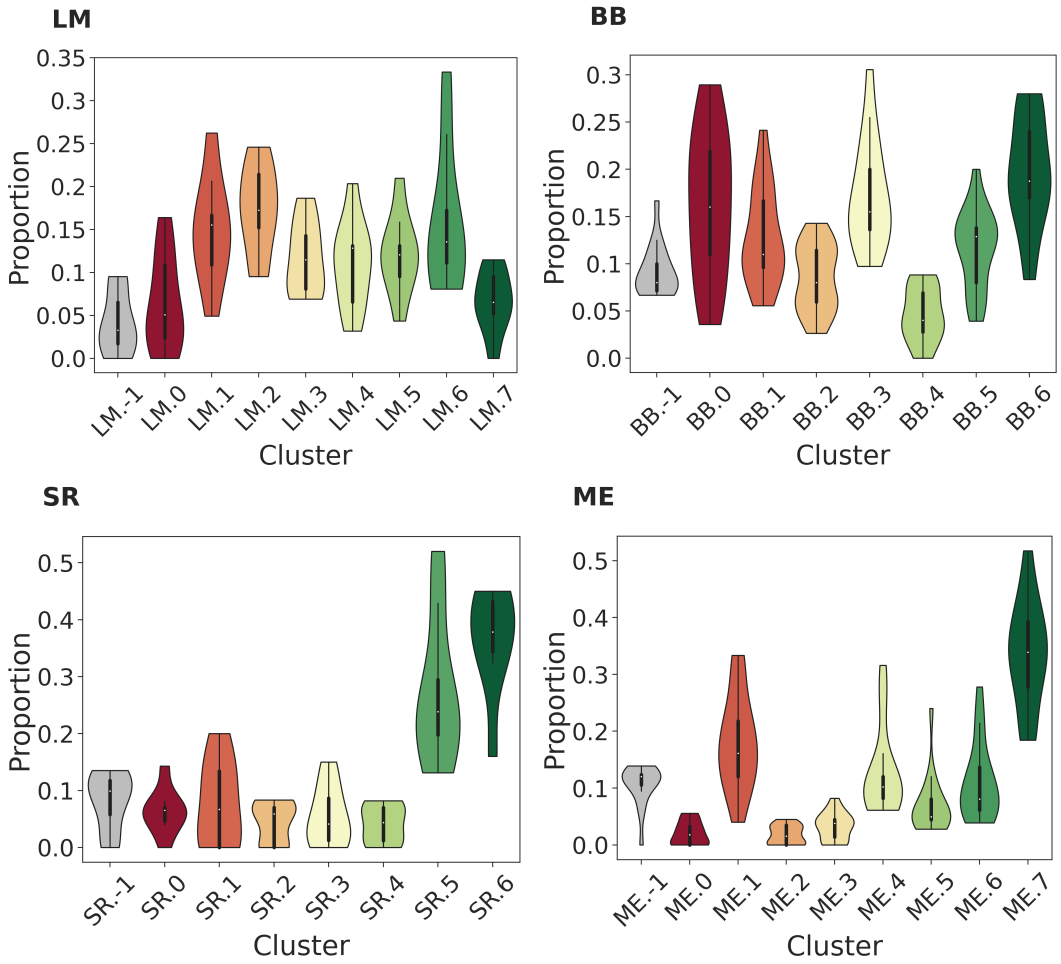

Figure 3.13: **Violin plots for the distribution of the number of articles in each cluster over time.**

Living Machines	**1-grams**	**Cluster -1 Unlabeled**	**Cluster 0 Tactile**	**Cluster 1 Soft Robotics**	**Cluster 2 Locomotion**	**Cluster 3 Social**	**Cluster 4 Legged Locomotion**	**Cluster 5 Sensing**	**Cluster 6 Cognition**	**Cluster 7 Sensorimotor**
		cerebellar	tactile	mechanical	locomotion	social	joint	consciousness	memory	fly
		cerebellum	sensor	cells	mechanism	children	walking	biomimetics	learning	visual
		motor	touch	muscle	body	interaction	leg	fish	hippocampal	loop
		acquisition	fingertip	plant	peristaltic	human	controller	research	navigation	interface
		search	whisker	swimming	snake	humanoid	muscles	odor	models	closed
	2-grams	**Cluster -1 Unlabeled**	**Cluster 0 Tactile**	**Cluster 1 Soft Robotics**	**Cluster 2 Locomotion**	**Cluster 3 Social**	**Cluster 4 Legged Locomotion**	**Cluster 5 Sensing**	**Cluster 6 Cognition**	**Cluster 7 Sensorimotor**
		odor source	biomimetic tactile	work present	control scheme	human robot	human balance	optic flow	neural networks	closed loop
		mobile treadmill	sensory augmentation	artificial muscle	snake like	social interaction	central pattern	compound eye	autobiographical memory	fly robot
		motor responses	tactile sensing	mechanical properties	decentralized control	robot interaction	neural network	machine learning	decision making	robot interface
		omnidirectional mobile	tactile sensors	muscle cells	peristaltic pumping	vocal fold	balance control	living machines	reinforcement learning	eye movements
		anticipatory postural	active touch	plant roots	like robot	social robots	hind limb	odor tracking	visual navigation	spike rate
Bioinspiration & Biomimetics	**1-grams**	**Cluster -1 Unlabeled**	**Cluster 0 Flying**	**Cluster 1 Sensing**	**Cluster 2 Soft Robotics**	**Cluster 3 Materials**	**Cluster 4 Behaviour**	**Cluster 5 Swimming**	**Cluster 6 Locomotion**	
		human	aerodynamic	sensor	soft	materials	visual	swimming	leg	
		hand	wings	sensing	actuator	surfaces	zebrafish	fin	locomotion	
		finger	lift	array	manipulator	properties	rat	propulsion	running	
		controller	flapping	line	muscles	structures	collective	undulatory	gait	
	2-grams	**Cluster -1 Unlabeled**	**Cluster 0 Flying**	**Cluster 1 Sensing**	**Cluster 2 Soft Robotics**	**Cluster 3 Materials**	**Cluster 4 Behaviour**	**Cluster 5 Swimming**	**Cluster 6 Locomotion**	
		closed loop	flapping wing	lateral line	soft robots	mechanical properties	optic flow	swimming speed	inverted pendulum	
		moment arm	leading edge	artificial lateral	soft actuators	bio inspired	spiking neural	swimming performance	robotic fish	
		control law	aerodynamic performance	compound eye	bio inspired	self cleaning	biologically inspired	self propelled	legged locomotion	
		feedback control	wing kinematics	vortex street	dielectric elastomer	electron microscopy	mobile robots	pectoral fins	loaded inverted	
		control strategy	micro air	active electrolocation	biological muscles	contact angle	dummy fish	underwater vehicles	spring loaded	
Soft Robotics	**1-grams**	**Cluster -1 Unlabeled**	**Cluster 0 Control and Model**	**Cluster 1 Bending**	**Cluster 2 Artificial Muscles**	**Cluster 3 Jamming**	**Cluster 4 Tactile**	**Cluster 5 Materials**	**Cluster 6 Grasping**	
		fin	catheter	knee	collagen	muscle	tactile	properties	grasping	
		heart	continuum	contractions	tcas	pss	tactip	materials	finger	
		whisker	kinematic	dielectric	sfm	pjtd	spl	stent	objects	
		segments	trajectory	pams	contraction	proprioceptive	touch	fabrication	gripper	
		thrust	pain	vsdea	growing	spa	manipulation	devices	locomotion	
	2-grams	**Cluster -1 Unlabeled**	**Cluster 0 Control and Model**	**Cluster 1 Bending**	**Cluster 2 Artificial Muscles**	**Cluster 3 Jamming**	**Cluster 4 Tactile**	**Cluster 5 Materials**	**Cluster 6 Grasping**	
		ribbon fin	model based	dielectric elastomer	contraction ratio	jamming structures	tactile sensing	soft robotics	soft robot	
		peristaltic locomotion	contact force	bending angle	artificial muscle	loop control	soft biomimetic	soft actuators	variable stiffness	
		long term	control framework	soft actuators	mckibben muscles	tubular jamming	biomimetic finger	material properties	article presents	
		bending moment	closed loop	soft rigid	collagen microparticles	optical sensor	body motion	soft materials	body length	
		median fins	proposed model	soft lens	free strokes	closed loop	texture discrimination	finite element	soft gripper	
Merged Corpus	**1-grams**	**Cluster -1 Unlabeled**	**Cluster 0 Social**	**Cluster 1 Learning and Sensing**	**Cluster 2 Bipedal Locomotion**	**Cluster 3 Tactile**	**Cluster 4 Flying**	**Cluster 5 Swimming**	**Cluster 6 Materials**	**Cluster 7 Soft Robotics**
		continuum	social	learning	human	tactile	flapping	fin	surfaces	actuators
		needle	children	neural	knee	sensor	wing	swimming	properties	locomotion
		dynamic	humanoid	visual	biarticular	sensing	aerodynamic	propulsive	adhesion	leg
		lens	human	information	plant	flow	flight	thrust	adhesive	pneumatic
		sonar	interaction	sensory	joint	line	lift	propulsion	materials	stiffness
	2-grams	**Cluster -1 Unlabeled**	**Cluster 0 Social**	**Cluster 1 Learning and Sensing**	**Cluster 2 Bipedal Locomotion**	**Cluster 3 Tactile**	**Cluster 4 Flying**	**Cluster 5 Swimming**	**Cluster 6 Materials**	**Cluster 7 Soft Robotics**
		control law	human robot	neural network	moment arm	lateral line	flapping wing	self propelled	mechanical properties	soft robots
		compound eye	robot interaction	real time	soft wearable	artificial lateral	leading edge	propulsive efficiency	bio inspired	soft actuators
		bio inspired	social interaction	electric fish	knee joint	vortex street	aerodynamic performance	swimming performance	electron microscopy	artificial muscles
		focal length	sensorimotor contingencies	weakly electric	artificial leaf	rm vortex	wing kinematics	swimming speed	self cleaning	variable stiffness
		time variant	social robots	bio inspired	human knee	flow sensing	micro air	pectoral fins	contact angle	bio inspired

Table 3.3: **Top n-grams for all corpora.**

Cluster -1	Cluster 0	Cluster 1	Cluster 2	Cluster 3
tau	income	financial	materials	wing
law	distribution	stock	soft	flapping
growth	wealth	price	structures	flight
mf	money	market	biomimetic	control
return	law	economic	surfaces	robot

Table 3.4: **1-grams for the CO dataset.**

Major Theme	Substrings
Sustainability	sustainab
Efficiency	energy effic, material effic, store energ, energy stor
Ethics	ethic
Agriculture	agricultur
Energy autonomy	energy auton, energy gather, harvesting energ, energy harvest, harvest energ, gather energ, energetically autonom
Self healing	self healing
Biohybrid	bio hybrid, biohybrid
Biorobotics	bio robotic, biorobotic
Mixed societies	animal robot, plant robot, mixed societ, robot plant, fish and robot, robot and fish
Microorganism-robot	microorganism robot, micro organism robot, robot micro organism
Ecosystem-active robots	ecosystemactive robot, ecosystem active robot, ecosystem
Organic control	organic actuator, organic control

Table 3.5: **Substrings used to find articles in the ME corpus relating to major themes as determined by Plenary Talks of the Living Machines 2021 conference.**

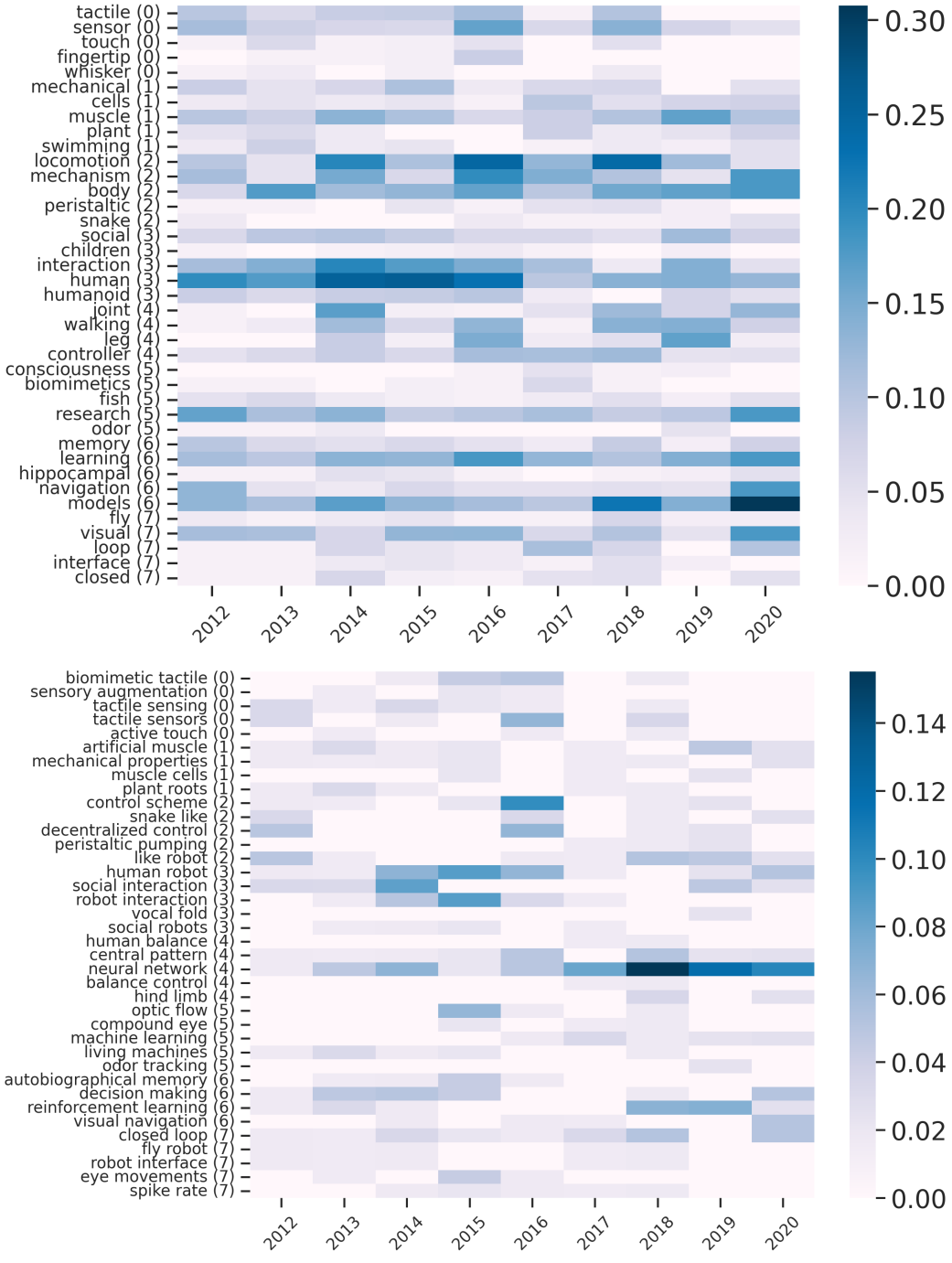

Figure 3.14: **Temporal evolution of the proportion of 1 (top) and 2-grams (bottom) in the LM corpus.** The number of 1- and 2-grams is normalized by the total number of articles for a given year in the conference or journal issue.

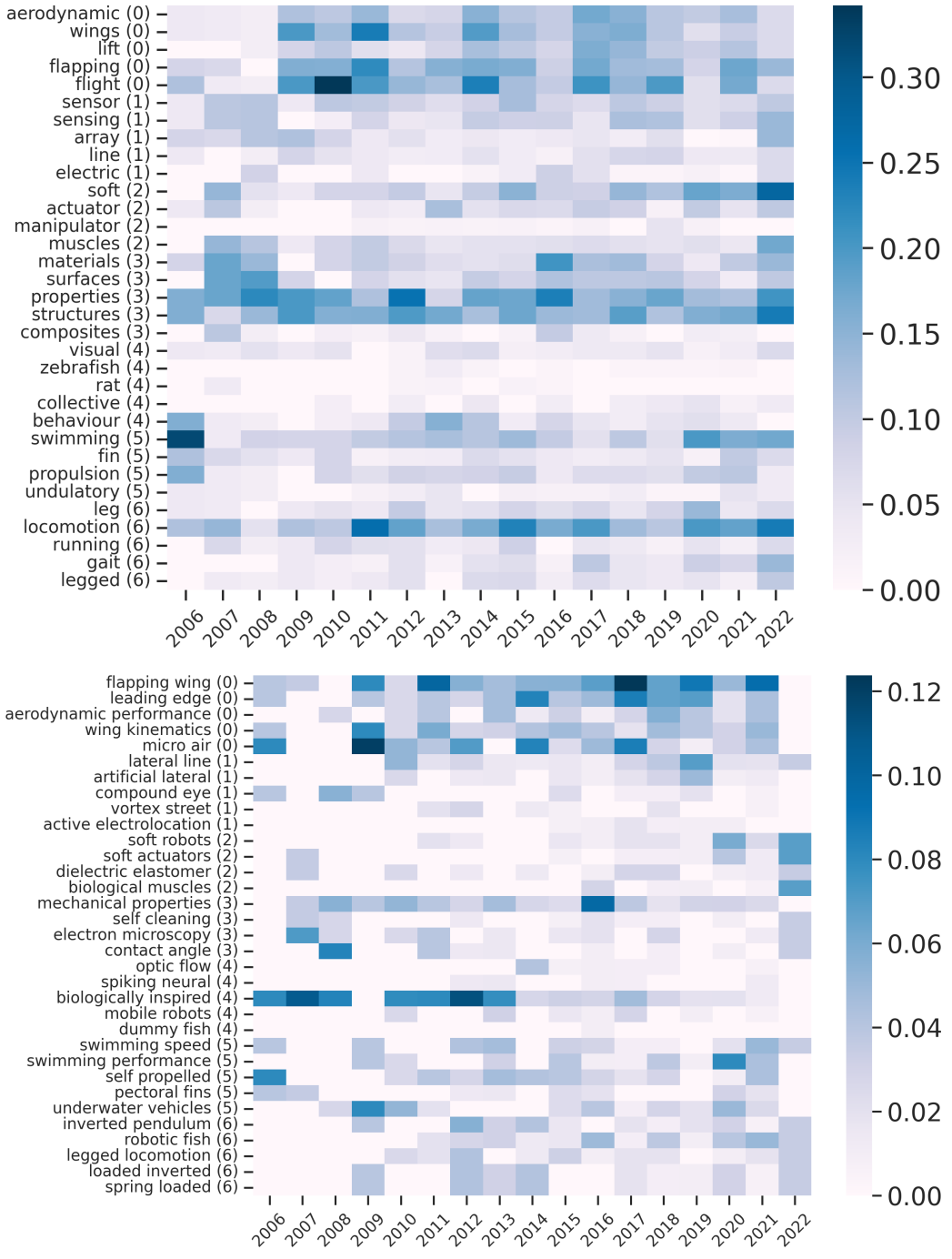

Figure 3.15: **Temporal evolution of the proportion of 1 (top) and 2-grams (bottom) in the BB corpus.** The number of 1- and 2-grams is normalized by the total number of articles for a given year in the conference or journal issue.

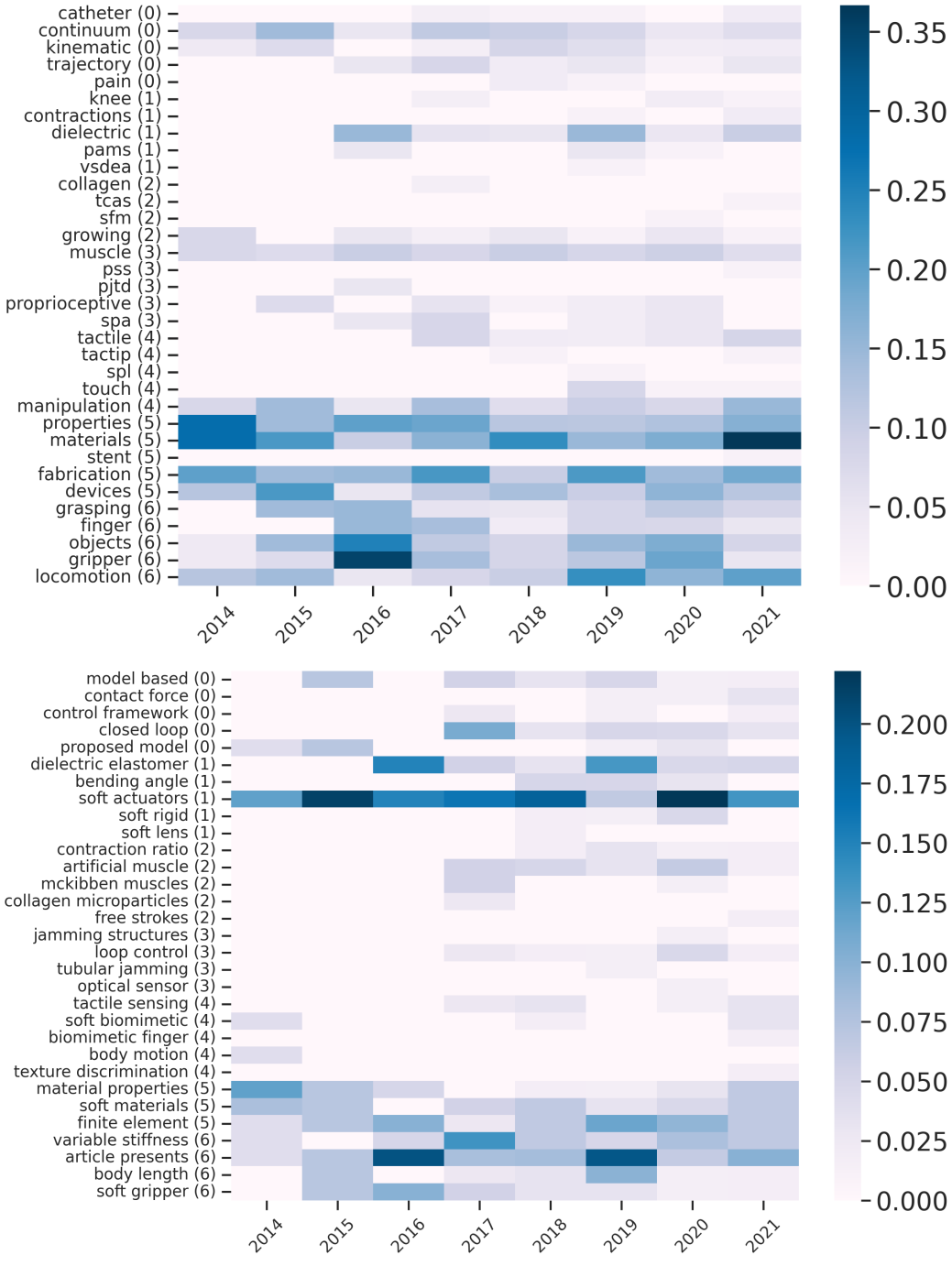

Figure 3.16: **Temporal evolution of the proportion of 1 (top) and 2-grams (bottom) in the SR corpus.** The number of 1- and 2-grams is normalized by the total number of articles for a given year in the conference or journal issue.

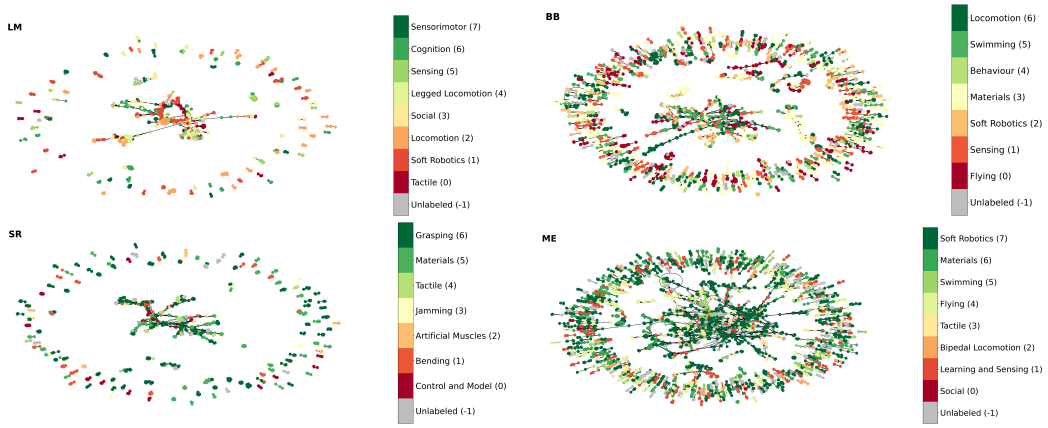

Figure 3.17: **Coauthorship network with nodes colored following the cluster allocation for individual authors.** An author is allocated to the cluster where he has the most publications. To reduce the network, only a few selected author nodes are annotated. We remove components that have less than 3 nodes in the graph to reduce visual clutter and, for each individual component with 10 nodes or more, we label the most active author.

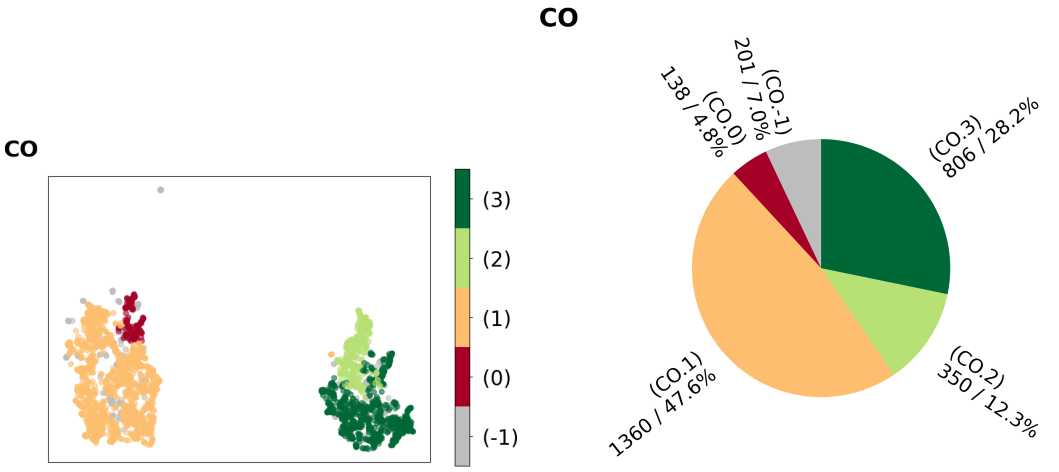

Figure 3.18: **Automatic analysis of a control (CO) dataset.** The CO dataset is made by mixing quantitative finance articles extracted from ArXiV in July 2021 with the BB corpus. The red and yellow points (CO.0 and CO.1) correspond to quantitative finance articles, the light and dark green points (CO.2 and CO.3) to BB articles. The split between the two datasets in the 2-D latent space is clear, suggesting that our algorithm is capable of correctly discriminating between two very different scientific domains.

To go further on keyword extraction

In this section, we will discuss an alternative method to extract the keywords most relevant to each cluster based on their underlying concepts rather than purely on their frequency.

We first compute the c-tf-idf scores and extract a large number of relevant keywords for each topic as previously described in this chapter (eq. 3.15). We then add an additional processing step to remove similar keywords (e.g. *attention* and *attentions*) by sequentially going through the keywords in descending order of c-TF-IDF score and only keeping those that have Levenshtein similarity with all keywords of higher score **lower** than a threshold (we set the threshold at 0.8) until we have the desired top k keywords.

We retrieve the top k 1-grams, 2-grams and 3-grams using the aforementioned method, and amalgamate them before beginning the second phase of our keyword extraction algorithm. All these n-grams combined together are henceforth referred to as *candidate keywords* for a given cluster (note that a given n-gram can be a candidate keyword for several clusters). As computing word embeddings is relatively slow, this first phase of reducing the number of candidate keywords is important in order to allow for the scaling of this method to relatively large corpora.

In the second phase, we endeavor to select a small number of most relevant and diverse keywords from all the n-grams previously extracted. In order to do so, we first compute for each cluster c_i the global embedding $\mathbf{p}_{c_i}$ of the cluster (Eq. 3.20) :

$$\mathbf{p}_{c_i} = \frac{\sum_{k=1}^{n_{c_i}} \mathbf{e}(c_i^k)}{n_{c_i}} \tag{3.20}$$

where n_{c_i} is the number of documents in cluster i and $\mathbf{e}(c_i^k)$ is the embedding of the k-th document in cluster i with $k \in [1, n_{c_i}]$.

We refer to all n-grams selected by the following method as *selected keywords*. At the beginning of this algorithm and for a given cluster, all n-grams obtained after the first phase are thus candidate keywords and the list of selected keywords is empty. For each candidate keyword w, we compute the similarity of its embedding $\mathbf{e}(w)$ with the cluster's embedding $sim_{candidate}(w)$ and the similarity of its embedding with the maximally-similar selected keyword $sim_{selected}(w)$ (Eq. 3.23) :

$$\cos(\mathbf{a}, \mathbf{b}) = \frac{\mathbf{a} \cdot \mathbf{b}}{||\mathbf{a}||\,||\mathbf{b}||} \tag{3.21}$$

$$sim_{candidate}(w) = \cos(\mathbf{p}_{c_i}, \mathbf{e}^{\intercal}(w)) \tag{3.22}$$

$$sim_{selected}(w) = \max(\cos(\mathbf{e}(w), \widehat{\mathbf{K}}^{\intercal})) \tag{3.23}$$

where $\cos(..., ...)$ is the cosine similarity, $\mathbf{e}(w)$ is the embedding vector of candidate keyword w, $\widehat{\mathbf{K}}$ is the matrix containing the embeddings of all selected keywords and $\max(...)$ corresponds to the maximum of the resulting similarity vector (we find the selected keyword most similar to candidate keyword w).

We then compute the mmr (Maximal Marginal Relevance [59]) value over all candidate keywords following Eq. 3.24 and pick the keyword with the maximum mmr value, which we then add to our list of selected keywords and remove from the list of candidate keywords :

$$mmr(w) = (1 - d) \times sim_{candidate}(w) - d \times sim_{selected}(w) \tag{3.24}$$

with $d \in [0, 1]$ a diversity parameter to control the diversity when sequentially selecting keywords ($d = 0$ corresponds to no diversity parameter, $d = 1$ corresponds to maximally diverse keywords).

We repeat this process until we have the desired number n of selected keywords for each cluster. This two-phased keyword selection method presents several advantages :

- by first performing a selection of relevant n-grams with c-TF-IDF, we greatly reduce the number of candidate keywords (and thus the time and hardware requirements) of the mmr phase.
- by using n-gram embeddings instead of statistical methods such as c-TF-IDF for our keyword selection, we can explicitly add a diversity parameter and select n-grams based on their *semantic meaning* rather than pure mathematical frequency.
- using embeddings yields a more diverse mixture of 1-, 2- and 3-grams compared to just using c-TF-IDF which tends to over-represent higher order n-grams. This bias is still present as higher order n-grams are more numerous and are structurally able to be more descriptive than 1-grams, but is not as proeminent compared to simply using c-TF-IDF.

Chapter 4

Investor-patent networks as topologically mutualistic networks

This chapter is based on Carniel, T., Cazenille, L., Dalle, J. M., & Halloy, J., 2023 : *Investor-patent networks as mutualistic networks* (arXiv preprint arXiv:2311.18625.).

In this chapter, we build and study the bipartite graph of venture-backed innovation where investors are connected to patents through the patent portfolios of the startups they finance. Leveraging the methodologies described in chapters 2 and 3, we perform clustering on both investors and patents, resulting in a coarse-grained graph representation where investor communities are linked to technologies (clusters of thematically similar patents). Drawing on metrics from ecology, we then characterize the structure of this graph and find it to be topologically mutualistic due to the prevalence of links between generalist investors, whose portfolios are technologically diversified, and general-purpose technologies, characterized by a broad spectrum of use. As a consequence, the robustness of venture-funded technological innovation against different types of crises is affected by the high nestedness and low modularity, with high connectance, associated with mutualistic networks.

4.1 Introduction

Due to the increased role of startups in various technological fields, from biotechnology to artificial intelligence or quantum computing [12], venture capital has *de facto* become an important driver of technological innovation [174, 103, 104, 181]. Indeed, because of their inability to self-finance during the early years of their operations, startups need to rely on the investments that they themselves receive from specialized investors called venture capitalists (VCs). In this context, the relationships between VC funding and innovation have gradually become a topic of interest, mostly approached through patent data [224, 119, 145, 84]. This evolution has occurred in a world where crises have become more and more frequent [132], increasing the need to analyze the resilience of socio-economic systems [111]. These studies have notably preliminarily shown that VC funding could be negatively im-

pacted by local and global crises, be they financial [33], health [23] or geopolitical [175]. However, it is quite surprising that, although the VC network has been an active topic of study for the past 15 years [140, 186], and although [104] had pioneered a complex network approach focused on robustness, the direct interactions between VCs and the innovations they fund, based explicitly on the patents filed [2] by the startups funded, have not been explicitly studied, both in their own right and in relation to the robustness of the network they constitute. This VC-patent interaction network is bipartite, with nodes of a first class (VCs) interacting with nodes of a second class (patents) through investments in startups that file the patents.

In this study, considering the line of analysis of financial markets suggested by [183, 184] and echoing also the approaches that have started to directly address economic complexity [136], we combine large-scale financial datasets on the rounds of VC funding received by startups with patent data to explicitly analyze this bipartite investor-patent network and its emergent structure. We identify clusters of investors and clusters of patents and observe that their bipartite network is topologically mutualistic, *i.e.* that the structure of the network shares metric characteristics with mutualistic networks in ecology. This is due to the prevalence of investors whose financial incentives make them diversify their portfolios with respect to technological innovations in order to reduce the risks taken [50] and to the existence of a large number of general-purpose technologies, *i.e.* technologies whose use spreads widely across economic sectors [156].

With respect to the robustness of this network, we analyze its nestedness [17] and modularity [21] metrics, as they have been developed by the ecology and physics literature. Nestedness measures the existence of a matryochka-like structure of interaction, where specialist nodes interact with nodes that the generalist nodes also interact with. Modularity estimates the propensity of nodes in a module (a set of nodes allocated to the same group) to interact with nodes in the same module. Both metrics have been linked to the system response to perturbations. We find the investor-patent network to be strongly nested and weakly modular, which is consistent with its topologically mutualistic nature. As a consequence, this network is characterized by distinct responses to different system perturbations [292]. Crises that affect investors randomly or specialized investors tend to have relatively little impact, due to the redundant nested structure of the network, whereas events that target generalist investors tend to present a higher risk for related technological innovations [51].

4.2 Objectives

We study the interactions between investors and the technologies developed by startup companies in their portfolios. To do so, we combine financial and patent data in a network analysis framework. We first present the methodology used to build this bipartite network. We then analyze its topology and, using metrics developed by the literature in ecology and physics, notably nestedness and modularity, we discuss the implications of this topology for its robustness against crises.

4.3 Materials and methods

Further technical and implementation details are provided in the section 4.7.

Figure 4.1 presents the approach followed in this chapter. Using financial and patent databases, we retrieve information concerning startups, the patents they own and the investors that financed them. Investor communities are then identified using a pairwise investor similarity methodology from the literature and patent clusters are identified using Topic Modeling methods. These aggregate-level descriptions are then used to build a bipartite graph linking investor communities to patent clusters by using the startups as the bridge between the two (startups can be linked both to the patents they own and to the investors that financed them). We then use the biadjacency matrix of the network to quantitatively investigate its structure through network-level structural metrics frequently used in the ecological literature.

4.3.1 Datasets

The startup dataset used for this study was extracted through the Crunchbase[1] API on February 14, 2023. It contains information on 2 597 998 startups (name, headquarters location, creation date, sectoral tags), 396 506 funding events (funded startup, date of the funding round, investors involved, funding amount, investment stage), 241 489 investors (name, creation date, investor type, headquarter location) and 1 631 627 individuals (name, professional experiences, academic education, company board memberships and advisory roles). We removed the *Software* and *Other* sectors from the 47 original sectoral tags as they were found to be redundant, highly non-specific and over-represented (representing a combined total of roughly 13% of all tags in the dataset, first and fourth tags in terms of number of occurrences). We filtered out all companies founded before January 1st, 1998 to remove all companies that were not startups and all companies for which geographical information was not available. Funding rounds that were not VC funding (such as debt financing or grants) were also filtered out as they are carried by other actors than VCs. Since we focus on the interactions between investors and technological innovation, using companies as linking agents between both, we filtered out all companies that did not raise funds. After applying these filters, 234 358 companies remained in our dataset.

The patent dataset also supplied by Crunchbase and IPqwery[2] contains a total of 15 713 946 patents from WIPO[3], USPTO[4] and CIPO[5] with their title, abstract, filing date, owner identification, and International Patent Classification (IPC) codes. It provides a matching with the startup dataset that links patent owner IDs to Crunchbase company IDs, allowing us to determine the patent portfolios of startups. We filtered out all patents filed before January 1st, 2000 and all patents that were not owned by companies from our filtered startup dataset,

[1] https://www.crunchbase.com
[2] https://ipqwery.com
[3] World Intellectual Property Organization
[4] United States Patent and Trademark Office
[5] Canadian Intellectual Property Office

resulting in a final dataset of 835 763 patents.

4.3.2 Networks

To study the structure of interactions between investors and technological innovations, we could consider a network where investors interact with technologies based on their investments in startups and on the patents owned by startups : an investor and a patent would simply be linked if the investor has financed the startup owning the patent. At this level of granularity, the interactions between patents and investors are however too sparse to study fundamental behaviors. Since individual patents are known to belong to classes or clusters associated with different fields of technological innovation [2], and since investors belong to different types [60], we cluster investors and patents in order to aggregate them into coarser-grained communities.

Investor communities

We detect investor communities following the methodology described in [60]. We select investors with 60 or more investments bringing the number of investors down to 2017 and, for each of them, build the 5 characteristic distributions as described in [60] : temporal, amount, geographical, series and sectoral investment distributions. We compute a similarity metric between all pairs of investors to build a weighted similarity network where all investors are linked and the edge weights correspond to the pairwise similarities. As all nodes are linked to each other in a pairwise similarity network, the network is then pruned to reduce link density (weak links are removed to transform our highly mixed community structure into a simpler, lowly mixed community structure [166]) in order to run a community detection algorithm [34]. This yields $c = 16$ investor communities as shown in Fig. 4.2A. The community results are in line with those presented in [60], with novel communities emerging as the dataset used was extracted more recently. [60] have shown that the investor communities provided by this methodology are stable with regard to perturbations to the characteristic distributions of individual investors, suggesting that the underlying investor clustering is robust.

Patent clusters

We apply topic modeling to our patent dataset following the methodology described in [125, 61]. In order to thematically cluster similar patents together based on their textual abstracts, we create vector representations (embeddings) of individual patents using the PatentSBERT [22] model specialized in patent modeling. Each patent is thus represented by a 768-dimensional embedding. We then create a low-dimensional representation of all embeddings using parametric UMAP [262]. Using these UMAP vectors in conjunction with HDBSCAN [210], a density-based spatial clustering algorithm, we perform the clustering of individual patents. The HDBSCAN algorithm works in two phases : first, a clustering is performed by identifying regions of high density and grouping the points in these regions

together and a hierarchical approach is then taken to return a flat clustering able to take into account the variable cluster densities. As spatial density-based algorithms perform better when the dimensionality of the data is not too high, the dimensionality reduction step is performed to improve the performance of the HDBSCAN clustering. This process results in $p = 98$ patent clusters. To characterize the patent clusters, we perform a keyword extraction procedure for each cluster using c-TF-IDF, extracting the n-grams that are representative of each patent cluster. Hyperparameters used for the HDBSCAN and UMAP algorithms are provided in Table 4.1.

Investor-patent network

We allocate each startup to the patent cluster most represented in its patent portfolio. Investors that did not invest in any startup that holds a patent and startups that do not own patents are removed, resulting in 1937 investors and 12 007 startups after filtering. We then build the weighted investor-patent network based on the funding events between investors and startups : the weight of the edge connecting an investor community and a patent cluster corresponds to the number of times members of the investor community have invested in startups allocated to the patent cluster. The resulting weighted investor-patent network is bipartite since it is composed of two different classes of nodes, with nodes of one class (in our case, investor communities) being only linked to nodes of the other class (here, patent clusters). Other examples of bipartite networks in socio-economic research include the country-product network [137] or the country-food production network [298]. Specific metrics have been developed to characterize bipartite networks [302] that are presented in the following section.

4.3.3 Network metrics

Nestedness

Nestedness [17] is a structural property of networks that characterizes to what extent specialist species interact with subsets of the species generalist species interact with, meaning that in nested networks, specialist-specialist interactions are infrequent. Mutualistic networks, *i.e.* networks where both species involved in an interaction have a net benefit such as plant-pollinator or seed dispersal networks have been shown to be significantly nested [17]. Possible explanations such as system tolerance to species extinctions have been suggested as a reason for the prevalence of nestedness in mutualistic systems [51], but the origin of nestedness in these networks remains an open question [234]. Furthermore, the nested architecture of networks has been shown to be positively correlated with structural robustness (studied for instance through the lens of species extinction in ecology) when it is assumed that species with lower degree are more at risk of extinction, meaning that nested networks are maximally robust when the least linked species (specialists) become extinct but more fragile when the most linked species (generalists) face systematic extinction [51]. Following up on these findings, [254] have shown that maximally nested networks maximize the structural stability of mutualistic systems and that most observed networks were

close to this optimum architecture.

Nestedness measures have been widely studied as one of the key metrics characterizing interactions in bipartite networks. Measuring it, however, has been a topic of ongoing investigations [235, 47, 7, 299], and a number of different methods have been developed [199]. Among the latter, several nestedness metrics such as the Atmar & Patterson temperature or the overlap and decreasing fill (NODF) [235] do not take into account the quantitative nature of the interaction matrix, reducing it to a binary interaction matrix. Since the nature of our data allows us to access detailed information about the frequency of interaction between nodes, we see that link weights span several orders of magnitudes and therefore opt for the spectral radius ρ [279] metric, a nestedness measure that can handle weighted networks. This is of particular relevance as it has been shown that networks that were thought to be significantly nested in binary form were found not to be nested when accounting for interaction weights [279].

Bipartite modularity

The modularity Q of a network is a structural measure of how frequently nodes in defined subgroups of the network (modules) interact with each other compared to their frequency of interactions with nodes of other subgroups. The adjacency matrix of a modular network thus presents blocks of dense interactions between nodes of a given subgroup, and few links with nodes of other subgroups. Here, we use a modularity measure developed specifically to take into account the bipartite nature of the network of study [21].

Connectance

The connectance of a network is defined as the number of realized links divided by the number of possible links in the network. This structural metric has been shown to be linked to network complexity, degree distribution and network stability [239, 204]. Furthermore, in bipartite ecological networks, the level of connectance of the network has been shown to impact the relationship between modularity and nestedness [108], with low connectance networks displaying a positive correlation between modularity and nestedness and networks with a high connectance value displaying a negative correlation between modularity and nestedness.

4.4 Results

4.4.1 Investor communities

Starting from 2017 investors, we apply the clustering methodology described in the methods section. We obtain 16 investor communities (Fig. 4.2A) described in Table 4.2. The communities are relatively few in number and are heterogeneous in size (the smallest one, *C.13*, is comprised of 19 investors and the largest, *C.04*, is comprised of 239 investors). Com-

munity *C.00* is composed of investors specialized in the *Health Care* sector, *C.01* of historic investors that have been active relatively homogeneously throughout the whole period of study, *C.02* of generalist investors capable of investing different amounts at different stages without displaying a strong sectoral specialization, *C.03* of early-stage cryptocurrency investors that started being active around 2020, *C.04* of United Kingdom (UK) and Germany (DE)-focused investors, *C.05* of late-stage and private equity (PE) investors, *C.06* of early-stage and business angels (BAs), *C.07* of a specific type of very early-stage investors called *accelerators*, *C.08* of Canada-focused early-stage investors and incubators, *C.09* of France-focused investors with a slight preference for the *Health Care* sector, *C.10* of China-focused investors, *C.11* of early-stage actors that started being active around 2013, *C.12* of investors that started being active in 2014 capable of investing throughout all stages of the VC cycle, *C.13* of Latin America (Brazil, Mexico, Colombia)-focused investors, *C.14* of India and Southeast Asia (SEA)-focused investors (India, Indonesia, Singapore), *C.15* of Japan-focused investors.

These investor communities present relatively straightforward identities, with some of them being mainly defined by their geography of investments (*C.10* and *C.14*), others being defined by their sectors of investment (*C.00* and *C.03*), others by their stage of investment (*C.05* and *C.07*), others by their temporal patterns of investment (*C.01* and *C.11*) and others by a combination of several dimensions (e.g. *C.08* with a mix of the stage and geographical dimensions or *C.09* with a mix of the sectoral and geographical dimensions). Previous work [60] has shown these clustering results to be robust to the decimation of the characteristic dimensions used to compute the similarities.

4.4.2 Patent clusters

Individual patents are grouped into clusters (Fig. 4.2B), and we extract keywords and apply labels to describe the resulting patent clusters as shown in Table 4.3. The size of the patent clusters is strongly heterogeneous, with the smallest patent cluster containing 226 patents and the largest cluster containing 82 513 patents. Our patent clusters cover a wide range of specific technologies, the top 5 clusters by number of patents being : cluster 53 (*Pharmaceutical compositions/therapy*) with 82 513 patents, cluster 41 (*Wireless Communication Technology*) with 51 139 patents, cluster 55 (*Image Processing & Autonomous Vehicles*) with 41 507 patents, cluster 4 (*Pharmaceutical Compound Therapy*) with 30 981 patents and cluster 64 (*Semiconductor Device Fabrication*) with 20 700 patents. Examples of other patent clusters include cluster 12 (*Seismic Survey Techniques*, 322 patents), cluster 38 (*Nucleic Acid Analysis*, 14 600 patents) and cluster 15 (*Social Media Content*, 424 patents).

Technologies can be thought of as roughly being linked to three overarching groups : hardware-based, Health Care-related and software-based, each with its own specific challenges and constraints. Patented technologies can of course draw elements from several of these general fields, but we manually allocated each patent cluster to one of the three groups based on their label and keywords. We then colored the patent cluster nodes in Fig. 4.3 following this allocation : *Manufacturing* in green (38 patent clusters), *Information Technology* (IT) in blue (35 patent clusters), and *Health Care* in red (24 patent clusters). Even though there is some heterogeneity in the number of patent clusters in each group, all three

groups are well-represented.

4.4.3 Investor-patent network

A bipartite network linking investor communities and patent clusters (Fig. 4.3) is built using investor communities and patent clusters. The degree distribution of the network is highly heterogeneous (truncated power law, Fig. 4.6) both for investor communities and patent clusters, meaning that a small number of nodes have a large number of connections to other nodes while most others have a low number of connections. Broad-scale networks, which exhibit a truncated power law degree distribution, are commonly found in abiotic and biotic systems, and are the result of finite size effects of the studied underlying network. The tail of the distribution (nodes with high degree) by definition has few observations, and as real processes are often bounded by the constraints of the system (in our case, a finite number of funding events), a bounded distribution is better suited to describe the system.

The biadjacency matrix associated with the bipartite network is reordered following the descending node degree on both investor community and patent cluster nodes (upper-left packing) to show the nested interaction patent of the network (Fig. 4.9). Community *C.02* is the most active investor community, with a fairly diversified patent portfolio. Community *C.00* is the second most active community, with a strong specialization in *Health Care*-related patent clusters (clusters 53, 57, 4 and 38). On the other end of the matrix, communities *C.13* and *C.03* are the least active, with *C.13* showing no specific pattern and *C.03* showing IT and finance-related patent activity (clusters 26, 36, 30 and 96). The nested interaction pattern of the network is visible, with a strong density of interaction in the upper-left corner of the matrix and few interactions in the lower-right corner of the matrix. We visually observe that the nested structure, when rearranging the biadjacency matrix by descending order of degree, is imperfect in part due to the specificity of community *C.00*. Indeed, since we work with quantitative rather than binary data, it boasts both a high degree (high number of interactions) and a high specialization (relatively few patent clusters with which it interacts).

4.4.4 Connectance

The measured connectance of our network is $C = 0.72$, a high value compared to ecological bipartite networks (Fig. 4.7B). As there are in theory no forbidden interactions (interactions that are structurally impossible in a network for physiological or phenological reasons) in our system and as our study covers a long period of time, this high value is not surprising. The magnitude of the connectance has strong implications on other structural network metrics such as degree distribution, nestedness and modularity.

4.4.5 Modularity

Using a modularity-based community detection algorithm, we measure the modularity value of our network and retrieve 4 modules. The measured modularity of our network is $Q_m = 0.19$, meaning that the network is weakly modular (modularity ranges from -1 to 1, with negative values corresponding to anti-modular networks and positive values to modular networks). The 4 retrieved modules are shown by text color for patent cluster nodes and investor communities in Fig. 4.3. Three of these modules show a strong technological focus (module 1 around *Manufacturing*, module 2 around *Information Technologies* and module 3 around *Health Care*), with the fourth one (module 0) containing a mix of technologies. A matrix-based view of the biadjacency matrix reordered by modules is shown in Fig. 4.10. We also computed the normalized modularity $\overline{Q}$ of a number of bipartite ecological networks (Fig. 4.7A) and compared it to the normalized bipartite modularity of our network $\overline{Q}_m$. We find that our modularity is lower than most networks it was compared to, potentially due to the different underlying nature of this socio-economic network compared to ecological networks.

4.4.6 Relevance tests and ecological metrics

Relevance tests for the nestedness and modularity of our network are performed, and the investor community-patent cluster network is found to be significantly more nested (Fig. 4.4A, $\rho_m = 5662$, $\rho_{null} = 3799 \pm 295$, mean $\pm$ std, $z_\rho = 6.32$) and significantly less modular (Fig. 4.4B, $Q_m = 0.19$, $Q_{null} = 0.61 \pm 0.02$, $z_Q = -21$) compared to the null model. This specific network topology has strong implications on the properties of the network, notably in terms of robustness to external perturbations such as species extinction and in terms of species diversity. The statistically high nestedness and low modularity (compared to the null models) of the interaction structure between investor communities and patent clusters is in line with previous findings in the literature as nestedness and modularity have been shown to be anticorrelated for networks with high connectance [108]. We also perform this analysis on a network where interactions are weighted by total funding amounts rather than number of interactions, and, even though the ordering of investor communities by degree is different, we find similar results (Fig. 4.11 and 4.12) in terms of nestedness and modularity.

4.5 Discussion

We observe that the bipartite network of interactions between startup investors and the patents in which they indirectly invest exhibits an emergent topological mutualism. This mutualistic topology, commonly found in ecological systems, results here from the presence of many investors whose generalist nature is induced by their portfolio diversification strategies, on one side of the network, and of general-purpose technologies with a broad spectrum of use, on the other. On the investor side, portfolio diversification is a fundamental and basic idea of modern portfolio theory [200]. Investors are statistically better

off if and when they diversify the risks they take among their investments: typically here, by supporting startups that develop different kinds of technological innovations. On the other side of the network, general purpose technologies [44] are technologies characterized by their pervasiveness because they are used as inputs by many downstream sectors: here, by numerous startups with many different products that benefit from investments from many different types of investors. Although we do not address here the complexity of economies properly speaking and notably trade networks [53], this result is closely related to the studies that have shed new light on the workings of socio-economic systems within the framework of economic complexity, and shares with them the rationale according to which the understanding of many societal issues implies to look at the systemic interactions that produce them [136, 15]. In addition, but with a less pronounced relevance with respect to our approach, it should also be noted that another line of investigation has also attempted to draw parallels between the study of mutualistic systems and economic issues by introducing economic market effects to explain the evolution and stability of mutualistic interactions in ecological systems [229, 48].

Topologically mutualistic networks have been shown to be significantly nested [17], a property that has been related to network robustness both for socio-economic systems and in ecology. Both literatures concur that the observed nested structure of the bipartite matrices describing topologically mutualistic networks contributes to their robustness and stability [135, 199] and study the vulnerabilities of such systems, with the general conclusion that a nested system reacts very differently to perturbations depending on which types of nodes they affect.

Studies of the bipartite network of interactions between designer and contractor firms in the New York City garment industry [260, 261], following Uzzi's seminal work [300] on the sources and consequences of embeddedness [122] have typically highlighted the fact that since the nested architecture of mutualistic networks implies that nodes contribute heterogeneously to their vulnerability, the removal of some nodes that especially contribute to the global nestedness of the network is consequentially associated with stronger vulnerabilities. In a similar vein, [135] have studied the Boulogne-sur-Mer Fish Market, focusing on the bipartite interactions between buyers and sellers, and studied its resistance to perturbations that would affect high-degree buyers or sellers with the conclusion that the auction part of this market was more robust. The theory of economic complexity has concurrently associated nestedness and the dynamics of industrial ecosystems [53], notably in relation to the resilience of economies to external shocks [133, 15].

In ecology, where the nested network structure tends to minimize competition between species and support greater biodiversity [18], maximizing structural stability [254], perturbations impacting generalist species have been shown to lead to faster species depletion at the network level [51] by isolating specialist species due to the nested structure.

Furthermore, and with respect to the stability of nested bipartite networks, ecological studies have further shown that nested interaction networks emerge by considering an optimization principle aimed at maximizing species abundance [285] and that nested mutualistic interactions boost equilibrium population densities and increase the resilience of communities [283] : typically, when analyzing the short-term dynamics following a strong population perturbation, mutualist networks are associated with an ability to replete af-

fected communities when species numbers fall dangerously low [283].

With respect to the investor-patent network, the attrition of generalist communities of investors should therefore be associated with a potentially severe impact on the entire network, an impact that could put technological fields and the associated emerging technologies at risk. Perturbations targeting generalist investors such as communities *C.02* (investors capable of investing different amounts at different stages without a strong sectoral specialization) and *C.06* (early-stage and business angels) stand out as the highest vulnerabilities, which highlights the role that these investors play for the system as a whole and for the diversity of technological fields of innovation that receive funding. In addition to these two communities, using funding amounts to weight the bipartite network links instead of the number of interactions (see Fig. 4.12), community *C.05* (late-stage and private equity investors) also stands out as the most generalist investor community in this nested bipartite network, an observation which could be of special relevance since this community has been subject to a decrease in activity since 2022 [240, 70, 106]. Such a perturbation, as it affects a community crucial to the nested patent-investor network, could be expected to have not only quantitative consequences, as is generally foreseen, but also qualitative ones, putting fields of technological innovation at risk.

In parallel, investor communities such as *C.00* (investors specialized in the *Health Care* sector) and *C.03* (early-stage cryptocurrency investors) act as specialist investors in the bipartite network, whose emergence could be related to the need for specific skillsets in these sectors: for instance, the Health Care sector (that corresponds here to module 3 in Fig. 4.3) is well-known to be associated with very specific regulation and R&D cycles.

A relevant question here is whether the number of technological and sectoral specificities will increase in relation to the development of commonly named "deep techs", a category that includes medtechs, quantum computing or artificial intelligence, each of which is associated with specific and emerging regulations. Such a phenomenon could lead to an increased number of specialist investor communities, which would in turn increase the modularity of the patent-investor network, a phenomenon negatively correlated with network robustness as mentioned above but which could also mitigate perturbations affecting key generalist investors. A more modular structure of the bipartite patent-investor network could result in fields of technological innovations being dependent on a limited number of investor communities for their funding, but conversely less dependent upon generalist investors.

Finally, and with respect to future research, startup databases are known to underrepresent early-stage funding rounds compared to later-stage ones due to an easier tracking of the latter. Although we do not expect such a bias to affect the results of our study, complementary analyses on reduced but more exhaustive datasets could further clarify this issue. In addition, VC funding tends to target novel, emerging and potentially disruptive technologies, while others are funded by a more varied panel of investments which could also warrant more comprehensive investigations, notably innovations that spin-off from academia in the context of its specific set of institutions and incentives [86]. Further studies could also attempt to directly assess the robustness of the investor-startup network with respect to technological diversity [137, 333] when facing different types of crises.

4.6 Conclusion

In this work, by leveraging large-scale financial, startup and patent datasets, we have built a bipartite network directly linking investors and technologies. Using network metrics, we have found this network to display an emergent topological mutualism associated with a heterogeneous degree distribution, a significant nestedness and a significantly lower modularity compared to null models. This has relevant implications for the robustness of the ability of startups and investors to support technological innovation when facing crises. We notably expect the system to react differently depending on perturbations. In particular, perturbations affecting investor communities that contribute strongly to the nestedness of the patent-investor network could have a far-reaching impact on technological innovation.

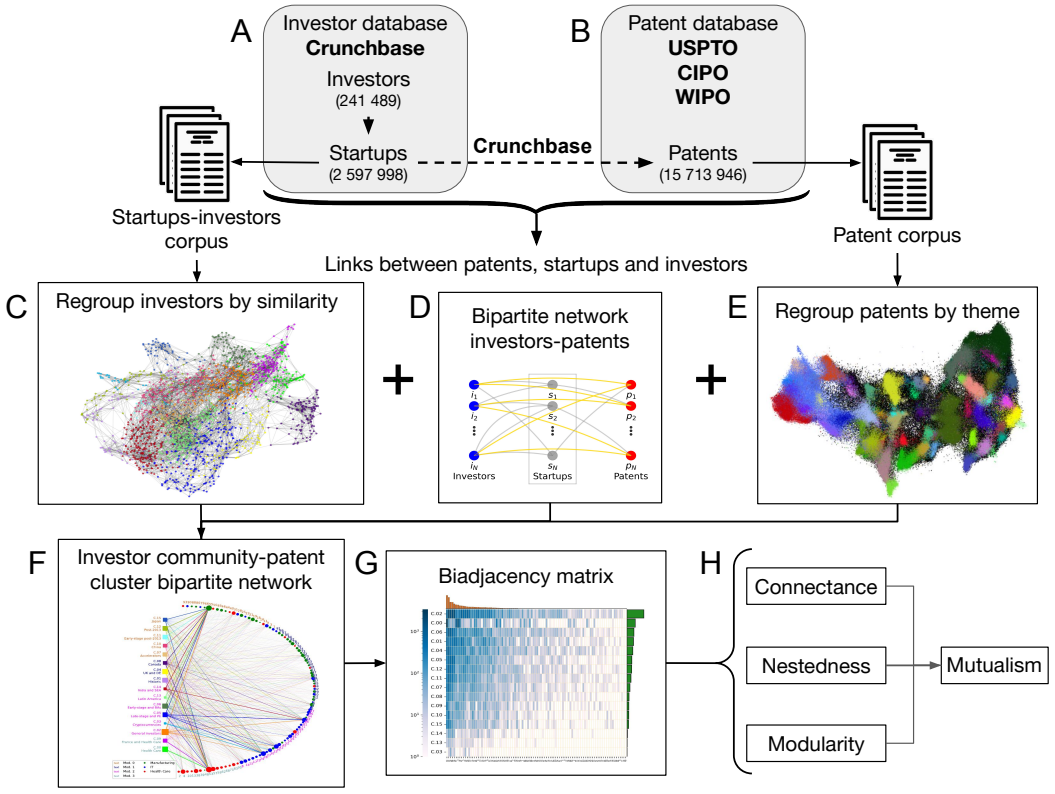

Figure 4.1: **Workflow presenting the approach followed in this chapter. (A)** Investor and company information is extracted from Crunchbase. **(B)** Patent data from USPTO, CIPO and WIPO is extracted and matched with the Crunchbase company information. **(C)** 16 investors communities are retrieved using a similarity metric between pairs of investors. **(D)** The bipartite network between investors and the patents of the companies they invested in is built. **(E)** NLP-based topic modeling of patents is performed on their abstracts and 98 patent clusters are retrieved. **(F)** The investor community-patent cluster graph is built based on the investor-patents bipartite network by combining the results from steps **(C)**, **(D)** and **(E)**, *i.e.* by aggregating investors into their investor communities and patents into their patent clusters on the investor-patents bipartite network. **(G)** The biadjacency matrix of the investor community-patent cluster graph is extracted to quantitatively visualize the interaction patterns and compute network structure metrics. **(H)** Network structure metrics (connectance, nestedness, modularity) are computed using the biadjacency matrix to study the topology of the network and the properties deriving from it.

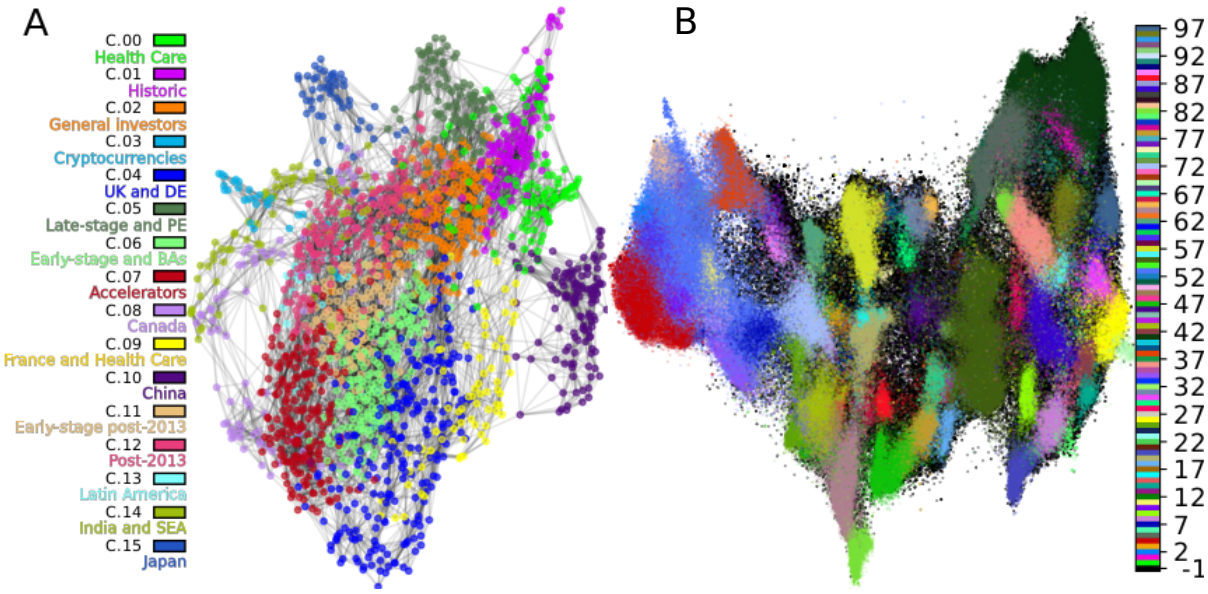

Figure 4.2: **Investor communities and patent clusters. A.** Pruned investor similarity network. Each node corresponds to an investor, and its color corresponds to the investor community it is allocated to. All investors can be grouped into 16 communities that define types of "investor species" (C.00 to C.15, presented in Fig. 4.3 and in Tab. 4.2). **B.** Projected latent space (2 dimensions) of the patent data. Each point represents a patent and its color corresponds to its cluster allocation. The clustering defines 99 clusters, 98 thematic clusters and one unlabeled cluster. Cluster -1 (in black) corresponds to unlabeled data points.

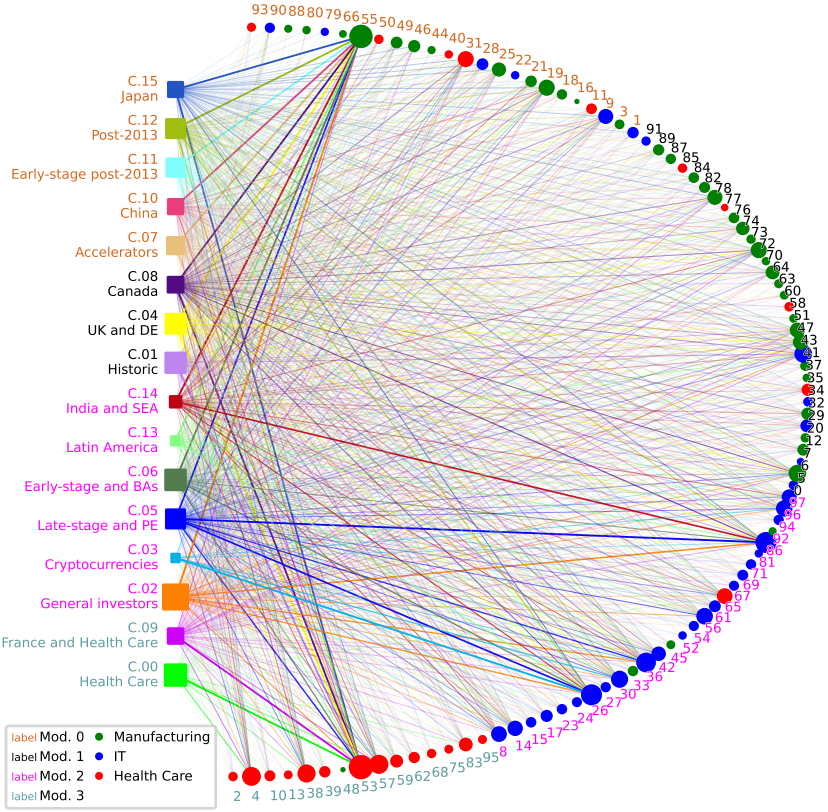

Figure 4.3: **The investor community-patent cluster bipartite network.** Square nodes represent investor communities and circle nodes patent clusters. Node sizes are a function of the node degrees. Link weights are normalized for each investor community by the maximum edge weight of the investor community, and the edge width shown is the logarithm of the normalized weight. A brief description of investor communities is provided under each investor community label, and a more extensive description is available in Table 4.2. Nodes were positioned following the 4 modules obtained by the bipartite modularity algorithm, and node label colors correspond to the module they were allocated to. Patent clusters are colored following a manual allocation of the high-level technological field they deal with (red for *Health Care*, blue for *Information Technology*, green for *Manufacturing*).

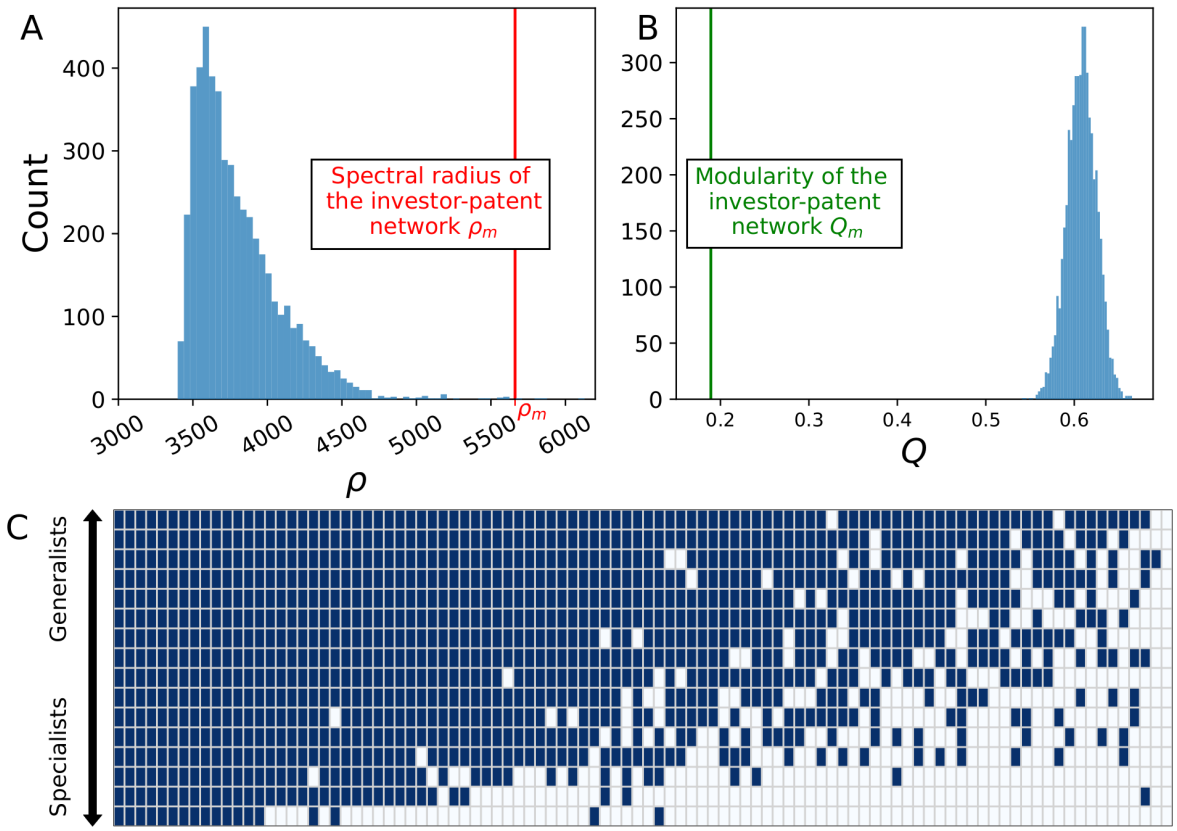

Figure 4.4: **Statistical relevance tests for the nestedness and the modularity of the network. (A)** Statistical relevance test for the nestedness ρ_m (red vertical line) of the investor community-patent cluster network compared with $5\,000$ iterations of the null model (blue histogram) described in the Appendix. We see that our network is significantly more nested compared to networks generated by the null model. **(B)** Statistical relevance test for the modularity Q_m (green vertical line) of the investor community-patent cluster network compared with $5\,000$ iterations of the null model (blue histogram). We see that our network is significantly less modular compared to networks generated by the null model. **(C)** Binarized representation of the biadjacency matrix. Investor communities correspond to the rows, patent clusters to the columns. The rows and columns are reordered by descending marginals (sums of the value of the row or column), yielding an upper-left packed matrix. The nested structure is displayed, with more specialist investor communities (bottom rows of the matrix) mostly interacting with a subset of the patent clusters the generalist species (top rows of the matrix) interact with.

4.7 Appendix

Methods

Investor-patent networks

To study the structure of investor-technology interactions, we first need to build a network where investors interact with technologies. As startups own patent portfolios and investors own startup portfolios through their investments, a link can be drawn between an investor and a patent if the investor has financed the startup owning the patent (Fig. 4.5). Since we wish to study investor-patent interactions, we filter out all patents that are not linked to startups that have raised funds. The number of patents owned by funded startups in our database and the number of investors, however, are large, with $835\,763$ patents and $113\,934$ individual investors remaining after filtering. This presents several problems, both from a computation and statistical point of view: this graph is too large to apply the usual matrix-based metrics, and the interactions between patents and investors are too sparse to study fundamental behaviors. To remediate this, we propose to separately cluster the investors and the patents to create coarser-grained communities, strengthening signal by grouping similar investors together (investor communities) and similar patents together (patent clusters).

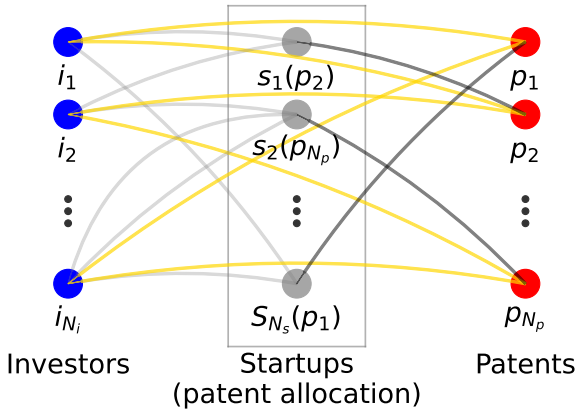

Figure 4.5: **Defining the network of investors and patents.** Investors are represented by a blue node and their investments by a link (grey lines) to the startup nodes (grey dots). The nodes of the startups are linked (dark grey lines) to the patents they own represented by red nodes. By transitivity, the investors are linked to these patents (yellow lines). The network is defined by the set of investor nodes linked through their investments to the set of patent nodes. This forms a possible bipartite network between investors and patents. N_i, N_s and N_p represent the number of investor, startup and patent nodes respectively.

Patent clusters

The IPC (International Patent Classification) is unfortunately not easily amenable to network analysis such as ours for several reasons [284, 306]. First, it is a fixed classification that is infrequently updated even though patents are being published increasingly fast, creating a grey area where new technologies are not accurately categorized due to the lack of a sufficiently suitable class amongst the existing ones. Second, a patent is a complex document that contains different aspects such as technological descriptions or fields of application. Some patents thus exist at the intersection of several classifications (a patent describing a recommender system for targeted advertising, for instance, exists in both the recommender system space and the targeted advertising space), which the IPC resolves by allowing the patent to have multiple classes. This is, however, problematic when trying to quantitatively study patent data as multi-class analysis is markedly more complex than binary analysis. Third, the patent class taxonomy is massive with over 70 000 subgroups; a method to reduce the number of categories is necessary in order to obtain a smaller graph so that relevant and easily interpretable analyses can be performed. Doing so by cutting the taxonomy closer to the root would be a potential way to proceed, but this would lose information as a patent can exist in different branches of the taxonomy. Finally, there is some inconsistency in the classification between different patent offices and countries [32, 62], where a patent will not necessarily have the same classification between the different offices and examiners. We thus endeavor to develop a patent classification method in order to homogenize the patent classification for our study.

We therefore apply topic modeling to our patent dataset, following the methodology described in [125, 61]. We first concatenate the title and abstract for all patents and embed the concatenated string using the PatentSBERT [22] which is a state-of-the-art language model specialized in patent modeling. This results in 768-dimensional embeddings for each patent. We then create a low-dimensional representation of all the embeddings using the parametric UMAP algorithm [262]. This algorithm differs from regular UMAP as it trains a neural network using a subset of the total dataset to learn the high-to-low dimensional mapping and uses this model to reduce the dimensionality of the complete dataset. This is necessary as the number of patents in the dataset makes it otherwise too large to use regular UMAP. HDBSCAN [210] is then used to perform the clustering based on the UMAP vectors of the patents. The dimensionality reduction step is performed to improve the performance of the HDBSCAN clustering and has the added benefit of reducing the memory requirements of the pipeline. The hyperparameters used for the algorithms are shown in Tab. 4.1.

Parametric UMAP parameters			HDBSCAN parameters			
n_components	n_neighbors	sample_size	min_samples	cluster_epsilon	min_cluster_size	cluster_selection_method
10	3	0.3	1	0	220	“eom"

Table 4.1: **Hyperparameters used for the parametric UMAP and HDBSCAN algorithms.**

Finally, as one cluster is much larger than the rest (over 3 times larger than the second largest) and relatively high-level (containing all organic and Health Care-related technologies), we run the clustering algorithm on the patents in this cluster with the same param-

eters. This process results in $p = 98$ patent clusters found by our process. Colors in the related figure in the main text of the chapter (Fig. 2) were computed using the distinctipy Python library [251].

To characterize the patent clusters, we perform a keyword extraction procedure for each cluster. We compute, for all n-grams present in the cluster, the c-TF-IDF score as defined in eq. 3.15 and we extract the top k n-grams in terms of the c-TF-IDF score. We then add an additional step to remove variations of the same keyword in a given cluster (*i.e.* "power" and "powers"). To do so, we compute the $3 \times k$ top n-grams in terms of c-TF-IDF score and iterate through the list in descending order of c-TF-IDF-score, keeping n-grams if their Levenshtein similarity with all other selected n-grams (*i.e.* with higher c-TF-IDF scores) is below 0.8 until we reach the desired number k of top distinct n-grams.

Building the investor-patent network

We apply a filter by removing all investors that are only linked to companies that do not own patents and by removing all startups that do not own patents, yielding a bipartite network with 1937 investors and $12\ 007$ startups. We allocate each startup to a patent cluster by choosing the patent cluster that is the most represented in its patent portfolio (*i.e.* if a startup has 2 patents in patent cluster 0 and 5 patents in patent cluster 1, the startup will be allocated to patent cluster 1; in the event of a tie, the startup is randomly allocated between all tied clusters). This is done for several reasons : first, a company can own patents that are not truly representative of its activity, with some patents being strategically filed to block rival companies rather than to exploit the patented innovation. A patent can also be a refiling of a patent previously detained by the company, thus artificially inflating the number of patents a company has in a given patent cluster. Furthermore, as some startups own a large number of patents whereas others own much fewer patents, this allows us to focus on investment decisions made by investors without introducing a bias resulting from startup patent portfolio sizes.

We then build an investor-patent network where, for each investment, an investor is directly linked to a patent cluster following the patent cluster allocation of the startup that received funding. The investor-startup (and by extension the investor-patent) network naturally gives rise to a bipartite structure as funding interactions can only link startups and investors.

We then cluster investors together following investor communities, resulting in a bipartite network linking investor communities on one side with patent clusters on the other. Colors in the related figure (Fig. 3) in the main text of the chapter were computed using the distinctipy Python library [251].

Network metrics

Nestedness

We use the spectral radius [279] to measure nestedness as our network has a quantitative structure. Large dominant eigenvalues have been shown to be associated with highly nested structures for both binary and quantitative matrices [279], and the spectral radius ρ (eq. 4.1) of a weighted network is defined as the largest eigenvalue of the weighted adjacency matrix $\mathbf{W}$ of the network.

$$\rho(\mathbf{W}) = \max\{|\lambda_1|, ..., |\lambda_n|\} \tag{4.1}$$

where $\{|\lambda_1|, ..., |\lambda_n|\}$ correspond to the real part of the eigenvalues of $\mathbf{W}$.

In order to compute the spectral radius of the network studied, we compute the weighted adjacency matrix $\mathbf{W}_{n\times n}$ with $n = c + p$ where c is the number of top nodes (investor communities) and p the number of bottom nodes (patent clusters). $\mathbf{W}$ is block-diagonal, *i.e.*

$$\mathbf{W} = \begin{pmatrix} 0_{c\times c} & \tilde{\mathbf{W}}_{c\times p} \\ \tilde{\mathbf{W}}^T_{p\times c} & 0_{p\times p} \end{pmatrix} \tag{4.2}$$

where $\tilde{\mathbf{W}}$ is called the *biadjacency matrix* of the network and $\tilde{\mathbf{W}}^T$ denotes its transpose.

Bipartite modularity

Following [21], community detection on the bipartite network can be performed by finding a configuration of node allocations into modules (sets of investor community nodes and patent cluster nodes) maximizing bipartite modularity Q as defined in eq. 4.3.

$$Q = \frac{1}{M}\sum_{u=1}^{c}\sum_{v=1}^{p}(\tilde{\mathbf{W}}_{uv} - \frac{y_u z_v}{M})\delta(g_u, g_v) \tag{4.3}$$

where M is the sum of edge weights, y and z the row and column marginals of $\tilde{\mathbf{W}}$, g_u and g_v the community allocation of nodes u and v, and δ a Kronecker delta *i.e.* $\delta(g_u, g_v) = 1$ if $g_u = g_v$ (nodes u and v are allocated to the same module) and $\delta(g_u, g_v) = 0$ if $g_u \neq g_v$.

As the amplitude of modularity is dependent on network parameters such as the network size and fill, the maximum modularity value Q_{max} can be computed for each network following eq. 4.4. We then compute, for each network, the normalized modularity $\overline{Q}$ [21] as defined in eq. 4.5.

$$Q_{max} = \frac{1}{M}(M - \sum_{u=1}^{c}\sum_{v=1}^{p}\frac{y_u z_v}{M})\delta(g_u, g_v) \tag{4.4}$$

$$\overline{Q} = \frac{Q}{Q_{max}} \tag{4.5}$$

Connectance

The connectance C of a network is defined as the number of realised links over the number of potential links as shown in eq. 4.6 for a bipartite network.

$$C = \frac{\sum_{u=1}^{c} \sum_{v=1}^{p} K(\tilde{\mathbf{W}}_{uv})}{c \times p}$$
$$K(\tilde{\mathbf{W}}_{uv}) = \begin{cases} 1 & \text{if } \tilde{\mathbf{W}}_{uv} \geq 1 \\ 0 & \text{otherwise} \end{cases} \tag{4.6}$$

Null models

Two different null models were tested to study the statistical significance of the measured nestedness and modularity. Null model 1 consists of a random rewiring of the weighted edges of the graph. Edges are randomly shuffled, and both ends are wired to pairs of nodes chosen at random. A constraint is applied such that no node remains unlinked. Null model 2, following [279], keeps the interaction structure of the biadjacency matrix while randomly shuffling the interaction weights, thus preserving the qualitative structure of the matrix but changing its quantitative structure. Both null models showed similar results in terms of spectral radius and modularity values, statistical results presented in this chapter were obtained using null model 1.

Additional results

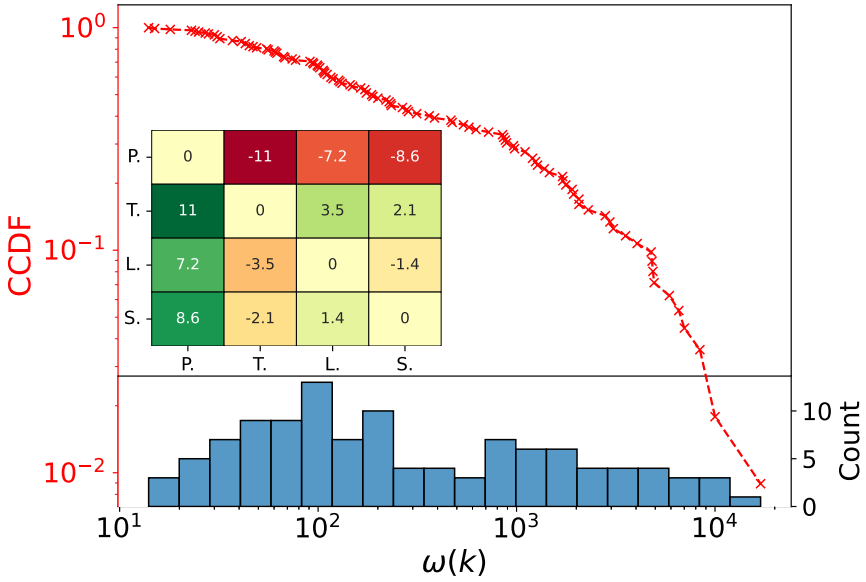

Figure 4.6: **Complementary Cumulative Distribution Function (CCDF, in red) of the degree distribution of the bipartite investor community-patent cluster network.** $\omega(k)$ is the degree of node k. The histogram shows the degree distribution, and the inset heatmap shows the most likely distribution when comparing pairs of candidate distributions (P. stands for Power Law, T. for Truncated Power Law, L. for Lognormal, S. for Stretched Exponential). All non-zero values shown are statistically significant values (*i.e.* $p \leq 0.05$), and the cells of the matrix correspond to the value of the R parameter. Positive values mean that the row candidate degree distribution is more likely than the column candidate degree distribution. The significance analysis was performed using the powerlaw package [8].

Figure 4.6 shows the Complementary Cumulative Distribution Function (CCDF) of the investor-patent network (red curve), statistical comparisons between several candidate distributions (inset heatmap) and the degree distribution (bar plot). The inset heatmap shows that the most likely distribution explaining the degree distribution is a truncated power law, which is a power law coupled with an exponential cutoff that can be due to finite size effects [52] induced by the limited number of investor communities and patent clusters in our samples. Several real-world natural and socio-economic networks display truncated power-law degree distributions [76, 30].

Panels A and B of Fig. 4.7 show two structural network metrics (normalized modularity and connectance) computed for all networks with over 20 species in the Web of Life database (histogram) and for the investor community-patent cluster network (magenta vertical lines). 253 of the original 300 networks in the Web of Life remain after filtering. Panel A shows the distribution of the normalized modularity $\overline{Q}$. The investor community-patent cluster network has a normalized modularity of $\overline{Q}_m = 0.26$, lower than almost all other networks. The modularity values are normalized (see the "Modularity" section in the Ap-

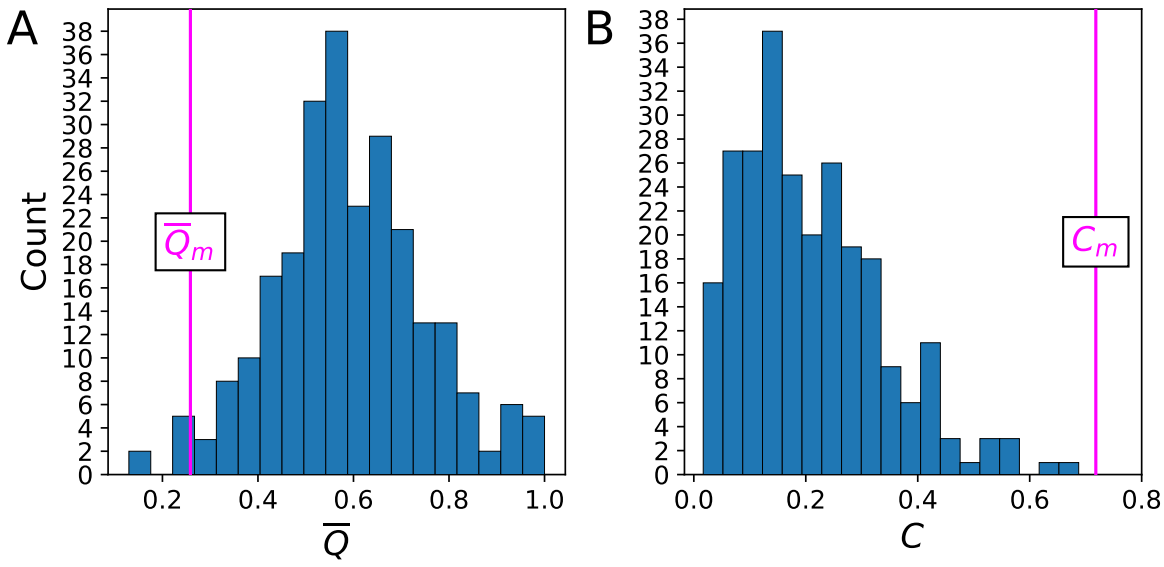

Figure 4.7: **Comparison with ecological networks for the normalized modularity and the connectance.** Ecological networks were extracted from the Web of Life database (https://www.web-of-life.es/), and only networks with 20 species or more were kept for this analysis. **(A)** Normalized modularity $\overline{Q}$. The normalized modularity of the investor community-patent cluster network ($\overline{Q}_m$, magenta vertical line) is compared to the normalized modularity of ecological networks (blue histogram). **(B)** Connectance C. The connectance of the investor community-patent cluster network (C_m, magenta vertical line) is compared to the connectance of ecological networks (blue histogram).

pendix) to allow comparison between networks of different topologies. Panel B shows the distribution of the connectance. We see that most networks have a low connectance value, with very few displaying a connectance over $C = 0.5$. The investor community-patent cluster network has a connectance of $C_m = 0.72$, higher than all the other networks in the database. This network can thus be said to have a high connectance in the context of the ecological literature.

Figure 4.8 shows the frequency of bipartite motifs in the investor community-patent cluster graph. Inferring the interaction type from the bipartite network structure has recently been a topic of interest in ecological communities, notably due to the public availability of a significant number of well-characterized bipartite ecological networks. Studies have found that this task is not straightforward [215] and that the network-level metrics such as nestedness or modularity might not be enough to discriminate between the different interaction types. Pichon *et al.* [238] proposed a multi-scale approach using both network-level metrics and network motifs to better differentiate antagonistic networks from mutualistic networks. Figure 4.8, following [238], shows the results of the bipartite motif frequency analysis performed on the investor community-patent cluster graph. The bipartite motif frequencies were computed using the bmotif package [272], and the square root of the motif frequency is shown for each motif ID. The bottom panel of the figure shows the shape of each motif ID. Comparing these motif frequencies with those shown in [238] does not lead to a strong conclusion as the motif frequencies of networks shown in [238] for the different

interaction types and ecologies show ample variations, and do not allow us to strongly favor an interaction type over the other based on the motif structure of our network. Further research would be needed in this direction.

Figure 4.9 shows the biadjacency matrix of the investor community-patent cluster graph reordered in descending order of degree for both rows and columns. Community *C.02* is the most active investor community, with a fairly diversified patent portfolio. Community *C.00* is the second most active community, with a strong specialization in *Health Care*-related patent clusters (clusters 53, 57, 4 and 38). On the other end of the matrix, communities *C.13* and *C.03* are the least active, with *C.13* showing no specific pattern and *C.03* showing IT and finance-related patent activity (clusters 26, 36, 30 and 96). The nested interaction pattern of the network is visible, with a strong density of interaction in the upper-left corner of the matrix and a few interactions in the lower-right corner of the matrix. We visually observe that the nested structure is imperfect, in part due to the specificity of community *C.00*. Indeed, since we work with quantitative rather than binary data, it boasts both a high degree (high number of interactions) and a high specialization (relatively few patent clusters with which it interacts). Patent clusters with the highest degrees (top rows of the matrix) are found to be technologies with a large number of applications (general purpose technologies), such as *Pharmaceutical compositions/therapy*, *Image Processing & Autonomous Vehicles* or *Payment & Transaction Systems*. This is not surprising as some aspects of these trending technologies see active development and applications from a wide range of actors, leading to a larger number of patents.

Figure 4.10 shows the reordered and normalized biadjacency matrix of the investor-patent network. Rows and columns are reordered to pack modules together (shown with the red rectangles), and patent clusters are colored following a manual allocation between 3 potential labels (*Health Care*, *Information Technology*, *Manufacturing*). We see that the first module containing communities *C.00* and *C.09* is strongly centered around *Health Care*-related patent clusters, the second module containing communities *C.02*, *C.03*, *C.05*, *C.06*, *C.13* and *C.14* is strongly centered around *Information Technology*-related patent clusters, the third module containing communities *C.01*, *C.04* and *C.08* is strongly centered around *Manufacturing*-related patent clusters and the fourth module has more of a mixed technological focus. These results are coherent with the identification of investor communities shown in Tab. 4.2, with communities *C.00* and *C.09* showing a *Health Care* focus, communities such as *C.03* (cryptocurrency investors), *C.13* and *C.14* (emerging regions) being allocated to the *Information Technology*-focused module and historic investors such as those of community *C.01* that have been active since the beginning of our period of study (coinciding with the strong focus on financing hardware startups following the crash of the dotcom bubble in the early 2000s) being allocated to the third module.

Investor community	# of investors	Brief community description
C.00	110	Health Care investors
C.01	132	Historic investors
C.02	200	Generalist investors active whole period
C.03	27	Cryptocurrency investors
C.04	239	EU-focused investors (UK and DE)
C.05	102	Late-stage investors and PE
C.06	236	Early-stage and BAs
C.07	189	Accelerators
C.08	80	Canada-focused investors
C.09	71	France-focused Health Care-focused investors
C.10	122	China-focused investors
C.11	158	Early-stage post-2013 investors
C.12	201	"New-generation" post-2013 investors
C.13	19	Latin America-focused investors
C.14	78	India and SEA-focused investors
C.15	53	Japan-focused investors

Table 4.2: **Description of investor communities. UK** stands for *United Kingdom*, **DE** for *Germany*, **PE** for *Private Equity*, **BA** for *Business Angel*, **SEA** for *Southeast Asia.* "Historic" investors are investors that have been active for a long period of time, since the late 1990s-early 2000s. "Generalist" investors are investors that do not display a significant sectoral focus, investing in all types of sectors and related technologies. "Cryptocurrency" investors are investors strongly specialized in cryptocurrencies and related financial sectors. "Late-stage" investors focus on the later stages of VC financing (series B and onwards), typically investing very large amounts. "Early-stage" investors focus on early stages of VC financing (pre-seed, seed and series A), investing relatively small amounts. "Business Angels" are individuals who invest their own money in startups, usually in early-stage rounds and low amounts. "Accelerators" are a specific type of early-stage investors that usually operate by selecting batches of companies for a short period, providing them with small amounts of money and an intensive mentoring program of a few months focused on developing specific aspects of the company. "Post-2013" investors are investors that started being active (or greatly increased their activity) around the 2013 period, where VC financing experienced sudden and significant growth.

patent cluster	# of connections	cluster label	1-grams	2-grams
0	74	Video Displaying Technology	video | format | stream | frames | media	video stream | video data | video content
1	186	Location-based Wireless Technology	location | wireless | positioning | mobile | satellite	mobile device | mobile station | wireless device
2	62	Cancer Treatment Therapies	cancer | treating | inhibitor | combination | treatment	treating cancer | methods treating | combination therapy
3	77	Fluid Valve Assembly	valve | piston | fluid | chamber | pressure	valve assembly | pressure tube | shock absorber
4	2801	Pharmaceutical Compound Therapy	compounds | formula | thereof | derivatives | diseases	pharmaceutical compositions | pharmaceutically acceptable | compounds formula
5	1938	Power Electronics Circuit	power | voltage | circuit | output | signal	power supply | clock signal | input signal
6	15	Wireless Network Technology	network | wireless | access | mobile | service	wireless network | network access | wireless device
7	232	Chemical Reaction Engineering	catalyst | process | gas | stream | carbon	gas stream | carbon dioxide | stream comprising
8	1104	Multimedia Streaming Services	media | content | video | playback | audio	media content | video content | playback device
9	862	Speech Processing Technology	audio | speech | sound | voice | microphone	audio signal | audio data | speech recognition
10	149	Pharmaceutical Formulations & Dosage	pharmaceutical | formulations | composition | release | oral	pharmaceutical composition | pharmaceutically acceptable | dosage form
11	132	Microbial Acid Production	acid | production | microorganisms | microbial | amino	amino acid | method producing | non naturally
12	60	Seismic Survey Techniques	seismic | sensor | measurement | subsurface | acoustic	seismic data | seismic trace | acoustic signal
13	41	Virus, Vaccine, Antigen	virus | vaccine | protein | recombinant | vectors	present invention | invention relates | nucleic acid
14	966	Online Advertising Services	advertisement | advertising | ad | campaign | advertiser	advertising campaign | advertising content | web page
15	119	Social Media Content	content | item | social | user | online	content item | social networking | social media
16	1	Vehicle Braking Systems	brake | disc | caliper | disk | lining	disc brake | brake disc | brake caliper
17	266	Cloud Computing Services	application | computing | cloud | software | service	cloud computing | computing environment | virtual machine
18	106	Augmented Reality Displays	light | display | pixel | image | eye	display device | light emitting | image light
19	1247	Motor Vehicle Assembly	rotor | motor | shaft | assembly | vehicle	steering column | aerial vehicle | motor vehicle
20	363	Video Compression Technology	video | block | coding | prediction | picture	video data | video coding | motion vector
21	200	Agricultural Management & Yield	crop | agricultural | yield | plant | field	agricultural field | crop yield | management zones
22	31	Location-based Tracking Technology	location | mobile | tracking | geo | fence	mobile device | tracking device | location information
23	111	Multi-Tenant Database	database | application | custom | tenant | object	access permissions | multi tenant | mechanisms methods
24	91	Network Management & Security	network | traffic | service | policy | proxy	communication network | network element | network agent
25	625	Battery Electrochemistry Technology	fuel | electrolyte | battery | anode | electrode	fuel cell | lithium ion | active material
26	4869	Payment & Transaction systems	payment | transaction | merchant | card | account	point sale | systems methods | mobile device
27	102	Web Page Management	web | page | content | tab | user	web page | context menu | tabs tab
28	220	Neural Network Technology	neural | training | network | input | output	neural network | convolutional neural | training neural
29	314	Ultrasound Medical Imaging	ultrasound | imaging | tissue | image | ultrasonic	ultrasound imaging | ultrasound device | ultrasound data
30	1713	Identity Authentication Technology	authentication | key | identity | user | server	user authentication | public key | private key
31	1204	Medical Monitoring Devices	physiological | patient | monitoring | heart | blood	heart rate | blood pressure | vital signs
32	68	Software Test Platform	application | software | platform | test | file	live multi | multi tenant | sdk platform
33	111	Footwear Assembly Tools	tubular | upper | footwear | portion | projectile	tubular element | projectile casing | entangling projectile
34	285	Polymer Composition Formulations	composition | polymer | weight | comprising | containing	composition comprising | invention relates | present invention
35	32	Heat Dissipation Technology	heat | cooling | air | thermal | coolant	heat dissipation | heat sink | heat exchanger

patent cluster	# of connections	cluster label	1-grams	2-grams
36	3542	Data Storage Systems	storage | memory | cache | file | data	data storage | encoded data | dispersed storage
37	102	HVAC Climate Control	temperature | hvac | thermostat | energy | setpoint	energy consumption | setpoint temperature | ambient temperature
38	2298	Nucleic Acid Analysis	nucleic | sample | acid | dna | sequencing	nucleic acid | methods compositions | invention provides
39	189	Medical Neural Stimulation	stimulation | nerve | tissue | electrical | electrode	electrical stimulation | nerve stimulation | peripheral nerve
40	30	Patient Support Equipment	support | patient | deck | frame | foot	patient support | support apparatus | hospital bed
41	2063	Wireless Communication Technology	wireless | ue | station | transmission | channel	base station | wireless communication | user equipment
42	723	Social Networking Platform	social | networking | users | online | content	social networking | user social | online social
43	977	Steel Cutting/Coating	steel | material | sheet | surface | coating	steel sheet | method producing | cutting edge
44	25	Light Imaging Technology	light | image | imaging | lidar | sensor	image sensor | light field | light pulses
45	43	Electronic Connectors	connector | electrical | electronic | plug | housing	electrical connector | electronic device | connector includes
46	280	Magnetic Sensor Devices	sensor | magnetic | field | sensing | current	magnetic field | magnetic sensor | field sensor
47	890	Organic LED/Solar	solar | layer | light | emitting | photovoltaic	light emitting | solar cell | emitting device
48	0	Integrated Circuit Devices	circuit | transistor | voltage | integrated | semiconductor	integrated circuit | semiconductor integrated | circuit includes
49	228	3D Printing Technology	printing | ink | dimensional | build | printer	dimensional printing | imprint lithonetworky | print head
50	62	Food Composition & Protein	protein | food | composition | soy | product	soy protein | protein solution | food product
51	37	Solar Energy Conversion	solar | photovoltaic | dc | power | inverter	photovoltaic power | dc power | solar power
52	32	Content Delivery Network	cdn | content | delivery | server | origin	content delivery | delivery network | network cdn
53	10004	Pharmaceutical compositions/therapy	compositions | invention | acid | cells | present	present invention | invention relates | invention provides
54	95	Web Content Management	content | folder | web | collection | item	collection folder | content management | content item
55	8393	Image Processing & Autonomous Vehicles	image | vehicle | object | display | camera	image data | aerial vehicle | unmanned aerial
56	1464	Cybersecurity & Threat Detection	security | malware | threat | malicious | risk	malware detection | security platform | anomalies threats
57	2954	Medical Devices and Implants	distal | catheter | implant | tissue | end	distal end | proximal end | devices methods
58	67	Medical Stimulation Devices	stimulation | tissue | ultrasound | nerve | transcutaneous	transcutaneous stimulation | adipose tissue | electrical stimulation
59	386	Pharmaceutical Treatment Methods	treating | treatment | administering | disease | compositions	methods treating | compositions methods | pharmaceutically acceptable
60	41	Fluid Management Systems	flow | valve | tubular | pipe | fluid	tubular element | tubular section | flow path
61	234	Electronic Messaging Platform	message | notification | messaging | user | email	electronic message | agent performance | contact information
62	175	Drug Delivery Devices	needle | drug | dose | syringe | delivery	delivery device | drug delivery | piston rod
63	47	Optical Imaging and Analysis	radiation | imaging | ray | detector | optical	absorption data | light source | radiation source
64	576	Semiconductor Device Fabrication	semiconductor | layer | substrate | region | gate	semiconductor device | semiconductor substrate | dielectric layer
65	1104	Healthcare Information Systems	patient | health | medical | healthcare | care	health care | medical information | patient data
66	18	Acoustic-Piezoelectric Devices	piezoelectric | acoustic | resonator | idt | surface	acoustic wave | piezoelectric plate | piezoelectric element
67	94	Messaging & Collaboration	message | messaging | user | chat | conversation	instant messaging | user device | electronic message
68	62	Stem Cell Research	cells | stem | pluripotent | progenitor | differentiation	stem cells | pluripotent stem | progenitor cells
69	145	Gambling Games	game | wager | player | gambling | entertainment	entertainment game | real world | hybrid game

patent cluster	# of connections	cluster label	1-grams	2-grams
70	31	Internal Combustion Engine	piston | valve | damper | crankshaft | engine	connecting rod | control valve | compression ratio
71	108	Media Content Recommendation	content | item | media | user | ratings	content item | media content | content based
72	1275	Fluid & Gas Systems	gas | fluid | liquid | chamber | air	exhaust gas | heat exchanger | compressed air
73	56	Touch Sensing Technology	touch | sensing | capacitive | sensor | capacitance	touch sensor | touch sensitive | touch panel
74	474	Lidar Optical Technology	optical | light | lidar | laser | beam	light source | light beam | optical signal
75	43	Surgical Robotics Cluster	surgical | instrument | robot | tool | effector	surgical instrument | end effector | surgical tool
76	129	Biosensor Analysis Technology	sample | analyte | electrode | test | biosensor	working electrode | liquid sample | flow cell
77	14	Nucleic Acid Biotechnology	nucleic | rna | acid | expression | sequence	nucleic acid | control elements | promoter control
78	878	Optical Networking Technology	optical | light | laser | waveguide | wavelength	optical signal | optical fiber | light source
79	27	Gaming & Accessory	game | player | battle | gaming | server	game content | server device | game program
80	30	Vehicle Tire Monitoring	tire | vehicle | sensor | pressure | door	tire pressure | pressure sensor | door lock
81	28	Television Program Guide	television | program | guide | interactive | schedule	program guide | television program | interactive television
82	164	Semiconductor Memory Devices	memory | semiconductor | layer | cell | bit	memory cell | memory device | semiconductor memory
83	540	Antibody Therapy Research	antibodies | antibody | binding | anti | cd	antigen binding | binding fragments | present invention
84	127	Measurement Sensor Technology	sensor | measuring | acoustic | measurement | gas	gas sensor | data sheet | tank floor
85	61	Fluid Pump Devices	fluid | pump | blood | gas | breast	breast pump | breathing gas | piezo air
86	4816	Search Engine Technology	search | query | document | results | queries	search results | search query | search engine
87	107	Fluid Analysis Technology	sample | fluid | flow | chamber | cartridge	flow cell | fluid sample | analytical instrument
88	23	Image Forming Devices	toner | sheet | member | forming | roller	image forming | forming apparatus | main body
89	172	Microfluidic Devices/Systems	microfluidic | fluid | sample | droplet | channel	microfluidic device | microfluidic channel | cell processing
90	99	Location Tracking Technology	location | mobile | devices | geonetworkic | determination	mobile device | location information | location data
91	55	Serial Bus Protocol	bus | serial | clock | signal | bit	serial bus | clock signal | digital data
92	24	Wireless Power Management	power | wireless | transmit | consumption | communication	power control | transmit power | power consumption
93	49	Fluid Dispensing Devices	dispensing | container | dispenser | outlet | liquid	dispensing apparatus | dispensing device | liquid medicine
94	100	User Interface Design	interface | user | ui | networkical | application	user interface | networkical user | computing device
95	45	RNAi Therapy Cluster	expression | compositions | gene | agents | treating	compositions methods | double stranded | invention provides
96	1380	Networking & Traffic Management	network | packet | traffic | routing | node	network traffic | network device | virtual network
97	850	Telecommunication Services & Devices	telephone | voice | caller | message | calls	telephone number | calling party | called party

Table 4.3: **Names of the patent clusters according to their ngrams.** Cluster labels were inferred from the top 20 1-grams and top 20 2-grams. The number of connections represents the number of connections between the patent cluster and all investor communities in the bipartite network.

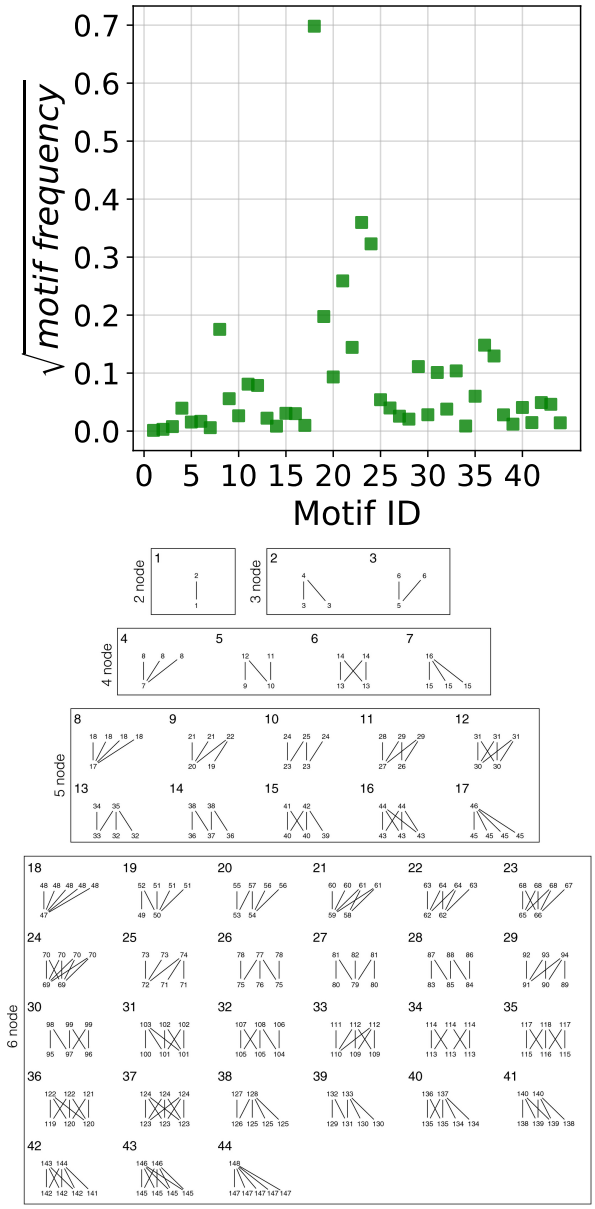

Figure 4.8: **Bipartite motif analysis of the investor community-patent cluster network.** Top : frequencies of bipartite network motifs found on the investor community-patent cluster network. Motif frequencies were computed using the bmotif package [272]. Bottom : shape corresponding to each motif ID (taken from [272]). Comparisons with the motif frequencies shown in [238] do not easily allow for the discrimination between antagonistic and mutualistic networks.

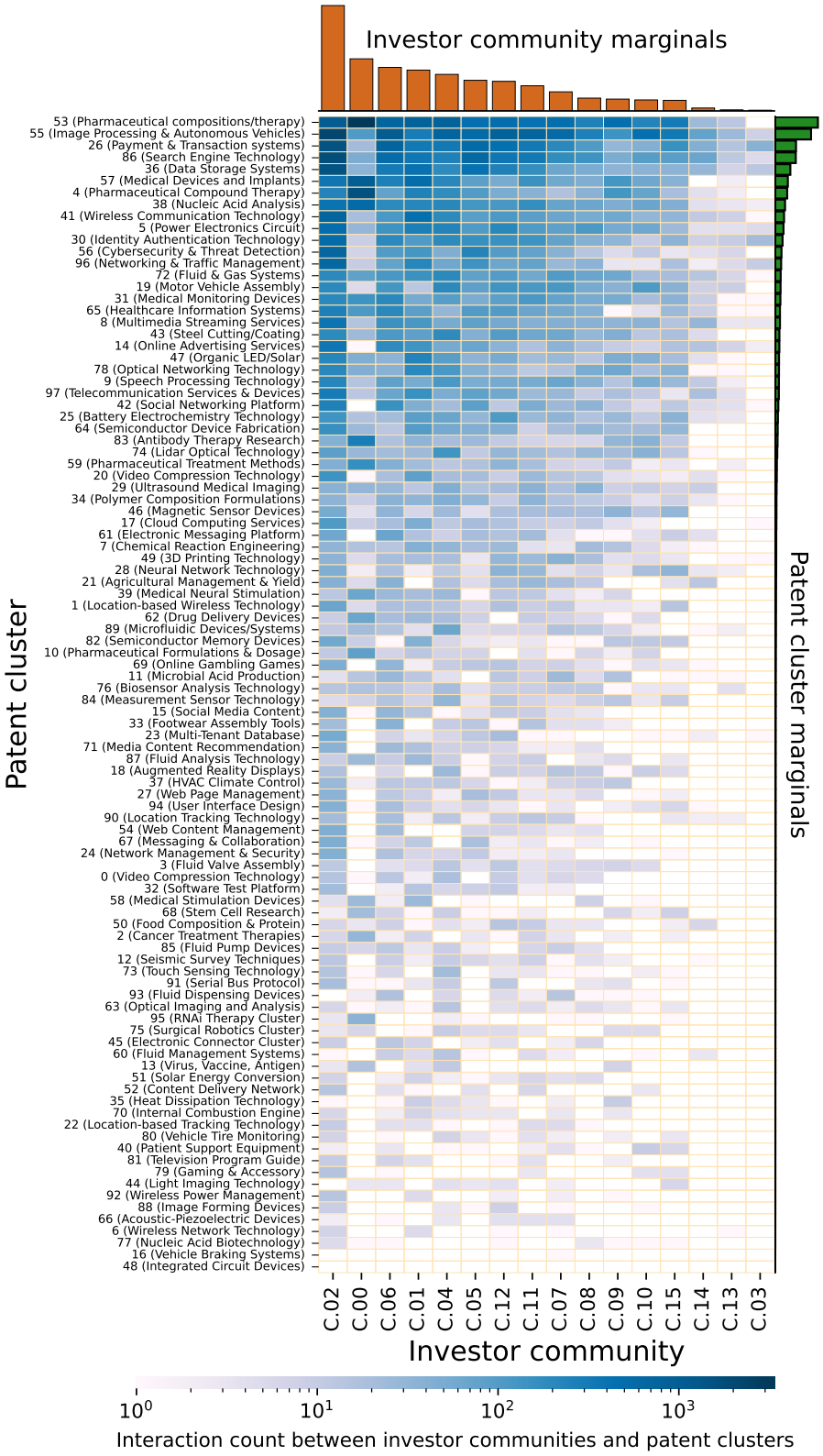

Figure 4.9: **Upper-left packed biadjacency matrix of the bipartite investor community-patent cluster network.** Patent clusters and their associated label correspond to the rows of the matrix, investor communities to the columns. The sum of each row and column (marginals) is computed and shown in the histograms on the top of the matrix for investor communities and on the right of the matrix for patent clusters. The matrix is then reordered (upper-left packed) by rearranging all rows and all columns by descending order of degree. Network-level structural metrics (such as nestedness, connectance and modularity) are computed based on this biadjacency matrix.

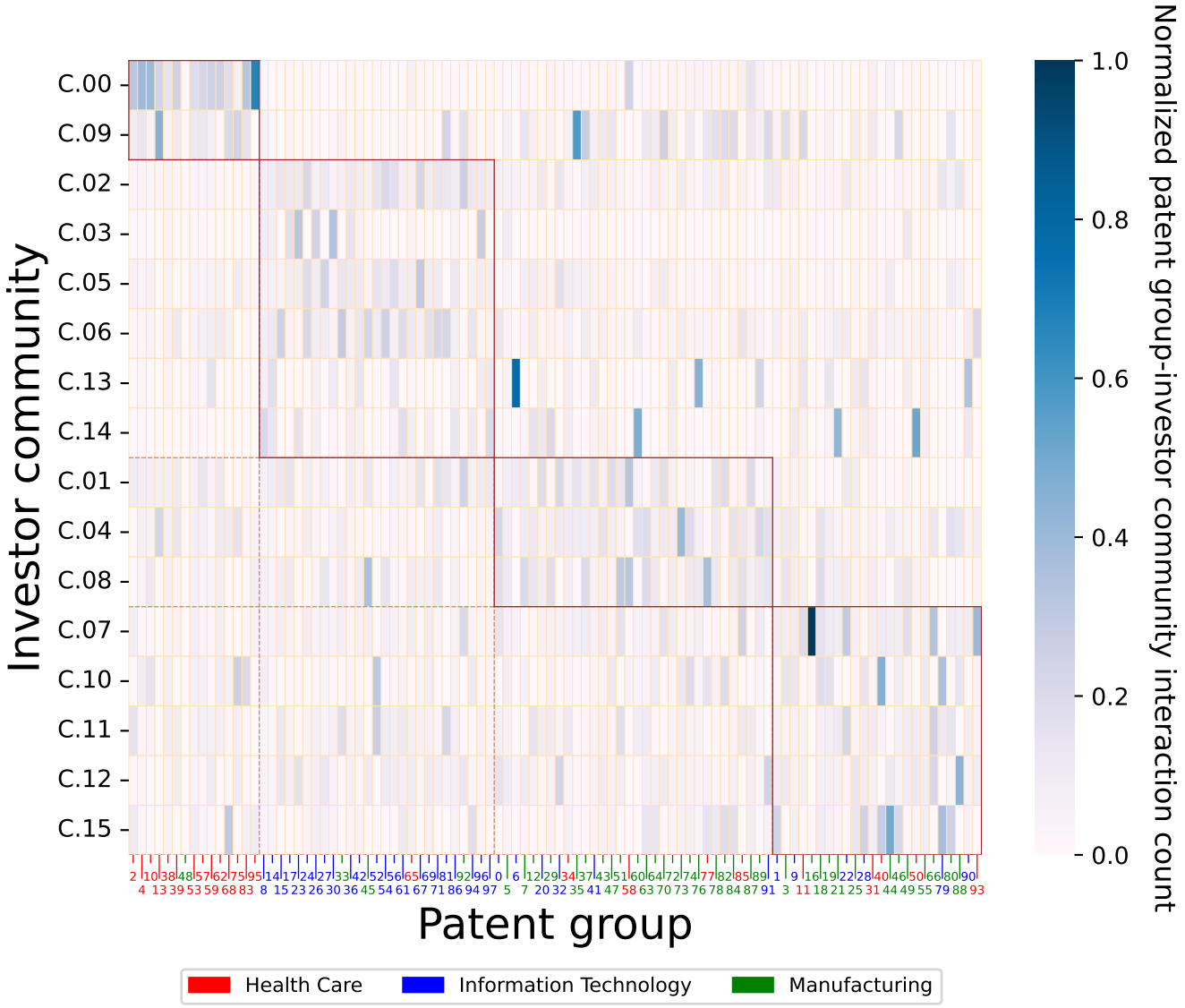

Figure 4.10: **Community-ordered biadjacency matrix using the bipartite modularity maximization algorithm.** The brown rectangle outlines show the modules retrieved by the algorithm. patent cluster tick colors represent the general technological field of the patent cluster (red corresponds to Health Care-related technologies, green to Manufacturing-related technologies and blue to Information Technology-related technologies). Note that patent cluster 48 has degree 0, and thus its allocation to the first module by the algorithm is purely random.

Weighting the network by financial amounts

We run the same analysis on a different version of the network, where interactions between investor communities and patent clusters are not weighted by the number of interactions but rather by the financing amounts, *i.e.* the weight of the interaction between community *C.00* and patent cluster 1 corresponds to the sum of the amounts invested by investors community *C.00* in patent cluster 1. The statistical relevance of the metrics (nestedness, modularity) is shown in Fig. 4.11, and we see that the results obtained for the network where interactions are weighted by interaction count hold for the network where interactions are weighted by amounts invested. This network is significantly nested ($z_\rho = 5.31$) and significantly less modular ($z_Q = -6.82$) than the networks generated by the null model.

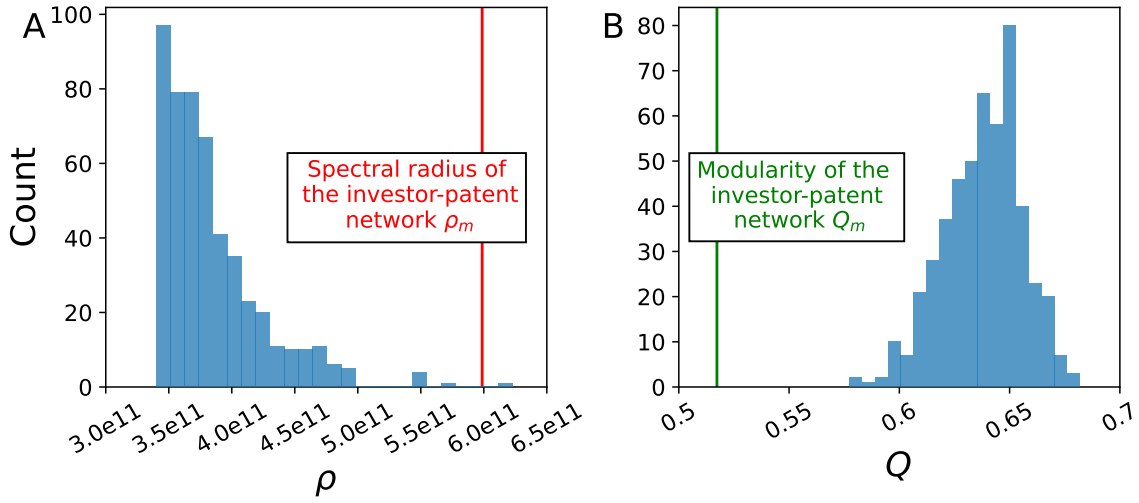

Figure 4.11: **Statistical relevance tests for the nestedness and the modularity of the network weighted by funding amounts. (A)** Statistical relevance test for the nestedness ρ_m (red vertical line) of the investor community-patent cluster network compared with 500 iterations of the null model (blue histogram) described in the Appendix. We see that our network is significantly more nested compared to networks generated by the null model, as was found in the network where only the number of interactions were studied. **(B)** Statistical relevance test for the modularity Q_m (green vertical line) of the investor community-patent cluster network compared with 500 iterations of the null model (blue histogram). We see that our network is significantly less modular compared to networks generated by the null model, as was found in the network where only the number of interactions were studied.

Figure 4.12 shows the upper-left packed biadjacency matrix of the network weighted by funding amounts. The top patent clusters remain in roughly the same order, but investor communities change markedly with community *C.05* (late-stage investors) now being the community with the highest degree.

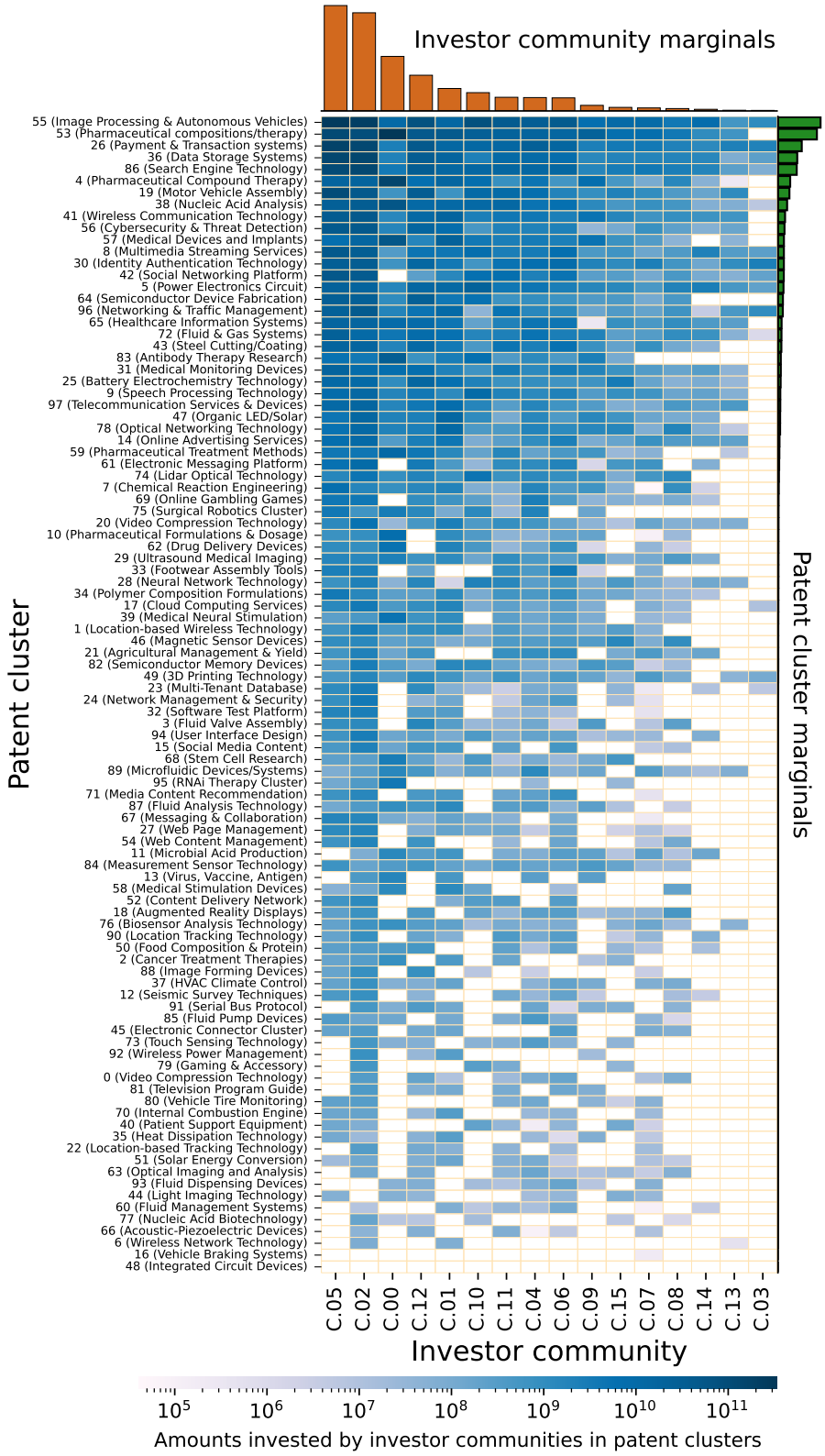

Figure 4.12: **Upper-left packed biadjacency matrix of the bipartite investor community-patent cluster network weighted by funding amounts.** The sum of each row and column (marginals) is computed and shown in the histograms on the top of the matrix for investor communities and on the right of the matrix for patent clusters. The matrix is then reordered (upper-left packed) by rearranging all rows and all columns by descending order of degree.

Conclusion

Summary

Chapter 2 presented a novel clustering method for venture capital investors. We computed characteristic distributions for each individual investors through their investment behavior, and computed the pairwise similarity for all investors. We then detected clusters based on similarity relationships between all investors. We showed that this method uncovered highly interpretable investor clusters, homogeneous in membership and heterogeneous in size. Furthermore, we showed that this approach was robust to feature decimation, as the high-level clusters were similar when computing clustering taking into account all investors characteristics or simply part of them, suggesting underlying complex investment patterns. Analysis of these results provided us with insights into the emergence of new actors of venture capital following events such as the 2008 financial crisis or the 2013 venture frenzy. Furthermore, this clustering approach provides a strong methodological tool to palliate the sparsity of interaction and heterogeneity of nodes in venture capital networks, which represents a significant step in studying their large-scale structure.

Chapter 3 presented a method towards the automatic detection of research topics in a large and complex technical domain. We assessed the validity of our analysis on a corpus of journal articles and conference proceedings (sources) on bioinspiration and biomimetics, a highly interdisciplinary subset of the scientific literature. We applied a natural language processing methodology that automatically extracts research topics directly from the titles and abstracts of the corpus. We characterized and presented each of the research themes automatically discovered in each of the sources, and analyzed their intersections between the different sources. We also examined research trends by studying the evolution of the number of articles in each of the research themes. This provided a snapshot of the current state of the bioinspiration and biomimetics literature, and validated the feasability of automatic detection of specific research themes and trends in scientific production. This both provides a validation of the methodology before applying it to patent corpora and a first step towards the integration of scientific trends in the study of entrepreneurial dynamics.

Chapter 4 presented a study of the startup-led innovation funding ecosystem. We built a bipartite network directly linking investors to patents owned by the startups they fund. We leveraged the approach described in chapter 2 to perform community detection on investor nodes and use topic modeling to perform clustering on patent nodes, creating a coarser-grained view of the network which reduces its size and sparsity and increases the heterogeneity of the nodes. Using structural metrics originally developed to study bipartite ecological networks, we found this network to be topologically mutualistic, with a heterogeneous degree distribution, a high nestedness and a low modularity. This specific structure is due to the prevalence of links between generalist investors and general purpose technologies, *i.e.* technologies with a broad spectrum of applications. This network structure implies non-linear response to crises, with the system weakly affected by negative events affecting specialist nodes and strongly affected by events targeting generalist nodes.

In the course of this thesis, we have provided novel insights using a complex networks approach to entrepreneurial ecosystems. We developed domain-specific methods that helped automatically uncover investor communities based on their behaviors and characteristics and automatically extract clusters from large corpora of text documents. By combining these two methods, we then built and characterized the structure of the startup-led innovation funding network.

In doing so, we have seen that explicitly taking into account the nature of the interactions in the networks allowed us to apply specific methodologies particularly suited for our research questions : investors characterized through the bipartite structure of the investor-startup network built using publicly-available data allowed for the extraction of highly meaningful communities, and the bipartite study of the investor-patent network permitted the use of specific metrics (such as nestedness and bipartite modularity) that were extensively linked with the robustness of the system in the literature, giving insights into its potential strengths and vulnerabilities for identifiable classes of actors. Furthermore, the topologically mutualistic nature of the investor community-patent cluster network places it within the general framework of mutualistic networks, allowing us to draw from this rich scientific literature. Careful consideration, however, must be given when doing so due the difference in the fundamental nature of interactions : conclusions deriving from the topological architecture can potentially be used in the study of investor-patent networks whereas those depending on the mutualistic nature of the interactions are not necessarily applicable.

Further avenues of research

Other types of complex networks

"There is a large element of compromise in mathematical modeling. The majority of interacting systems in the real world are far too complicated to model in their entirety"[6]. The systems studied here are no exception to this rule, and the data and network structures used in this thesis were chosen based on a trade-off between availability, complexity and estimated added value for the questions studied.

One direct improvement, given the data and hardware used in the course of these works, could potentially lie in using more elaborate modeling tools in order to reduce the compromises made, notably richer network structures, which could provide additional insights and alternatives to some of the works presented here. For instance, investor-startup (as modeled in chapter 2) or investor-patent (as modeled in chapter 4) relationships, rather than being modeled as bipartite networks, could be represented as multilayer networks or hypergraphs [77].

Multilayer networks are composed of several representations of the interacting nodes that exist in parallel, with each representation corresponding to one mode of interaction between the nodes. This allows for the integration of more information on the specific

[6]Taken from https://people.maths.bris.ac.uk/~madjl/course_text.pdf

interactions between two nodes, rather than simply linking two nodes in interaction. Furthermore, it has been found that, when a system is inherently multilayer, conclusions drawn from analysis of the aggregated graph can be misleading [79]. In the case of entrepreneurial ecosystems, multilayer representations can offer additional insights : take, for instance, the investment portfolio of a venture capital firm. Investors can fund companies at different stages of their lifecycle, an interaction network which would be more accurately modeled as a multilayer graph where each layer corresponds to a different investment stage and intralayer interactions to links between investors and startups if a funding round happened at the corresponding stage. For certain phenomena such as the estimation of information flow between investors, metrics computed on this graph, rather than on the aggregated graph where all stages are considered equal, could thus present significant differences.

Hypergraphs are graphs where a single edge (called a hyperedge) can connect any number of vertices, rather than simply connecting dyads. In the context of venture capital networks, as funding rounds often take the form of syndication events, these interactions are more accurately modeled as hyperedges connecting all investors involved in a funding round with the target startup rather than creating an edge linking each investor to the startup for the funding round. This also extends to syndication networks where the $\frac{n(n-1)}{2}$ dyads created between the n investors for each funding round (as is common in VC syndication networks) are instead more accurately represented as a single hyperedge directly connecting the n investors. The hypergraph representation is thus able to retain a larger amount of context, yielding information about all parties involved in an interaction rather than treating each dyadic interaction independently.

Research on these specific networks, however, is recent and there are still relatively few algorithms specifically designed for these specific structures. being increasingly developed [196, 75, 152, 19]. As these classes of graph make less compromises when building the representation of the underlying nature of the interactions in the venture capital network, we can hope to gain considerable insight from their use. Their analysis is, as the moment, fairly challenging due to the relative scarcity of available tools, but the methodological advances performed by the network science community are rapidly offering relevant tools and can certainly provide interesting insights.

Temporal dynamics

Compared to, for instance, ecological data, socio-economic datasets such as those used here often allow for longitudinal investigations of the phenomena studied, due to the usually more detailed and comprehensive nature of their records. Since most interactions modeled in this thesis are timestamped, we hoped to be able to study their temporal dynamics in a number of different contexts and representations. Even though some temporal evolutions were detected in our works (e.g. chapter 2), significant structural evolutions in terms of nestedness or modularity of the investor-patent network presented in chapter 4, for instance, could not be observed. These network dynamics have been observed in other socio-economic networks such as micro-blogging data [38] in the form of a self-adapting user-meme network that underwent a modular-to-nested structural transition. The reason for the absence of structural evolutions, with the measured nestedness and modularity

remaining relatively stable for the last decade, remains unclear. As the dynamics of the investor-patent network are rather slow compared to that of micro-blogging, the period studied might simply not be long enough to observe this transition, which could potentially happen in the future or have already happened in the past. Even though the data available goes back a number of decades, the exhaustiveness of the Crunchbase dataset becomes more and more questionable the further back we go in the past, as Crunchbase itself (the data provider) was founded in July 2007. Data prior to this date was collated from a number of sources, but is not as reliable as data following the creation of the company. Another candidate explanation is simpler : there is no guarantee that such a transition takes place in the investor-patent network, which could simply turn out to be structurally stable and remain in its measured state in the absence of significant perturbations. Studying the temporal dynamics of the networks linking the various agents in entrepreneurial ecosystems would benefit from further research leveraging data covering a longer timespan.

Collective motion

Popular assessments about *herding behaviours* or about investment fads and fashions are widespread, sometimes supported by anecdotal evidence, but observing, measuring or evaluating how, and in what respect, the investment strategies of investors coordinate and evolve through time is still challenging. Public financial markets have, on their part, been an active topic of study for physicists [74, 276], but the venture ecosystem has up to now only limitedly been subjected to scientific investigations on these matters. To put it differently, even though new ventures have been a corner topic of the literature on entrepreneurship for the past 20 years [65] and although investments strategies and related social processes play a major role in structuring the dynamics of startup ecosystems, we are still mostly missing methods and tools that would help us understand the processes affecting or governing herd behaviour in venture capital networks. This question will benefit from drawing on knowledge from different disciplines that have already studied similar concepts applied to different contexts [161], as the study of financial markets has benefited from the insights of physicists on certain matters [39, 197]. Indeed, this can be thought of as an application of collective motion [307], one of the higher-profile topics of interdisciplinary research [307]. In the course of this thesis, advances have been made to study herding behaviour in venture markets, but much remains to be done : our novel clustering method presented in chapter 2 helps deal with the heterogeneity of individual investors and their relatively temporally scarce activity patterns, but the *space* (in the mathematical sense of the word) in which to measure their collective behaviours has – so far – eluded us.

Integrating more data

The choice during this thesis was to perform analyses on data relatively easy to obtain for all companies and investors, using mainly public fundraising events and basic company data that is accessible through commercial APIs without rate limits. This allows for the study of large, representative datasets as the methods can be generalized to any number of companies and investors as long as they are represented in the dataset. There exist data

relevant to entrepeneurial dynamics that is much more detailed in nature such as company valuation during funding rounds or fund performance metrics, but it is difficult to access, being either much more expensive or private and thus accessible only for a small subset of companies. There is no doubt, however, that unlocking this data would open the door to a large number of studies highly valuable to entrepreneurship research, such as for instance relating node characteristics in the investor-startup or investor-patent network to fund performance at a large scale, or studying the impact of innovation metrics of a company (computed through its patent portfolio) on its valuation during funding rounds.

A longitudinal study of the geographically-embedded investor-startup temporal bipartite network appears as a potential candidate avenue of research. Indeed, preliminary observations have shown that even though the majority of investments take place locally, long-distance investing remains a sizeable portion of the total number of investments. This finding suggests a potentially richer mechanism than the common short-distance investment pattern and pleads for direct investigations of the circumstances under which investors decide to venture further away, in keeping with [275].

The examples given above remain strongly entrepreneurship-related, leveraging data that directly concerns actors directly interacting together in entrepreneurial ecosystems such startups and investors. One major avenue of research, however, lies in the integration of environmental characteristics in future studies of entrepreneurship. Indeed, as discussed in this thesis and in a number of research works [78], innovation does not happen in a vacuum, and its actors are strongly influenced by the environment they are involved with. Events such as the COVID-19 pandemic, public policies, or trends and dynamics in academia all have significant impacts on entrepreneurial dynamics. Here, we have taken a step towards expanding the system-environment boundary of entrepreneurial ecosystems through the automatic characterization of the evolution of trends in academia, with the distant goal being the study of their impact on startup-investor networks. Much, however, remains to be done to integrate the impact of these trends – amongst many others – on entrepreneurial ecosystems.

Appendix

Appendix A

The Crunchbase dataset

Chapters 2 and 4 are based on analyses using the Crunchbase dataset, which has recently become a standard in data-driven studies of entrepreneurship [85]. In the context of their API program, Crunchbase offers the possibility of directly downloading the entirety of their data. The data is continually updated, both for new events pertaining to entrepreneurship (e.g. new organizations being founded, organizations raising funds, individuals moving from one organization to another, new funds being created). Furthermore, as Crunchbase is a US-based organization, its data can potentially present biases in terms of exhaustivity depending on the region covered. North American data, for instance, might be more readily available due to the geographical proximity and shared English language between the data supplier and the public communications and databases they use to build the dataset. As the works presented here span several years, the underlying Crunchbase data can vary from one chapter to another, due to the different dates of extraction.

The strength and weaknesses of existing commercial entrepreneurial databases have been studied [249] with VentureSource, Pitchbook and Crunchbase emerging as the best data providers in terms of coverage and accuracy in key dimensions related to company, financing and founders data. How the date of extraction impacts the contents of the extracted dataset, however, has –to our knowledge– not been studied. Here, after briefly presenting the contents of the Crunchbase database in section A.1, section A.2 shows comparisons of extracts of the database performed at different points in time in order to better understand how the database itself evolved through the course of this thesis.

A.1 Contents of the dataset

Figure A.1 shows the relationships between the different datasets supplied in the Crunchbase database used in this thesis. *People* work in *Startups* that raise funds from *Investors* through *Funding Rounds*. Detailed information is provided for each of these datasets describing the structure or event, allowing us to link the actors in the entrepreneurial ecosystems (for instance, investors can be linked with a startup through the funding round targeting the startup in which the investors have participated). To give an order of magnitude, the

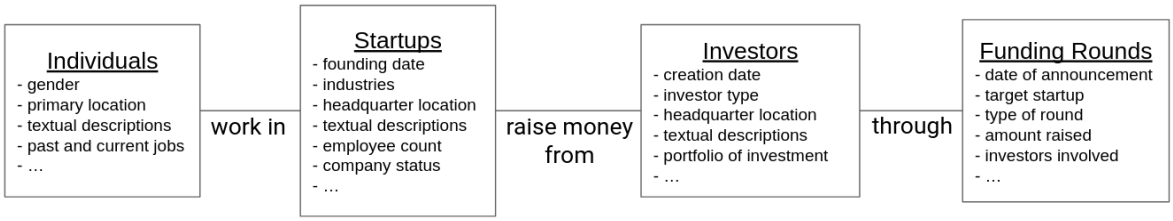

Figure A.1: **Structure and excerpts of the fields of the Crunchbase datasets.**

Crunchbase database extracted on January the 8th, 2024 contains 3 274 470 entries in the organizations table, 1 830 126 in the people table, 464 846 in the funding rounds table and 270 712 in the investors table. Out of the 3 274 470 organizations, roughly 10% (315 091) have raised funds.

Using this database presents a significant difference compared to, for instance, ecological data that is specifically gathered to answer a specific research question and where all data is considered relevant. Here, our database covers a wide array of entrepreneurship-related fields that do not necessarily directly relate to our studies. Relevant subsets of the data present in the database must thus be selected to answer specific research questions.

Table A.1 shows the list of fields in each dataset and their associated data type.

A.2 Temporal evolution of the database

To better understand these differences, we will present here several aspects of the Crunchbase dataset at the continent level extracted at 5 different points in time. The matching between countries and continents was performed automatically using the *pycountry* Python library, resulting in 6 continents : Oceania (OC), South America (SA), North America (NA), Europe (EU), Asia (AS) and Africa (AF). The datasets were extracted on the following dates : dataset *2019* extracted on October 28th, 2019 (28-10-2019), dataset *2020* extracted on October 7th, 2020 (07-10-2020), dataset *2021* extracted on October 28th, 2021 (28-10-2021), dataset *2023* extracted on February 14th, 2023 (14-02-2023) and dataset *2024* extracted on January the 8th, 2024 (08-01-2024).

A.2.1 New companies

Figure A.2 shows the temporal evolution of the number of founded companies for each of the continents for the various extracted datasets. Several different patterns can be seen in this figure.

First, we see that there is a first peak in the number of founded companies around 2000 observed for all continents except Africa, followed by a small decrease until 2002 and then an increase until the peak of new companies is reached in 2015. This property is shared between all datasets.

Second, we see that the number of new companies varies strongly between the different

Organizations	dtype
uuid	string
name	string
type	string
permalink	string
cb_url	string
rank	int64
created_at	string
updated_at	string
legal_name	string
roles	string
domain	string
homepage_url	string
country_code	string
state_code	string
region	string
city	string
address	string
postal_code	float64
status	string
short_description	string
category_list	string
category_groups_list	string
num_funding_rounds	float64
total_funding_usd	float64
total_funding	float64
total_funding_currency_code	string
founded_on	string
last_funding_on	string
closed_on	float64
employee_count	string
email	string
phone	string
facebook_url	string
linkedin_url	string
twitter_url	string
logo_url	string
alias1	float64
alias2	float64
alias3	float64
primary_role	string
num_exits	float64

People	dtype
uuid	string
name	string
type	string
permalink	string
cb_url	string
rank	int64
created_at	string
updated_at	string
first_name	string
last_name	string
gender	string
country_code	string
state_code	string
region	string
city	string
featured_job_organization_uuid	string
featured_job_organization_name	string
featured_job_title	string
facebook_url	string
linkedin_url	string
twitter_url	string
logo_url	string

Funding Rounds	dtype
uuid	string
name	string
type	string
permalink	string
cb_url	string
rank	float64
created_at	string
updated_at	string
country_code	string
state_code	string
region	string
city	string
investment_type	string
announced_on	string
raised_amount_usd	float64
raised_amount	float64
raised_amount_currency_code	string
post_money_valuation_usd	float64
post_money_valuation	float64
post_money_valuation_currency_code	string
investor_count	float64
org_uuid	string
org_name	string
lead_investor_uuids	string
investors	string
org_permalink	string

Investors	dtype
uuid	string
name	string
type	string
permalink	string
cb_url	string
rank	int64
created_at	string
updated_at	string
roles	string
domain	string
country_code	string
state_code	string
region	string
city	string
investor_types	string
investment_count	int64
total_funding_usd	float64
total_funding	float64
total_funding_currency_code	float64
founded_on	string
closed_on	float64
facebook_url	string
linkedin_url	string
twitter_url	string
logo_url	string

Table A.1: **Descriptive table of the data supplied in the Crunchbase dataset.** The list of fields and their associated type (string, float or integer) is presented for each dataset (Organizations, People, Funding Rounds, Investors).

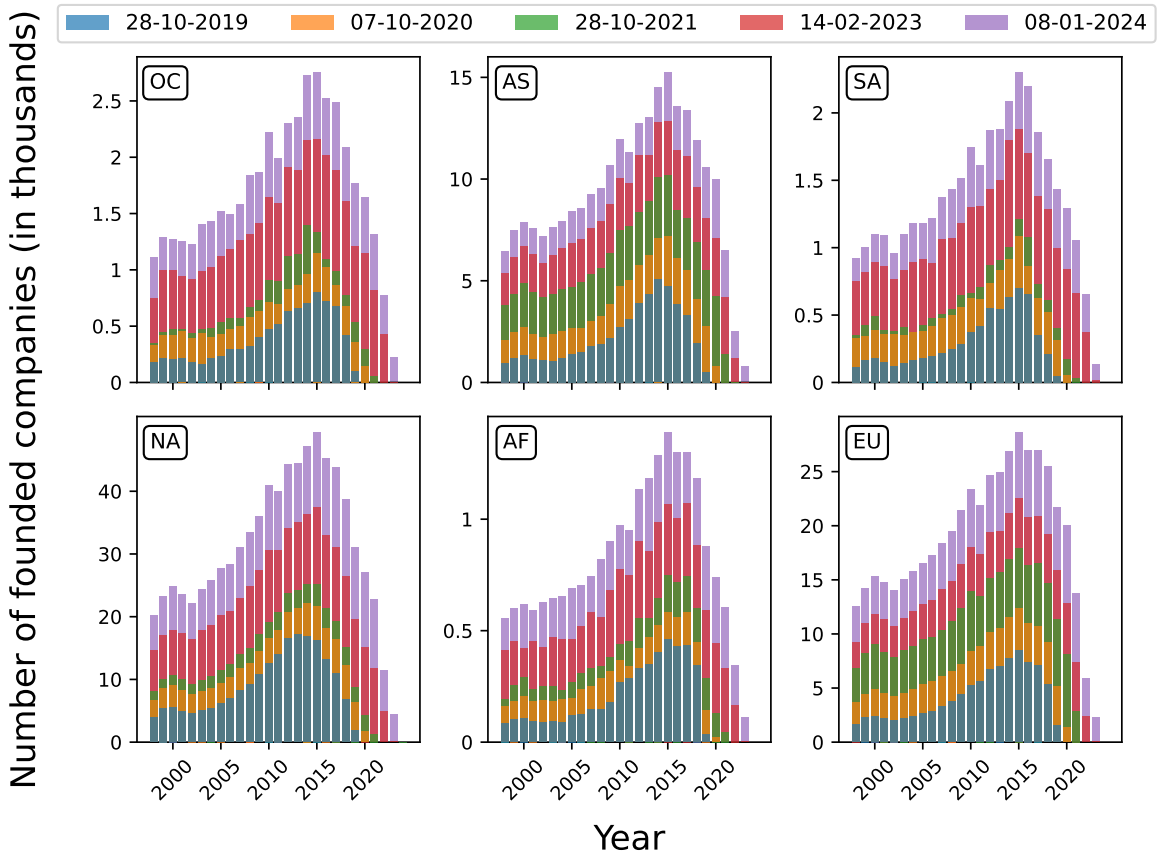

Figure A.2: **Temporal evolution of the number of founded companies in each of the continents.** We count, for each year and each continent, the total number of companies founded in the year.

ecosystems, with North America clearly housing the highest number of company creation, followed by Europe and Asia. Significantly less companies are being founded each year in Oceania, South America and Africa.

Third, we see a strong *a priori* consolidation of the data on newly-founded companies : if we compare the 2019 and 2024 datasets, we see that the number of new companies is significantly higher for the 2024 dataset no matter the year and region. This phenomenon also strongly depends on the geographical location : the relative difference in the temporal series between dataset 2020 and dataset 2021, for instance, is much bigger for the European and Asian continents than for the others, suggesting a strong effort by Crunchbase in improving their coverage of these specific regions between the two extractions.

Last, we see that there is a strong reporting lag in terms of new company creation. If we focus, for instance, on the number of company creations in North America in 2020, we see a stark contrast between the 2021 dataset and the 2023 dataset even though both were extracted long after the end of the year in question.

A.2.2 New investors

Figure A.3 shows the temporal evolution of the number of new investors for each of the continents for the various extracted datasets. An investor is considered *new* for a given year if their first-ever investment in the dataset took place during that year, regardless of their geographical location (if investor A performed their first investment in North America in 2005 and their first investment in Europe in 2007, it will be considered a new investor for the year 2005 in North America and will not count for Europe).

Once again, we see that North America is leading in terms of number of new investors per year, followed by Asia and Europe and then distantly by the other 3 regions. We also see a stark increase in the number of new investors across all regions in 2021 and 2022, much higher than all previous years in all regions except for Asia.

We see a massive difference between Asia and other regions : the number of new investors for all years has drastically increased between the 2019 and 2020 datasets, and between the 2020 and 2021 datasets. For other regions, there is relatively little difference between the various datasets except for the year of extraction. This differs significantly from the comparison across datasets for new organizations. This difference could potentially be a consequence of the fact that investments (and thus investors) are public information that are typically widely communicated, and are thus well-referenced on databases such as Crunchbase. New companies, on the other hand, often do not exist on the databases until they reach certain milestones (such as raising funds) that can help them be referenced. The drastic improvement in coverage of the Asian continent between the 2020 and 2021 dataset is in line with the one seen on new companies.

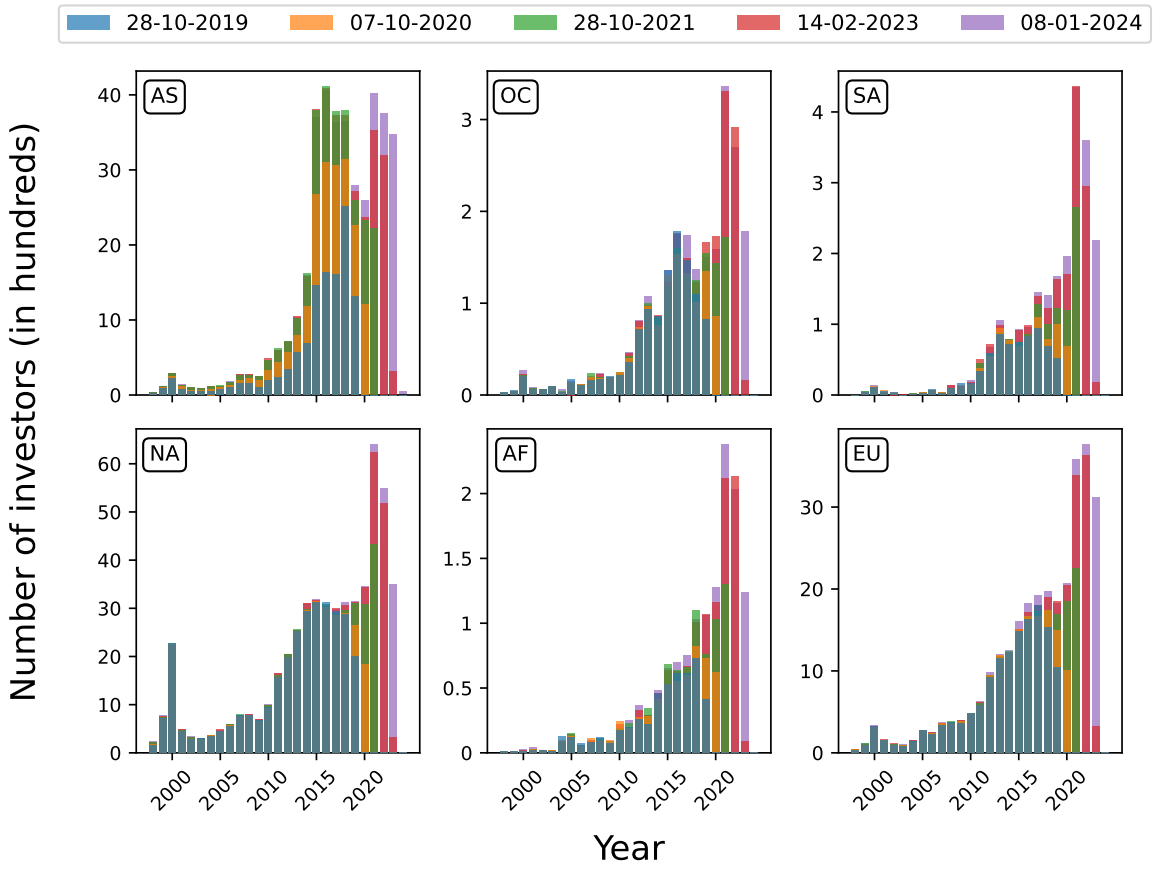

Figure A.3: **Temporal evolution of the total number of new investors in each of the continents.** We count, for each year and each continent, the number of investors with headquarters in the continent that perform their first-ever investment.

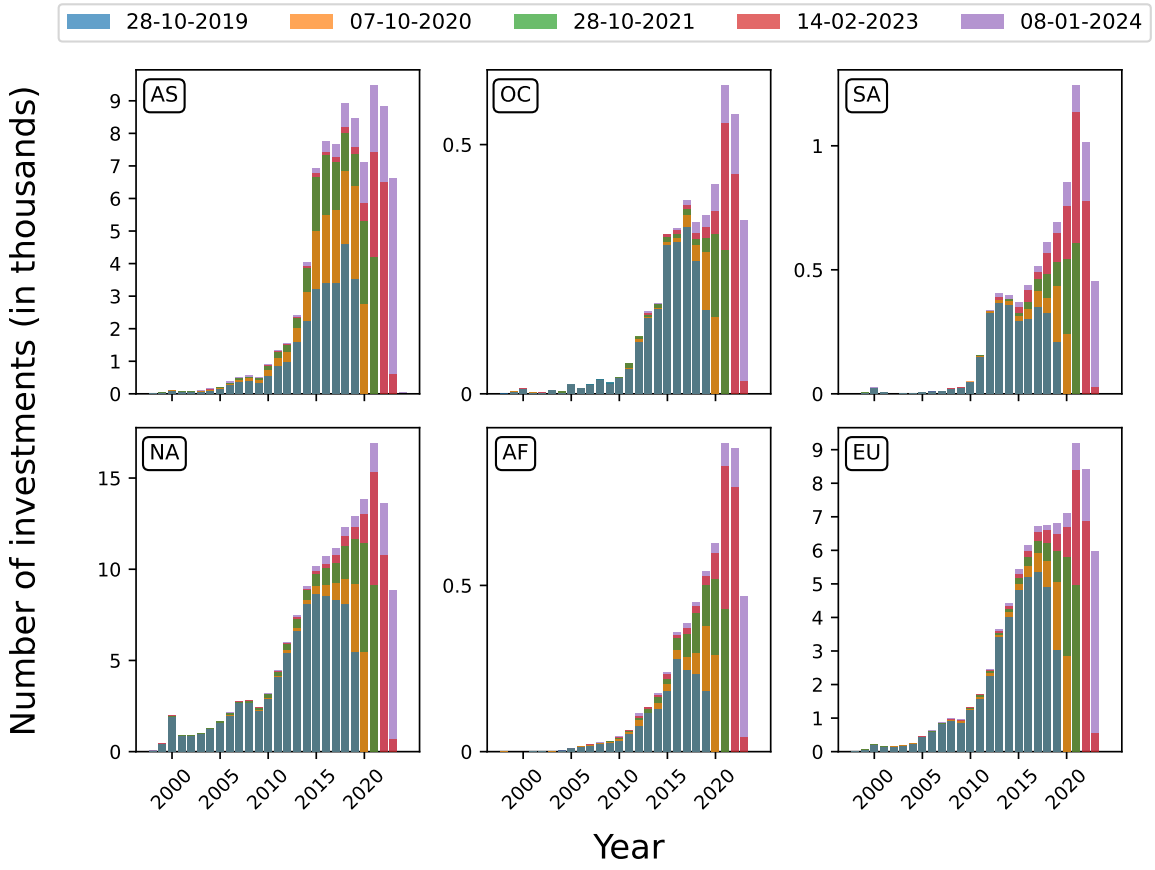

Figure A.4: **Temporal evolution of the total number of funding rounds in each of the continents.** We count, for each year and each continent, the number of venture capital funding rounds targeting companies with headquarters in the continent, regardless of the stage of investment.

A.2.3 Number of investments

Figure A.4 shows the temporal evolution of the total number of funding rounds in each continent, where a funding round is considered as taking place in a continent if the company raising funds has headquarters in the continent.

As previously seen, North America leads in terms of number of funding rounds, followed by Asia and Europe and once again distantly by the 3 other regions. The difference between the various datasets is larger than for new investors, but still relatively small for most regions except Asia. Asia, once again, sees a significant increase in investments over time between the 2019, 2020 and 2021 datasets.

The maximum number of funding rounds is reached in 2021 for all continents. We also see a lag in reporting of the funding rounds in the various regions, as evidenced for instance with the year 2019 for all datasets.

A.2.4 Stages of investment

Figure A.5 shows the temporal evolution of the number of Pre-seed rounds in each continent for the various datasets, computed by counting for each year the number of rounds labeled Pre-seed in the Crunchbase dataset that target a company with headquarters in the continents.

North America leads in number of Pre-seed rounds, followed by Europe, Asia, South America, Africa and Oceania. The maximum is reached after 2020 for all continents, with almost all Pre-seed rounds raised after 2010. We also see that there is a significant increase in the number of Pre-seed rounds between the various datasets for all years, potentially suggesting an *a posteriori* re-classification of rounds as Pre-seed.

Figure A.6 shows the temporal evolution of the number of Seed funding rounds in each continent.

North America leads in terms of number of Seed funding rounds over the years, followed by Europe and Asia, and then distantly by South America, Africa and Oceania.

We observe relatively little difference between the various datasets except for Asia, which once again shows a high level of *a posteriori* consolidation of the funding data. The reporting lag is present but globally low, with for instance an increase in coverage of roughly 20% for year 2018 in North America when comparing the 2019 and 2024 datasets.

Figure A.7 shows the temporal evolution of the number of Series A funding rounds in each continent.

North America leads in terms of number of Series A funding rounds over the years, followed by Asia, Europe, South America, Oceania and Africa.

There is even less difference between the various datasets compared to Seed funding rounds, except once again for the Asian continent which shows a very strong increase in number of funding rounds between the 2019, 2020 and 2021 datasets for all years starting

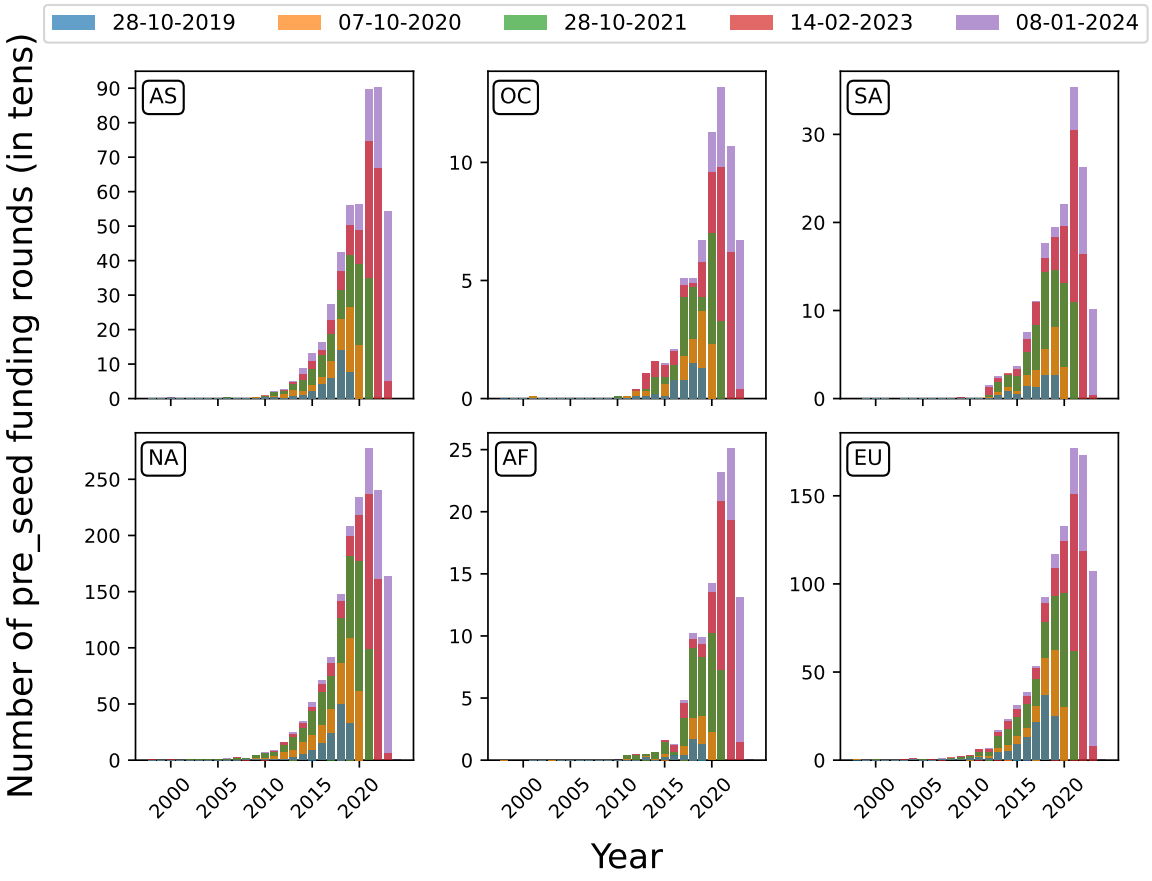

Figure A.5: **Temporal evolution of the number of Pre-seed funding rounds in each of the continents.** We count, for each year and each continent, the number of Pre-seed funding rounds targeting companies with headquarters in the continent.

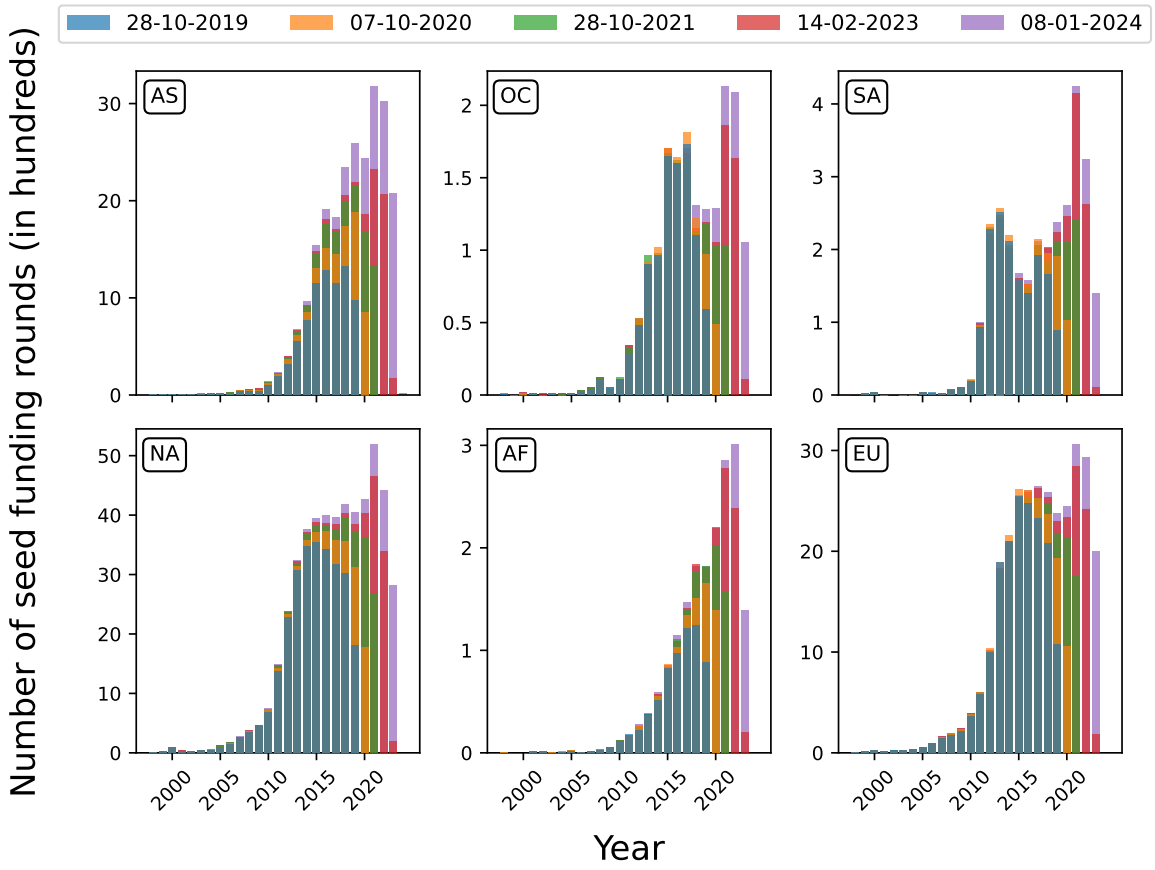

Figure A.6: **Temporal evolution of the number of Seed funding rounds in each of the continents.** We count, for each year and each continent, the number of Seed funding rounds targeting companies with headquarters in the continent.

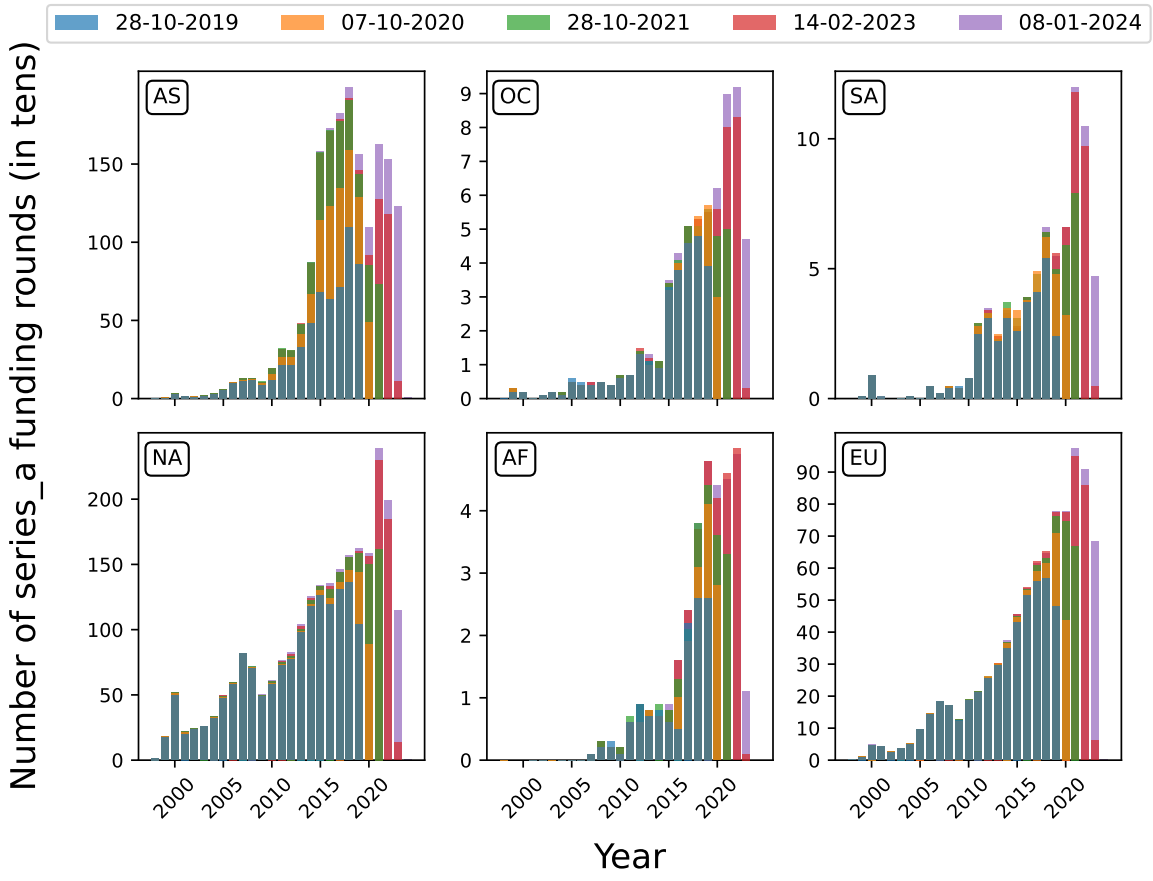

Figure A.7: **Temporal evolution of the number of Series A funding rounds in each of the continents.** We count, for each year and each continent, the number of Series A funding rounds targeting companies with headquarters in the continent.

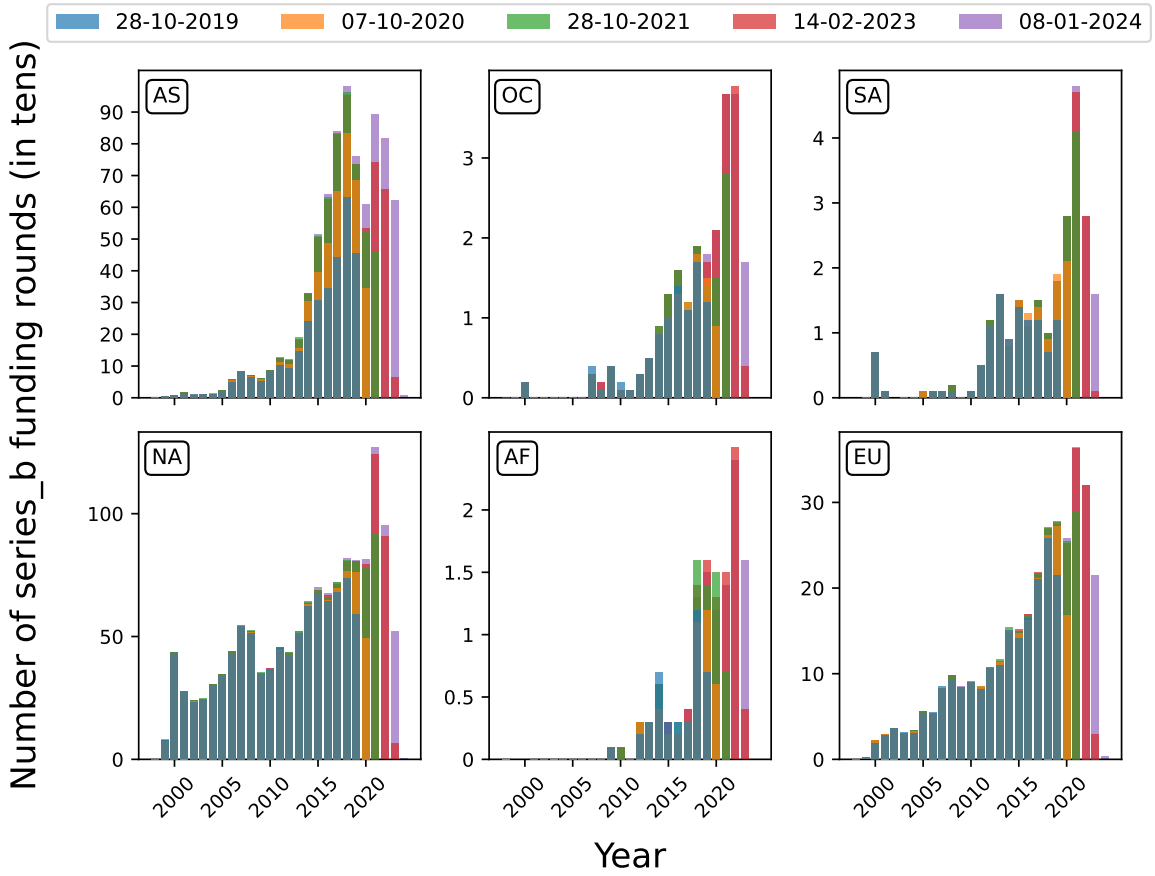

Figure A.8: **Temporal evolution of the number of Series B funding rounds in each of the continents.** We count, for each year and each continent, the number of Series B funding rounds targeting companies with headquarters in the continent.

from 2010 onwards. The high coherence between datasets suggests that information on Series A funding is generally reliable, with most data being correctly referenced as it is communicated. This can be due to several reasons : companies that raise Series A rounds have usually previously raised funds and thus are more likely to already be referenced in Crunchbase, making tracking their company news easier. Furthermore, news concerning Series A rounds tend to be published in more outlets than Seed rounds, reaching wider audiences and thus increasing the likelihood of the information being captured in Crunchbase, regardless of the previous status of the company in the database.

Another observation of note is that, at least in North America, Asia and Europe, Series A funding rounds slowed down in 2008 and 2009 compared to the previous years, a pattern that not observed for Seed funding rounds that grew monotonously until the mid-2010s. This decrease in activity could be a consequence of the 2008 financial crisis that had a specific negative impact on this funding stage.

Figure A.8 shows the temporal evolution of the number of Series B funding rounds in each continent.

North America leads in number of Series B funding rounds, followed closely by Asia and

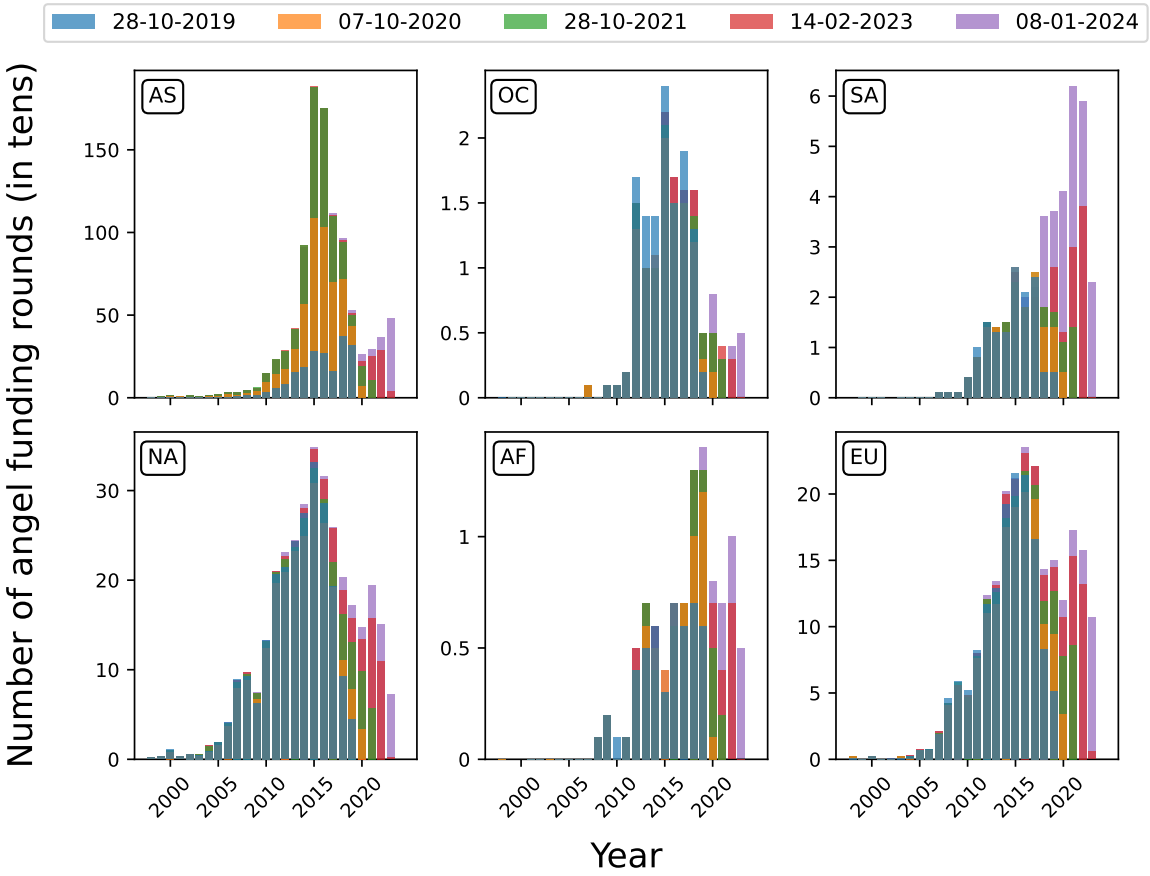

Figure A.9: **Temporal evolution of the number of Angel funding rounds in each of the continents.** We count, for each year and each continent, the number of Angel funding rounds targeting companies with headquarters in the continent.

then more distantly by Europe, with South America, Oceania and Africa-based companies raising few Series B funding. Apart from Asia where a strong improvement in data coverage is yet again observed, this effect is very weak in the other geographical regions.

Similarly to our observations for Series A rounds, we see a decrease in Series B activity after 2007 for Asia and North America. Due to the relative sparsity of Series B rounds around this period in all continents except North America, however, this observation could simply result from statistical fluctuations.

Figure A.9 shows the temporal evolution of the number of angel funding rounds in each continent.

Asia leads in terms of number of angel rounds, followed distantly by North America and Europe, and then by South America, Oceania and Africa. The increase between various datasets is particularly important in the Asian continent for angel rounds, with a roughly 6-fold increase in 2015, for instance, between the 2019 and 2021 datasets. For the 2024 dataset, the maximum in terms of number of angel rounds is observed in 2015 for Asia, Oceania, North America and Europe, in 2021 for South America and in 2019 for Africa.

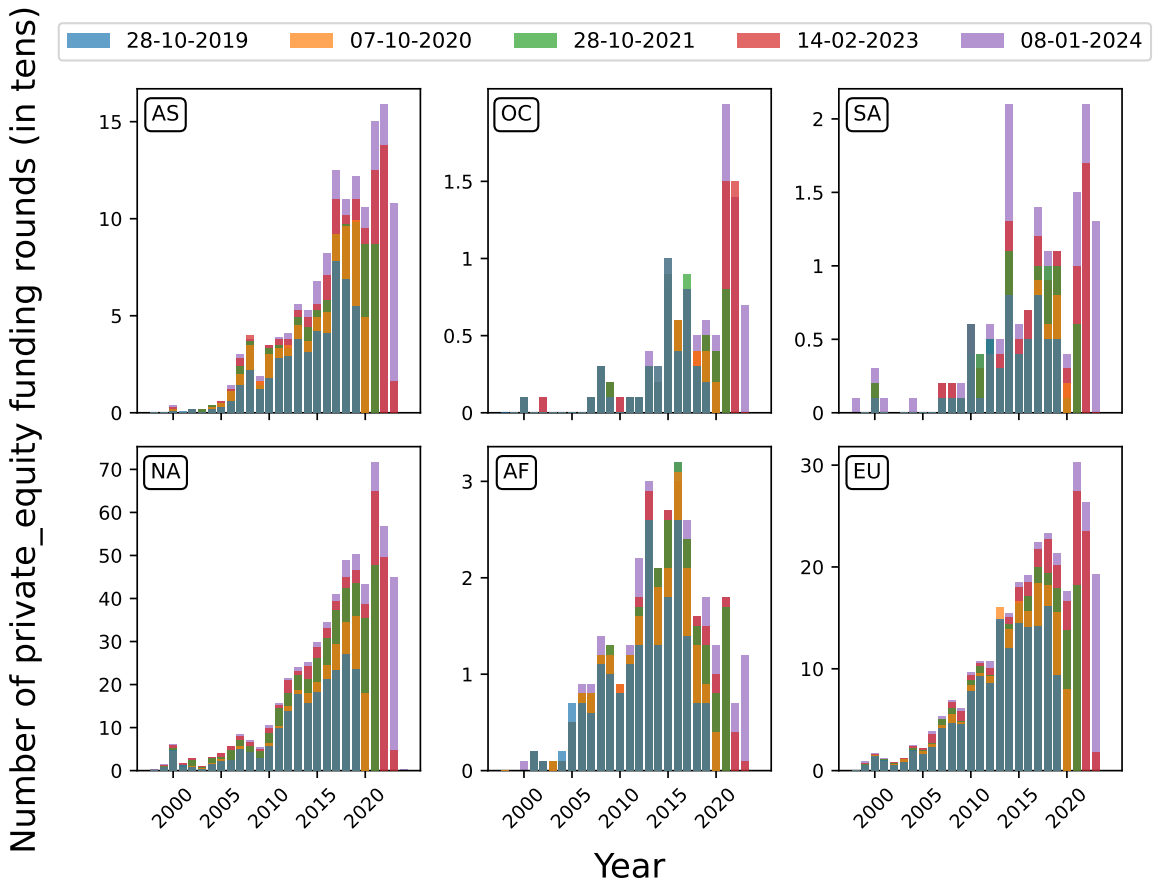

Figure A.10: **Temporal evolution of the number of private equity funding rounds in each of the continents.** We count, for each year and each continent, the number of private equity funding rounds targeting companies with headquarters in the continent.

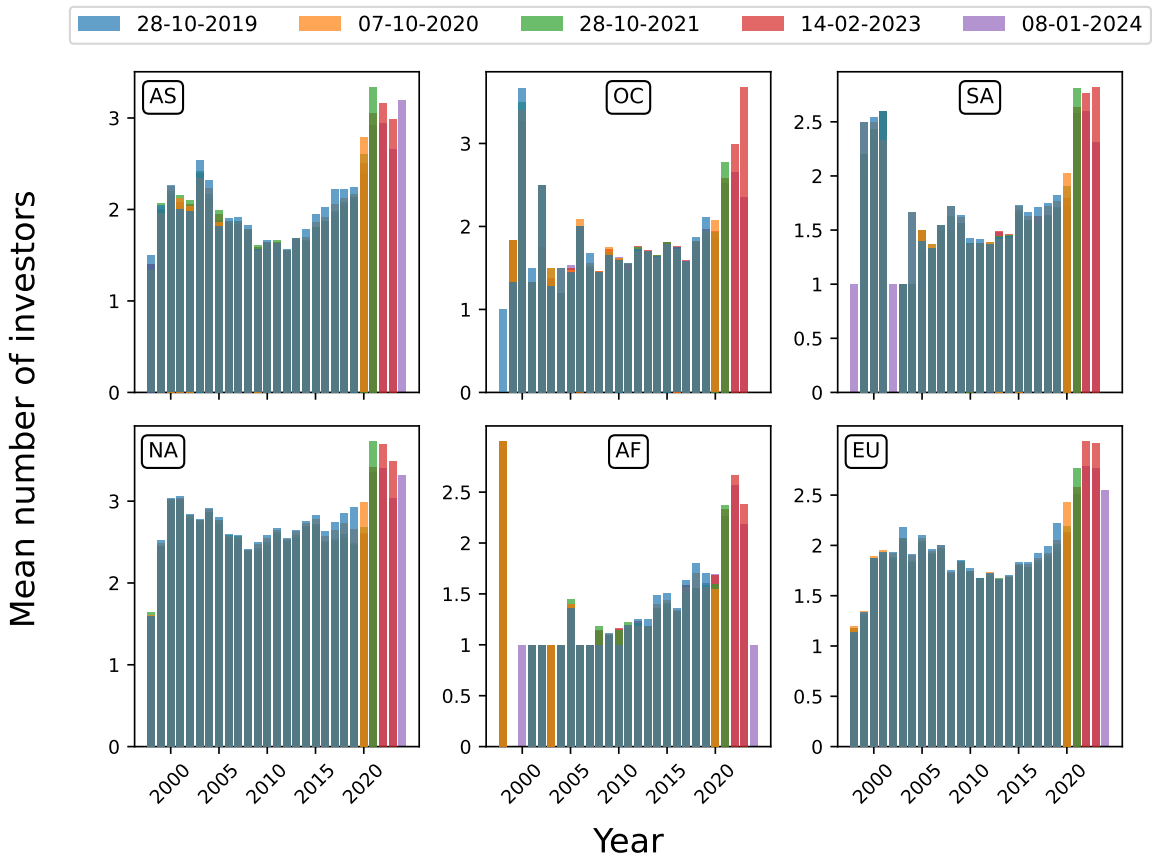

Figure A.11: **Temporal evolution of the mean number of investors involved in each funding round in each of the continents.** We count, for each year and each continent, the number of investors involved in each funding round targeting companies with headquarters in the continent and compute the mean of the number of investors.

Figure A.10 shows the temporal evolution of the number of private equity funding rounds in each continent.

North America leads in terms of number of private equity rounds, followed by Europe, Asia, Africa, South America and Oceania. The increase in coverage between the various datasets is significant and seems roughly similar in proportion for all regions, contrary to other funding rounds where Asia experienced significant differences.

A.2.5 Mean syndication size

Figure A.11 shows the temporal evolution of the mean numbers of investors involved in each funding round for each continent. It is computed by counting, for each dataset, the number of investors involved in each funding round targeting companies with headquarters in the continent and averaging over the number of investors.

The results seem difficult to analyze in the smaller ecosystems (Oceania, South America and Africa) for early years due to the strong fluctuations displayed. The mean number of

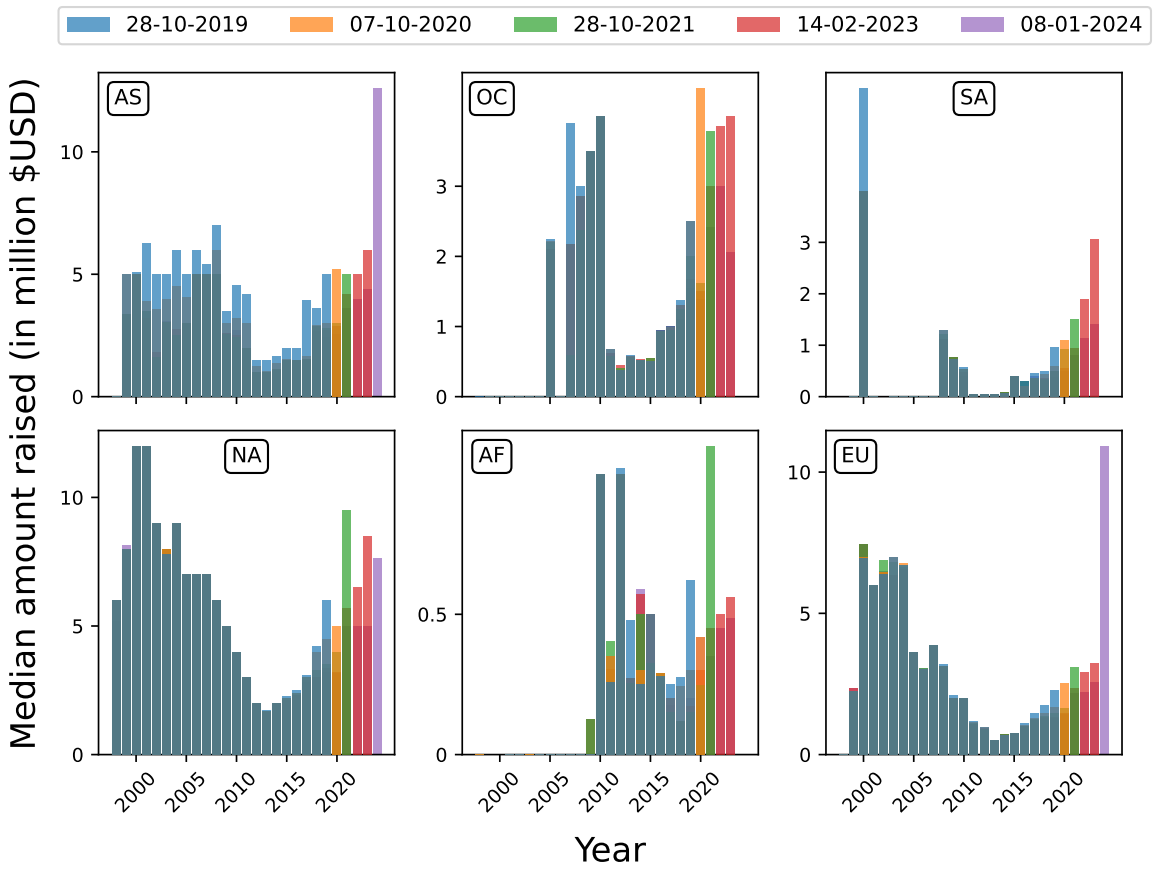

Figure A.12: **Temporal evolution of the median amount raised in each of the continents.** We count, for each year and each continent, the amount raised in each funding round targeting companies with headquarters in the continent and compute the median of all amounts for the year. We only represent years for which the median value was computed on at least 10 funding rounds.

investors is globally higher in North America compared to other ecosystems, followed by Asia and Europe and then by Oceania, South America and Africa. We see that the maximum mean number of investors is reached after 20020 for all continents, but shows little difference between the various datasets. One possible explanation could be that, even though a significant amount of funding rounds is added between the different dataset downloads, the information on the composition of the investor syndicates is relatively accurate, thus yielding a stable result in the mean number of investors.

A.2.6 Median amount raised

Figure A.12 shows the temporal evolution of the median amount of fund raised for each continent. It is computed by measuring, for each dataset, the number of investors involved in each funding round targeting companies with headquarters in the continent and taking the median over all amounts raised. We opted to use the median as funding round amounts

span several orders of magnitude, which leads to strongly fluctuating mean values.

The median amount raised is generally higher in North America, followed by Asia, Europe, Oceania, South America and Africa. Two separate trends are observed : first, the minimum median amount is reached around 2013 for all regions and all datasets except Africa even though the total number of funding rounds grows throughout the period of study, suggesting a growing share of early-stage funding compared to the total number of funding rounds. Second, when comparing newer datasets (such as the 2023 or 2024 datasets) to older datasets (such as 2019 or 2020), the median amount tends to be lower (see for instance year 2018 in Asia, North America or Europe), which suggests that the funding rounds added between the two datasets are rounds where lower amounts were raised, and thus most likely early-stage funding rounds. This is coherent with the observations presented for early-stage funding (Figs. A.5 and A.6) that were more subject to *a posteriori* consolidation of the data coverage.

A.3 General conclusions on the dataset

We have studied several properties of the Crunchbase dataset, giving us a better view of both its strengths and weaknesses.

First, we see that the dataset is large, with thousands of events (such as funding rounds or company creations) taking place each year in the various continents. These events, however, are unevenly geographically spread with the vast majority taking place in North America, Asia and Europe.

Second, we see that there are stark contrasts in the representation of the different events over time : there can be significant differences in the number of companies in the dataset for a given year depending on when the dataset was accessed, no matter the year. Funding rounds, on the other hand, tend to be less prone to this uncertainty, with the number of funding rounds for a given year showing small variations regardless of when the data was extracted. This holds for all continents except Asia where coverage before 2021 was much smaller.

Third, there is a significant lag in the reporting of the data, meaning that information temporally close to the extraction date tends to be more uncertain across regions and datasets. Quantifying the extent of this reporting lag is difficult, as it depends on a number of factors such as the type of data and the geographical region. For funding data, there seems to be relatively little new information added 2 years after the date of extraction (e.g. looking at year 2018 on Fig. A.4 in North America, we see that many new funding rounds are added in the 2020 dataset and 2021 dataset, with comparatively few funding rounds added in the 2023 and 2024 datasets).

Fourth, there is a massive difference in funding round data coverage raised by Asian companies between the 2019 and 2020 datasets and other datasets, suggesting that results pertaining to Asian ecosystems based on data extracted before 2021 could be impacted by funding rounds being underrepresented.

When working with real-world data, it is important, when feasible, to investigate its limitations in order to understand the associated methodological constraints. No dataset is perfect and the one used here is, as we have shown, no exception to the rule, in spite of its being constantly updated and consolidated to increase the quality of its coverage. Such analyses do not necessarily invalidate existing results, but rather provide us with context for accurate analysis.

Bibliography

[1] Aly Abdelrazek et al. “Topic modeling algorithms and applications: A survey”. In: *Information Systems* 112 (2023), p. 102131.

[2] Daron Acemoglu, Ufuk Akcigit, and William R Kerr. “Innovation network”. In: *Proceedings of the National Academy of Sciences* 113.41 (2016), pp. 11483–11488.

[3] R Adner and DA Levinthal. “The emergence of emerging technologies California Management Review”. In: (2002).

[4] Zsuzsa Ákos et al. “Thermal soaring flight of birds and unmanned aerial vehicles”. In: *Bioinspiration & biomimetics* 5.4 (2010), p. 045003.

[5] Hassanin Al-Fahaam et al. “The design and mathematical model of a novel variable stiffness extensor-contractor pneumatic artificial muscle”. In: *Soft robotics* 5.5 (2018), pp. 576–591.

[6] Gursel Alici et al. “Modeling and experimental evaluation of bending behavior of soft pneumatic actuators made of discrete actuation chambers”. In: *Soft robotics* 5.1 (2018), pp. 24–35.

[7] Mário Almeida-Neto et al. “A consistent metric for nestedness analysis in ecological systems: reconciling concept and measurement”. In: *Oikos* 117.8 (2008), pp. 1227–1239.

[8] Jeff Alstott, Ed Bullmore, and Dietmar Plenz. “powerlaw: a Python package for analysis of heavy-tailed distributions”. In: *PloS one* 9.1 (2014), e85777.

[9] Cristiano Antonelli. *The economics of innovation, new technologies and structural change*. Routledge, 2014.

[10] Rhodri Armour et al. “Jumping robots: a biomimetic solution to locomotion across rough terrain”. In: *Bioinspiration & biomimetics* 2.3 (2007), S65.

[11] David Audretsch. “Entrepreneurship research”. In: *Management decision* 50.5 (2012), pp. 755–764.

[12] David Audretsch et al. “Innovative start-ups and policy initiatives”. In: *Research Policy* 49.10 (2020), p. 104027.

[13] Yousef Bahramzadeh and Mohsen Shahinpoor. “A review of ionic polymeric soft actuators and sensors”. In: *Soft Robotics* 1.1 (2014), pp. 38–52.

[14] Guochao Bai, Jieyu Wang, and Xianwen Kong. “A two-fingered anthropomorphic robotic hand with contact-aided cross four-bar mechanisms as finger joints”. In: *Conference on Biomimetic and Biohybrid Systems*. Springer. 2016, pp. 28–39.

[15] Pierre-Alexandre Balland et al. “The new paradigm of economic complexity”. In: *Research policy* 51.3 (2022), p. 104450.

[16] Andrea Baronchelli et al. “Networks in cognitive science”. In: *Trends in cognitive sciences* 17.7 (2013), pp. 348–360.

[17] Jordi Bascompte et al. “The nested assembly of plant–animal mutualistic networks”. In: *Proceedings of the National Academy of Sciences* 100.16 (2003), pp. 9383–9387.

[18] Ugo Bastolla et al. “The architecture of mutualistic networks minimizes competition and increases biodiversity”. In: *Nature* 458.7241 (2009), pp. 1018–1020.

[19] Federico Battiston, Vincenzo Nicosia, and Vito Latora. “The new challenges of multiplex networks: Measures and models”. In: *The European Physical Journal Special Topics* 226 (2017), pp. 401–416.

[20] Joel AC Baum and Brian S Silverman. “Picking winners or building them? Alliance, intellectual, and human capital as selection criteria in venture financing and performance of biotechnology startups”. In: *Journal of business venturing* 19.3 (2004), pp. 411–436.

[21] Stephen J Beckett. “Improved community detection in weighted bipartite networks”. In: *Royal Society open science* 3.1 (2016), p. 140536.

[22] Hamid Bekamiri, Daniel S Hain, and Roman Jurowetzki. “PatentSBERTa: A Deep NLP based Hybrid Model for Patent Distance and Classification using Augmented SBERT”. In: *arXiv preprint arXiv:2103.11933* (2021).

[23] Cristiano Bellavitis, Christian Fisch, and Rod B McNaughton. “COVID-19 and the global venture capital landscape”. In: *Small Business Economics* (2021), pp. 1–25.

[24] Iz Beltagy, Kyle Lo, and Arman Cohan. “SciBERT: A Pretrained Language Model for Scientific Text”. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 2019, pp. 3615–3620.

[25] Antonin Bergeaud, Yoann Potiron, and Juste Raimbault. “Classifying patents based on their semantic content”. In: *PloS one* 12.4 (2017), e0176310.

[26] Shai Bernstein, Arthur Korteweg, and Kevin Laws. “Attracting early-stage investors: Evidence from a randomized field experiment”. In: *The Journal of Finance* 72.2 (2017), pp. 509–538.

[27] Fabio Bertoni, Massimo G Colombo, and Luca Grilli. “Venture capital financing and the growth of high-tech start-ups: Disentangling treatment from selection effects”. In: *Research policy* 40.7 (2011), pp. 1028–1043.

[28] Luís MA Bettencourt and Jasleen Kaur. “Evolution and structure of sustainability science”. In: *Proceedings of the National Academy of Sciences* 108.49 (2011), pp. 19540–19545.

[29] Lidong Bing et al. “Abstractive Multi-Document Summarization via Phrase Selection and Merging”. In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 2015, pp. 1587–1597.

[30] Bernd Blasius. “Power-law distribution in the number of confirmed COVID-19 cases”. In: *Chaos: An Interdisciplinary Journal of Nonlinear Science* 30.9 (2020).

[31] Peter Michael Blau. *Inequality and heterogeneity: A primitive theory of social structure*. Vol. 7. Free Press New York, 1977.

[32] Victor I Blinnikov, VV Belov, and MA Makarov. “Some problems in the use of the international patent classification”. In: *World Patent Information* 6.2 (1984), pp. 63–68.

[33] Jorn H Block, Geertjan De Vries, and Philipp G Sandner. “Venture capital and the financial crisis: An empirical study across industries and countries”. In: (2010).

[34] Vincent D Blondel et al. “Fast unfolding of communities in large networks”. In: *Journal of statistical mechanics: theory and experiment* 2008.10 (2008), P10008.

[35] Yvonne Blum et al. “Swing leg control in human running”. In: *Bioinspiration & biomimetics* 5.2 (2010), p. 026006.

[36] Moreno Bonaventura et al. “Predicting success in the worldwide start-up network”. In: *Scientific reports* 10.1 (2020), p. 345.

[37] Maxime Bonelli. “The adoption of artificial intelligence by venture capitalists”. In: *Available at SSRN 4362173* (2022).

[38] Javier Borge-Holthoefer et al. “Emergence of consensus as a modular-to-nested transition in communication dynamics”. In: *Scientific reports* 7.1 (2017), p. 41673.

[39] Jean-Philippe Bouchaud et al. “Fluctuations and response in financial markets: the subtle nature ofrandom’price changes”. In: *Quantitative finance* 4.2 (2003), p. 176.

[40] Joseph L Bower and Clayton M Christensen. “Disruptive technologies: catching the wave”. In: (1995).

[41] Frédéric Boyer and Vincent Lebastard. “Exploration of objects by an underwater robot with electric sense”. In: *Conference on Biomimetic and Biohybrid Systems*. Springer. 2012, pp. 50–61.

[42] James A Brander, Raphael Amit, and Werner Antweiler. “Venture-capital syndication: Improved venture selection vs. the value-added hypothesis”. In: *Journal of Economics & Management Strategy* 11.3 (2002), pp. 423–452.

[43] Pontus Braunerhjelm, Sameeksha Desai, and Johan E Eklund. “Regulation, firm dynamics and entrepreneurship”. In: *European Journal of Law and Economics* 40 (2015), pp. 1–11.

[44] Timothy F Bresnahan and Manuel Trajtenberg. “General purpose technologies ‘Engines of growth’?” In: *Journal of econometrics* 65.1 (1995), pp. 83–108.

[45] S Brin and L Page. “The anatomy of a large-scale hypertextual web search engine”. In: *Computer networks and ISDN systems* (1998).

[46] Ross Brown and Colin Mason. “Looking inside the spiky bits: a critical review and conceptualisation of entrepreneurial ecosystems”. In: *Small business economics* 49 (2017), pp. 11–30.

[47] Matteo Bruno et al. “The ambiguity of nestedness under soft and hard constraints”. In: *Scientific reports* 10.1 (2020), p. 19903.

[48] Redouan Bshary and Ronald Noë. “A marine cleaning mutualism provides new insights in biological market dynamics”. In: *Philosophical Transactions of the Royal Society B* 378.1876 (2023), p. 20210501.

[49] Amit Bubna, Sanjiv R Das, and Nagpurnanand Prabhala. “Venture capital communities”. In: *Journal of Financial and Quantitative Analysis* 55.2 (2020), pp. 621–651.

[50] Axel Buchner, Abdulkadir Mohamed, and Armin Schwienbacher. “Diversification, risk, and returns in venture capital”. In: *Journal of Business Venturing* 32.5 (2017), pp. 519–535.

[51] Enrique Burgos et al. “Why nestedness in mutualistic networks?” In: *Journal of theoretical biology* 249.2 (2007), pp. 307–313.

[52] Stephen M Burroughs and Sarah F Tebbens. “Upper-truncated power law distributions”. In: *Fractals* 9.02 (2001), pp. 209–222.

[53] Sebastián Bustos et al. “The dynamics of nestedness predicts the evolution of industrial ecosystems”. In: *PloS one* 7.11 (2012), e49393.

[54] Marcello Calisti, Michele Giorelli, and Cecilia Laschi. “A locomotion strategy for an octopus-bioinspired robot”. In: *Conference on Biomimetic and Biohybrid Systems*. Springer. 2012, pp. 337–338.

[55] Marcello Calisti et al. “An octopus-bioinspired solution to movement and manipulation for soft robots”. In: *Bioinspiration & biomimetics* 6.3 (2011), p. 036002.

[56] Ricardo JGB Campello, Davoud Moulavi, and Jörg Sander. “Density-based clustering based on hierarchical density estimates”. In: *Pacific-Asia conference on knowledge discovery and data mining*. Springer. 2013, pp. 160–172.

[57] Lele Cao et al. “Sourcing Investment Targets for Venture and Growth Capital Using Multivariate Time Series Transformer”. In: *arXiv preprint arXiv:2309.16888* (2023).

[58] Lele Cao et al. “Using deep learning to find the next unicorn: A practical synthesis”. In: *arXiv preprint arXiv:2210.14195* (2022).

[59] Jaime Carbonell and Jade Goldstein. “The use of MMR, diversity-based reranking for reordering documents and producing summaries”. In: *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*. 1998, pp. 335–336.

[60] Théophile Carniel, José Halloy, and Jean-Michel Dalle. “A novel clustering approach to bipartite investor-startup networks”. In: *Plos one* 18.1 (2023), e0279780.

[61] Théophile Carniel et al. “Using natural language processing to find research topics in Living Machines conferences and their intersections with Bioinspiration & Biomimetics publications”. In: *Bioinspiration & Biomimetics* 17.6 (2022), p. 065008.

[62] Anne M Carpenter, Meriel Jones, and Charles Oppenheim. “Consistency of use of the International Patent Classification”. In: *KO KNOWLEDGE ORGANIZATION* 5.1 (1978), pp. 30–32.

[63] Daniel R Cavagnaro et al. “Measuring institutional investors’ skill at making private equity investments”. In: *The Journal of Finance* 74.6 (2019), pp. 3089–3134.

[64] Angelo Cavallo, Antonio Ghezzi, and Raffaello Balocco. “Entrepreneurial ecosystem research: Present debates and future directions”. In: *International entrepreneurship and management journal* 15 (2019), pp. 1291–1321.

[65] Yanto Chandra. “Mapping the evolution of entrepreneurship as a field of research (1990–2013): A scientometric analysis”. In: *PloS one* 13.1 (2018), e0190228.

[66] Bernard Chazelle. “An optimal convex hull algorithm in any fixed dimension”. In: *Discrete & Computational Geometry* 10.4 (1993), pp. 377–409.

[67] Henry Chen et al. “Buy local? The geography of venture capital”. In: *Journal of Urban Economics* 67.1 (2010), pp. 90–102.

[68] Wenbin Chen et al. “Fabrication and dynamic modeling of bidirectional bending soft actuator integrated with optical waveguide curvature sensor”. In: *Soft robotics* 6.4 (2019), pp. 495–506.

[69] Nadia Cheng et al. “Prosthetic jamming terminal device: A case study of untethered soft robotics”. In: *Soft robotics* 3.4 (2016), pp. 205–212.

[70] *CNBC. (2023, June 21). SoftBank to shift from ‘defense mode’ to ‘offense mode,’ says CEO Masayoshi Son.* Retrieved from https://www.cnbc.com/2023/06/21/softbank-to-shift-from-defense-mode-to-offense-mode-says-ceo-masayoshi-son.html.

[71] Daniel Cockayne. “What is a startup firm? A methodological and epistemological investigation into research objects in economic geography”. In: *Geoforum* 107 (2019), pp. 77–87.

[72] Susan Cohen and Yael V Hochberg. “Accelerating startups: The seed accelerator phenomenon”. In: (2014).

[73] Susan Cohen et al. “The design of startup accelerators”. In: *Research Policy* 48.7 (2019), pp. 1781–1797.

[74] Rama Cont and Jean-Philipe Bouchaud. “Herd behavior and aggregate fluctuations in financial markets”. In: *Macroeconomic dynamics* 4.2 (2000), pp. 170–196.

[75] Martina Contisciani, Federico Battiston, and Caterina De Bacco. “Inference of hyperedges and overlapping communities in hypergraphs”. In: *Nature communications* 13.1 (2022), p. 7229.

[76] Álvaro Corral and Álvaro González. "Power law size distributions in geoscience revisited". In: *Earth and Space Science* 6.5 (2019), pp. 673–697.

[77] Michele Coscia. *The Atlas for the Aspiring Network Scientist*. 2021. arXiv: 2101.00863 [cs.CY].

[78] Jerry Courvisanos and Stuart Mackenzie. "Innovation economics and the role of the innovative entrepreneur in economic theory". In: *Journal of Innovation Economics & Management* 2 (2014), pp. 41–61.

[79] Emanuele Cozzo et al. "Contact-based social contagion in multiplex networks". In: *Physical Review E* 88.5 (2013), p. 050801.

[80] Luke Cramphorn, Benjamin Ward-Cherrier, and Nathan F Lepora. "A biomimetic fingerprint improves spatial tactile perception". In: *Conference on Biomimetic and Biohybrid Systems*. Springer. 2016, pp. 418–423.

[81] G Christopher Crawford et al. "Power law distributions in entrepreneurship: Implications for theory and research". In: *Journal of Business Venturing* 30.5 (2015), pp. 696–713.

[82] Douglas Cumming and Na Dai. "Local bias in venture capital investments". In: *Journal of empirical finance* 17.3 (2010), pp. 362–380.

[83] Oscar M Curet et al. "Mechanical properties of a bio-inspired robotic knifefish with an undulatory propulsor". In: *Bioinspiration & biomimetics* 6.2 (2011), p. 026004.

[84] Silvia Dalla Fontana and Ramana Nanda. "Innovating to Net Zero: Can Venture Capital and Start-Ups Play a Meaningful Role?" In: *Entrepreneurship and Innovation Policy and the Economy* 2.1 (2023), pp. 79–105.

[85] Jean-Michel Dalle, Matthijs Den Besten, and Carlo Menon. *Using Crunchbase for economic and managerial research*. 2017.

[86] Partha Dasgupta and Paul A David. "Toward a new economics of science". In: *Research policy* 23.5 (1994), pp. 487–521.

[87] Antonio Davila et al. "The rise and fall of startups: Creation and destruction of revenue and jobs by young companies". In: *Australian Journal of Management* 40.1 (2015), pp. 6–35.

[88] David Devigne et al. "The role of domestic and cross-border venture capital investors in the growth of portfolio companies". In: *Small Business Economics* 40.3 (2013), pp. 553–573.

[89] Jacob Devlin et al. "Bert: Pre-training of deep bidirectional transformers for language understanding". In: *arXiv preprint arXiv:1810.04805* (2018).

[90] Michel-Marie Deza and Elena Deza. *Dictionary of distances*. Elsevier, 2006.

[91] Raymon van Dinter, Bedir Tekinerdogan, and Cagatay Catal. "Automation of systematic literature reviews: A systematic literature review". In: *Information and Software Technology* 136 (2021), p. 106589.

[92] Carsten F Dormann et al. “Indices, graphs and null models: analyzing bipartite ecological networks”. In: (2009).

[93] Will Drover et al. “A review and road map of entrepreneurial equity financing research: venture capital, corporate venture capital, angel investment, crowdfunding, and accelerators”. In: *Journal of management* 43.6 (2017), pp. 1820–1853.

[94] JW Duparré and FC Wippermann. “Micro-optical artificial compound eyes”. In: *Bioinspiration & biomimetics* 1.1 (2006), R1.

[95] Zaeem-Al Ehsan. “Defining a startup-a critical analysis”. In: *Available at SSRN 3823361* (2021).

[96] Thomas R Eisenmann. “Entrepreneurship: A working definition”. In: *Harvard Business Review* 10.5 (2013), pp. 1–3.

[97] Balasubramanian Elango et al. “How venture capital firms differ”. In: *Journal of business venturing* 10.2 (1995), pp. 157–179.

[98] Seymour Epstein. “Integration of the cognitive and the psychodynamic unconscious.” In: *American psychologist* 49.8 (1994), p. 709.

[99] Christian Esposito et al. “Venture capital investments through the lens of network and functional data analysis”. In: *Applied Network Science* 7.1 (2022), p. 42.

[100] Peter R Fallon and Robert EB Lucas. “The impact of financial crises on labor markets, household incomes, and poverty: A review of evidence”. In: *The World Bank Research Observer* 17.1 (2002), pp. 21–45.

[101] Kara L Feilich and George V Lauder. “Passive mechanical models of fish caudal fins: effects of shape and stiffness on self-propulsion”. In: *Bioinspiration & biomimetics* 10.3 (2015), p. 036002.

[102] Samuel Fernando et al. “Optimising robot personalities for symbiotic interaction”. In: *Conference on Biomimetic and Biohybrid Systems*. Springer. 2014, pp. 392–395.

[103] Michel Ferrary and Mark Granovetter. “Social networks and innovation”. In: *The Elgar companion to innovation and knowledge creation*. Edward Elgar Publishing, 2017, pp. 327–341.

[104] Michel Ferrary and Mark Granovetter. “The role of venture capital firms in Silicon Valley’s complex innovation network”. In: *Economy and society* 38.2 (2009), pp. 326–359.

[105] Francesco Ferrati, Moreno Muffatto, et al. “Entrepreneurial finance: emerging approaches using machine learning and big data”. In: *Foundations and Trends® in Entrepreneurship* 17.3 (2021), pp. 232–329.

[106] *Financial Times. (2022, May 17). The mauling of Tiger Global.* Retrieved from https://www.ft.com/content/2a393020-4c43-4bd8-8a88-221459bcee58.

[107] Maurizio Follador et al. “Octopus-inspired innovative suction cups”. In: *Conference on Biomimetic and Biohybrid Systems*. Springer. 2013, pp. 368–370.

[108] Miguel A Fortuna et al. “Nestedness versus modularity in ecological networks: two sides of the same coin?” In: *Journal of animal ecology* (2010), pp. 811–817.

[109] Santo Fortunato. “Community detection in graphs”. In: *Physics reports* 486.3-5 (2010), pp. 75–174.

[110] Santo Fortunato et al. “Science of science”. In: *Science* 359.6379 (2018), eaao0185.

[111] Morgan R Frank et al. “Toward understanding the impact of artificial intelligence on labor”. In: *Proceedings of the National Academy of Sciences* 116.14 (2019), pp. 6531–6539.

[112] Brandon Freiberg and Sandra C Matz. “Founder personality and entrepreneurial outcomes: A large-scale field study of technology startups”. In: *Proceedings of the National Academy of Sciences* 120.19 (2023), e2215829120.

[113] Mariia Garkavenko et al. “Where do you want to invest? predicting startup funding from freely, publicly available web information”. In: *arXiv preprint arXiv:2204.06479* (2022).

[114] Clement Gastaud, Theophile Carniel, and Jean-Michel Dalle. “The varying importance of extrinsic factors in the success of startup fundraising: competition at early-stage and networks at growth-stage”. In: *arXiv preprint arXiv:1906.03210* (2019).

[115] Thomas George Thuruthel et al. “Learning closed loop kinematic controllers for continuum manipulators in unstructured environments”. In: *Soft robotics* 4.3 (2017), pp. 285–296.

[116] Shikhar Ghosh and Ramana Nanda. “Venture capital investment in the clean energy sector”. In: *Harvard Business School Entrepreneurial Management Working Paper* 11-020 (2010).

[117] Paul Gompers and Josh Lerner. “The venture capital revolution”. In: *Journal of economic perspectives* 15.2 (2001), pp. 145–168.

[118] Paul A Gompers et al. “How do venture capitalists make decisions?” In: *Journal of Financial Economics* 135.1 (2020), pp. 169–190.

[119] Juanita González-Uribe. “Exchanges of innovation resources inside venture capital portfolios”. In: *Journal of Financial Economics* 135.1 (2020), pp. 144–168.

[120] Juan P González-Varo and Anna Traveset. “The labile limits of forbidden interactions”. In: *Trends in Ecology & Evolution* 31.9 (2016), pp. 700–710.

[121] Michael Gorman and William A Sahlman. “What do venture capitalists do?” In: *Journal of business venturing* 4.4 (1989), pp. 231–248.

[122] Mark Granovetter. “The problem of embeddedness”. In: *American journal of sociology* 91.3 (1985), pp. 481–510.

[123] Christian Greiner and Michael Schäfer. “Bio-inspired scale-like surface textures and their tribological properties”. In: *Bioinspiration & biomimetics* 10.4 (2015), p. 044001.

[124] Jacopo Grilli, Tim Rogers, and Stefano Allesina. "Modularity and stability in ecological communities". In: *Nature communications* 7.1 (2016), p. 12031.

[125] Maarten Grootendorst. "BERTopic: Neural topic modeling with a class-based TF-IDF procedure". In: *arXiv preprint arXiv:2203.05794* (2022).

[126] Maarten Grootendorst. *KeyBERT: Minimal keyword extraction with BERT.* Version v0.3.0. 2020. DOI: 10.5281/zenodo.4461265. URL: https://doi.org/10.5281/zenodo.4461265.

[127] Guo-Ying Gu et al. "A survey on dielectric elastomer actuators for soft robots". In: *Bioinspiration & biomimetics* 12.1 (2017), p. 011003.

[128] Weiwei Gu, Jar der Luo, and Jifan Liu. "Exploring small-world network with an elite-clique: Bringing embeddedness theory into the dynamic evolution of a venture capital network". In: *Social Networks* 57 (2019), pp. 70–81. ISSN: 0378-8733. DOI: https://doi.org/10.1016/j.socnet.2018.11.002. URL: https://www.sciencedirect.com/science/article/pii/S0378873318302272.

[129] Paulo R Guimaraes Jr. "The structure of ecological networks across levels of organization". In: *Annual Review of Ecology, Evolution, and Systematics* 51 (2020), pp. 433–460.

[130] J. Halloy. "Sustainability of living machines". In: *Living machines: A handbook of research in biomimetics and biohybrid systems (eds, Prescott, Tony and Lepora, Nathan and Verschure, Paul FMJ).* Oxford University Press, 2018.

[131] Kwanghyun Han, Nam-Ho Kim, and Dongjun Shin. "A novel soft pneumatic artificial muscle with high-contraction ratio". In: *Soft robotics* 5.5 (2018), pp. 554–566.

[132] Angela Harris and James J Varellas. "Law and political economy in a time of accelerating crises". In: *Journal of Law and Political Economy* 1.1 (2020).

[133] Ricardo Hausmann and César A Hidalgo. "The network structure of economic output". In: *Journal of economic growth* 16 (2011), pp. 309–342.

[134] Li He et al. "Identifying the gene signatures from gene-pathway bipartite network guarantees the robust model performance on predicting the cancer prognosis". In: *BioMed research international* 2014 (2014).

[135] Laura Hernandez, Annick Vignes, and Stéphanie Saba. "Trust or robustness? An ecological approach to the study of auction and bilateral markets". In: *PloS one* 13.5 (2018), e0196206.

[136] César A Hidalgo. "Economic complexity theory and applications". In: *Nature Reviews Physics* 3.2 (2021), pp. 92–113.

[137] César A Hidalgo and Ricardo Hausmann. "The building blocks of economic complexity". In: *Proceedings of the national academy of sciences* 106.26 (2009), pp. 10570–10575.

[138] César A Hidalgo et al. "The principle of relatedness". In: *Unifying Themes in Complex Systems IX: Proceedings of the Ninth International Conference on Complex Systems 9*. Springer. 2018, pp. 451–457.

[139] César A Hidalgo et al. "The product space conditions the development of nations". In: *Science* 317.5837 (2007), pp. 482–487.

[140] Yael V Hochberg, Alexander Ljungqvist, and Yang Lu. "Whom you know matters: Venture capital networks and investment performance". In: *The Journal of Finance* 62.1 (2007), pp. 251–301.

[141] Yael V Hochberg, Michael J Mazzeo, and Ryan C McDevitt. "Specialization and competition in the venture capital industry". In: *Review of Industrial Organization* 46.4 (2015), pp. 323–347.

[142] John H Holland. "The global economy as an adaptive process". In: *The economy as an evolving complex system* 5 (1988), pp. 117–124.

[143] Thomas Homer-Dixon. "Complexity science". In: *Oxford Leadership Journal* 2.1 (2011), pp. 1–15.

[144] Christian Hopp and Finn Rieder. "What drives venture capital syndication?" In: *Applied Economics* 43.23 (2011), pp. 3089–3102.

[145] Sabrina T Howell et al. *How resilient is venture-backed innovation? evidence from four decades of us patenting*. Tech. rep. National Bureau of Economic Research, 2020.

[146] David H Hsu. "Experienced entrepreneurial founders and venture capital funding". In: *Available at SSRN 584702* (2004).

[147] Jiaqi V Huang and Holger G Krapp. "Closed-loop control in an autonomous biohybrid robot system based on binocular neuronal input". In: *Conference on Biomimetic and Biohybrid Systems*. Springer. 2015, pp. 164–174.

[148] Vernon Ireland and Alex Gorod. *Contribution of complex systems to entrepreneurship*. 2016.

[149] Mikko Jääskeläinen. "Venture capital syndication: Synthesis and future directions". In: *International Journal of Management Reviews* 14.4 (2012), pp. 444–463.

[150] AB Jaffe and G De Rassenfosse. "Patent citation data in social science research: Overview and best practices". In: *J. Assoc. Inf. Sci. Technol.* (2017).

[151] Jasper W James et al. "Tactile Model O: Fabrication and testing of a 3d-printed, three-fingered tactile robot hand". In: *Soft Robotics* 8.5 (2021), pp. 594–610.

[152] Wenkai Jiang et al. "HyperX: A scalable hypergraph framework". In: *IEEE Transactions on Knowledge and Data Engineering* 31.5 (2018), pp. 909–922.

[153] Zhengyang Jiang, Cameron Peng, and Hongjun Yan. "Personality differences and investment decision-making". In: *Journal of Financial Economics* 153 (2024), p. 103776.

[154] Luis O Jimenez and David A Landgrebe. "Supervised classification in high-dimensional space: geometrical, statistical, and asymptotical properties of multivariate data". In: *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 28.1 (1998), pp. 39–54.

[155] Yonghong Jin, Qi Zhang, and Sai-Ping Li. "Topological properties and community detection of venture capital network: Evidence from China". In: *Physica A: Statistical Mechanics and Its Applications* 442 (2016), pp. 300–311.

[156] Boyan Jovanovic and Peter L Rousseau. "General purpose technologies". In: *Handbook of economic growth*. Vol. 1. Elsevier, 2005, pp. 1181–1224.

[157] Kwangmok Jung et al. "Artificial annelid robot driven by soft actuators". In: *Bioinspiration & biomimetics* 2.2 (2007), S42.

[158] Dana Kanze et al. "Evidence that investors penalize female founders for lack of industry fit". In: *Science Advances* 6.48 (2020), eabd7664.

[159] Steven N Kaplan, Berk A Sensoy, and Per Strömberg. "Should investors bet on the jockey or the horse? Evidence from the evolution of firms from early business plans to public companies". In: *The Journal of Finance* 64.1 (2009), pp. 75–115.

[160] Steven N Kaplan and Per ER Strömberg. "Characteristics, contracts, and actions: Evidence from venture capitalist analyses". In: *The journal of finance* 59.5 (2004), pp. 2177–2210.

[161] Màrton Karsai. "Computational Human Dynamics: People, Networks, and Collective Phenomena". PhD thesis. Ecole normale supérieure de Lyon; Laboratoire de l'Informatique du Parallélisme, 2019.

[162] Robert K Katzschmann, Andrew D Marchese, and Daniela Rus. "Autonomous object manipulation using a soft planar grasping manipulator". In: *Soft robotics* 2.4 (2015), pp. 155–164.

[163] Stuart A Kauffman. *Investigations*. Oxford University Press, 2000.

[164] William R Kerr, Ramana Nanda, and Matthew Rhodes-Kropf. "Entrepreneurship as experimentation". In: *Journal of Economic Perspectives* 28.3 (2014), pp. 25–48.

[165] John Maynard Keynes. "The general theory of employment". In: *The quarterly journal of economics* 51.2 (1937), pp. 209–223.

[166] Jeongseon Kim, Soohwan Jeong, and Sungsu Lim. "Link Pruning for Community Detection in Social Networks". In: *Applied Sciences* 12.13 (2022), p. 6811.

[167] Hiroaki Kitano. "Nobel Turing Challenge: creating the engine for scientific discovery". In: *npj Systems Biology and Applications* 7.1 (2021), pp. 1–12.

[168] Mikko Kivelä et al. "Multilayer networks". In: *Journal of complex networks* 2.3 (2014), pp. 203–271.

[169] Virginia Klema and Alan Laub. "The singular value decomposition: Its computation and some applications". In: *IEEE Transactions on automatic control* 25.2 (1980), pp. 164–176.

[170] Anthony W Knapp. *Basic real analysis*. Springer Science & Business Media, 2005.

[171] Frank Hyneman Knight. *Risk, uncertainty and profit*. Vol. 31. Houghton Mifflin, 1921.

[172] Shoichiro Koizumi et al. "Recurrent braiding of thin McKibben muscles to overcome their limitation of contraction". In: *Soft robotics* 7.2 (2020), pp. 251–258.

[173] Tobias Kollmann and Andreas Kuckertz. "Evaluation uncertainty of venture capitalists' investment criteria". In: *Journal of Business Research* 63.7 (2010), pp. 741–747.

[174] Samuel Kortum and Josh Lerner. "Assessing the contribution of venture capital to innovation". In: *RAND journal of Economics* (2000), pp. 674–692.

[175] Helmut Krämer-Eis et al. *Entrepreneurial finance and the Russian war against Ukraine: A survey of European venture capital and private equity investors*. Tech. rep. EIF Working Paper, 2023.

[176] Pietro Landi et al. "Complexity and stability of ecological networks: a review of the theory". In: *Population Ecology* 60 (2018), pp. 319–345.

[177] Chris Larson et al. "A deformable interface for human touch recognition using stretchable carbon nanotube dielectric elastomer sensors and deep neural networks". In: *Soft robotics* 6.5 (2019), pp. 611–620.

[178] Nicole Lazzeri et al. "Towards a believable social robot". In: *Conference on biomimetic and biohybrid systems*. Springer. 2013, pp. 393–395.

[179] Pierre Legendre and Louis Legendre. *Numerical ecology*. Elsevier, 2012.

[180] Josh Lerner and Mark Baker. "An Empirical Analysis of Investment Return Dispersion in Emerging Market Private Equity". In: *The Journal of Private Equity (Retired)* 20.4 (2017), pp. 15–24.

[181] Josh Lerner and Ramana Nanda. "Venture capital's role in financing innovation: What we know and how much we still need to learn". In: *Journal of Economic Perspectives* 34.3 (2020), pp. 237–261.

[182] Josh Lerner, Antoinette Schoar, and Wan Wongsunwai. "Smart institutions, foolish choices: The limited partner performance puzzle". In: *The Journal of Finance* 62.2 (2007), pp. 731–764.

[183] Simon A Levin and Andrew W Lo. "A new approach to financial regulation". In: *Proceedings of the National Academy of Sciences* 112.41 (2015), pp. 12543–12544.

[184] Simon A Levin and Andrew W Lo. "Introduction to PNAS special issue on evolutionary models of financial markets". In: *Proceedings of the National Academy of Sciences* 118.26 (2021), e2104800118.

[185] Lusi Li and Haibo He. "Bipartite graph based multi-view clustering". In: *IEEE Transactions on Knowledge and Data Engineering* (2020).

[186] Ruiqi Li et al. "The evolution of k-shell in syndication networks reveals financial performance of venture capital institutions". In: *Social Networks* 76 (2024), pp. 191–202.

[187] Benyamin Lichtenstein. "Emergence and emergents in entrepreneurship: Complexity science insights into new venture creation". In: *Entrepreneurship Research Journal* 6.1 (2016), pp. 43–52.

[188] Andrew W Lo. "The adaptive markets hypothesis: Market efficiency from an evolutionary perspective". In: *Journal of Portfolio Management, Forthcoming* (2004).

[189] RJ Lock, SC Burgess, and R Vaidyanathan. "Multi-modal locomotion: from animal to application". In: *Bioinspiration & biomimetics* 9.1 (2013), p. 011001.

[190] Magnus Lofstrom, Timothy Bates, and Simon C Parker. "Why are some people more likely to become small-businesses owners than others: Entrepreneurship entry and industry-specific barriers". In: *Journal of Business Venturing* 29.2 (2014), pp. 232–251.

[191] Robert E Lucas Jr. "Econometric policy evaluation: A critique". In: *Carnegie-Rochester conference series on public policy*. Vol. 1. North-Holland. 1976, pp. 19–46.

[192] Ming Luo, Mahdi Agheli, and Cagdas D Onal. "Theoretical modeling and experimental analysis of a pressure-operated soft robotic snake". In: *Soft Robotics* 1.2 (2014), pp. 136–146.

[193] Laurens Van der Maaten and Geoffrey Hinton. "Visualizing data using t-SNE." In: *Journal of machine learning research* 9.11 (2008).

[194] Ian C MacMillan, Robin Siegel, and PN Subba Narasimha. "Criteria used by venture capitalists to evaluate new venture proposals". In: *Journal of Business venturing* 1.1 (1985), pp. 119–128.

[195] Ian C MacMillan, Lauriann Zemann, and PN Subbanarasimha. "Criteria distinguishing successful from unsuccessful ventures in the venture screening process". In: *Venture Capital*. Routledge, 2022, pp. 119–133.

[196] Matteo Magnani et al. "Community detection in multiplex networks". In: *ACM Computing Surveys (CSUR)* 54.3 (2021), pp. 1–35.

[197] Yannick Malevergne and Didier Sornette. *Extreme financial risks: From dependence to risk management*. Springer Science & Business Media, 2006.

[198] Andrew D Marchese, Robert K Katzschmann, and Daniela Rus. "A recipe for soft fluidic elastomer robots". In: *Soft robotics* 2.1 (2015), pp. 7–25.

[199] Manuel Sebastian Mariani et al. "Nestedness in complex networks: observation, emergence, and implications". In: *Physics Reports* 813 (2019), pp. 1–90.

[200] Harry Markowitz. "Portfolio Selection". In: *The Journal of Finance* 7.1 (1952), pp. 77–91.

[201] Alfred Marshall. *Principles of economics: unabridged eighth edition*. Cosimo, Inc., 2009.

[202] Matt Marx and Aaron Fuegi. "Reliance on science: Worldwide front-page patent citations to scientific articles". In: *Strategic Management Journal* 41.9 (2020), pp. 1572–1594.

[203] Colin Mason and Matthew Stark. "What do investors look for in a business plan? A comparison of the investment criteria of bankers, venture capitalists and business angels". In: *International small business journal* 22.3 (2004), pp. 227–248.

[204] Robert M May. "Will a large complex system be stable?" In: *Nature* 238.5364 (1972), pp. 413–414.

[205] Robert M May, Simon A Levin, and George Sugihara. "Ecology for bankers". In: *Nature* 451.7181 (2008), pp. 893–894.

[206] Daniele Mazzei et al. "I-clips brain: A hybrid cognitive system for social robots". In: *Conference on Biomimetic and Biohybrid Systems*. Springer. 2014, pp. 213–224.

[207] B Mazzolai. *The quest for bio-inspiration*. https://www.youtube.com/watch?v=0I_l5kPYiCU.

[208] Paul X McCarthy et al. "The impact of founder personalities on startup success". In: *Scientific Reports* 13.1 (2023), p. 17200.

[209] Paul X McCarthy et al. "The Science of Startups: The Impact of Founder Personalities on Company Success". In: *arXiv preprint arXiv:2302.07968* (2023).

[210] Leland McInnes, John Healy, and Steve Astels. "hdbscan: Hierarchical density based clustering". In: *Journal of Open Source Software* 2.11 (2017), p. 205.

[211] Leland McInnes, John Healy, and James Melville. "Umap: Uniform manifold approximation and projection for dimension reduction". In: *arXiv preprint arXiv:1802.03426* (2018).

[212] Bill McKelvey. "Toward a complexity science of entrepreneurship". In: *Journal of Business Venturing* 19.3 (2004), pp. 313–341.

[213] Penny Mealy, J Doyne Farmer, and Alexander Teytelboym. "Interpreting economic complexity". In: *Science advances* 5.1 (2019), eaau1705.

[214] Katja Mehlhorn et al. "Unpacking the exploration–exploitation tradeoff: A synthesis of human and animal literatures." In: *Decision* 2.3 (2015), p. 191.

[215] Matthew J Michalska-Smith and Stefano Allesina. "Telling ecological networks apart by their structure: A computational challenge". In: *PLoS Computational Biology* 15.6 (2019), e1007076.

[216] Axel Michelsen and Ole Næsbye Larsen. "Pressure difference receiving ears". In: *Bioinspiration & biomimetics* 3.1 (2007), p. 011001.

[217] Tomas Mikolov et al. "Efficient estimation of word representations in vector space". In: *arXiv preprint arXiv:1301.3781* (2013).

[218] Atieh Mirshahvalad et al. “Significant communities in large sparse networks”. In: *PloS one* 7.3 (2012), e33721.

[219] Ben Mitchinson et al. “Perception of simple stimuli using sparse data from a tactile whisker array”. In: *Conference on Biomimetic and Biohybrid Systems*. Springer. 2013, pp. 179–190.

[220] Christoph Molnar. *Interpretable Machine Learning. A Guide for Making Black Box Models Explainable*. https://christophm.github.io/interpretable-ml-book/. 2019.

[221] Marcello Mulas, Manxiu Zhan, and Jörg Conradt. “Integration of biological neural models for the control of eye movements in a robotic head”. In: *Conference on Biomimetic and Biohybrid Systems*. Springer. 2015, pp. 231–242.

[222] Seri Mastura Mustaza et al. “Dynamic modeling of fiber-reinforced soft manipulator: A visco-hyperelastic material-based continuum mechanics approach”. In: *Soft robotics* 6.3 (2019), pp. 305–317.

[223] T Nakata et al. “Aerodynamics of a bio-inspired flexible flapping-wing micro air vehicle”. In: *Bioinspiration & biomimetics* 6.4 (2011), p. 045002.

[224] Ramana Nanda and Matthew Rhodes-Kropf. “Investment cycles and startup innovation”. In: *Journal of financial economics* 110.2 (2013), pp. 403–418.

[225] Frank Neffke, Martin Henning, and Ron Boschma. “How do regions diversify over time? Industry relatedness and the development of new growth paths in regions”. In: *Economic geography* 87.3 (2011), pp. 237–265.

[226] A Nerkar and S Shane. “When do start-ups that exploit patented academic knowledge survive?” In: *Int. J. Ind. Organ.* (2003).

[227] Mark EJ Newman. “Complex systems: A survey”. In: *arXiv preprint arXiv:1112.1440* (2011).

[228] MEJ Newman. “Resource letter cs–1: Complex systems”. In: *American Journal of Physics* 79.8 (2011), pp. 800–810.

[229] Ronald Noë and Peter Hammerstein. “Biological markets”. In: *Trends in Ecology & Evolution* 10.8 (1995), pp. 336–339.

[230] Dimitri Ognibene et al. “Learning epistemic actions in model-free memory-free reinforcement learning: Experiments with a neuro-robotic model”. In: *Conference on Biomimetic and Biohybrid Systems*. Springer. 2013, pp. 191–203.

[231] Preston Ohta et al. “Design of a lightweight soft robotic arm using pneumatic artificial muscles and inflatable sleeves”. In: *Soft robotics* 5.2 (2018), pp. 204–215.

[232] P Oltermann. *Pfizer/BioNTech tax windfall brings Mainz an early Christmas present English. Name-Pfizer Inc; BioNTech SE; Copyright-Copyright Guardian News & Media Limited Dec 27, 2021; Last updated-2021-12-28.*

[233] Adam Paszke et al. “Pytorch: An imperative style, high-performance deep learning library”. In: *Advances in neural information processing systems* 32 (2019).

[234] Claudia Payrató-Borras, Laura Hernández, and Yamir Moreno. “Breaking the spell of nestedness: The entropic origin of nestedness in mutualistic systems”. In: *Physical Review X* 9.3 (2019), p. 031024.

[235] Claudia Payrato-Borras, Laura Hernandez, and Yamir Moreno. “Measuring Nestedness: A comparative study of the performance of different metrics”. In: *arXiv preprint arXiv:2002.00534* 10.21 (2020), pp. 11906–11921.

[236] P Phamduy et al. “Fish and robot dancing together: bluefin killifish females respond differently to the courtship of a robot with varying color morphs”. In: *Bioinspiration & biomimetics* 9.3 (2014), p. 036021.

[237] Hoang Vu Phan, Taesam Kang, and Hoon Cheol Park. “Design and stable flight of a 21 g insect-like tailless flapping wing micro air vehicle with angular rates feedback control”. In: *Bioinspiration & biomimetics* 12.3 (2017), p. 036006.

[238] Benoit Pichon et al. “Telling mutualistic and antagonistic ecological networks apart by learning their multiscale structure”. In: *bioRxiv* (2023), pp. 2023–04.

[239] Stuart L Pimm. “The structure of food webs”. In: *Theoretical population biology* 16.2 (1979), pp. 144–158.

[240] *Pitchbook. (2023, October 2). VCs hope plunging IRR is behind them.* Retrieved from https://pitchbook.com/news/articles/VC-performance-IRR-down-double-digits.

[241] T Prescott. *A Brief History of Living Machines.* https://www.youtube.com/watch?v=tj9Rf6EH58Y.

[242] Alec Radford et al. “Improving language understanding by generative pre-training”. In: (2018).

[243] Juan Ramos et al. “Using tf-idf to determine word relevance in document queries”. In: *Proceedings of the first instructional conference on machine learning*. Vol. 242. 1. Citeseer. 2003, pp. 29–48.

[244] Charles Rathkopf. “Network representation and complex systems”. In: *Synthese* 195 (2018), pp. 55–78.

[245] Bernardo Reisdorfer-Leite et al. “Startup definition proposal using product lifecycle management”. In: *IFIP international conference on product lifecycle management*. Springer. 2020, pp. 426–435.

[246] Zheng Ren and Kamran Mohseni. “A model of the lateral line of fish for vortex sensing”. In: *Bioinspiration & biomimetics* 7.3 (2012), p. 036016.

[247] Erwan Renaudo et al. “Design of a control architecture for habit learning in robots”. In: *Conference on Biomimetic and Biohybrid Systems*. Springer. 2014, pp. 249–260.

[248] Andre Retterath. “Human versus computer: benchmarking venture capitalists and machine learning algorithms for investment screening”. In: *Available at SSRN 3706119* (2020).

[249] Andre Retterath and Reiner Braun. “Benchmarking venture capital databases”. In: *Available at SSRN 3706108* (2020).

[250] Lionel Robbins. *An essay on the nature and significance of economic science*. Ludwig von Mises Institute, 2007.

[251] Jack Roberts and Jon Crall. *alan-turing-institute/distinctipy: v1.2.2*. Version v1.2.2. July 2022. DOI: 10.5281/zenodo.6803948. URL: https://doi.org/10.5281/zenodo.6803948.

[252] Hugo Rodrigue et al. “An overview of shape memory alloy-coupled actuators and robots”. In: *Soft robotics* 4.1 (2017), pp. 3–15.

[253] Mayra Z Rodriguez et al. “Clustering algorithms: A comparative approach”. In: *PloS one* 14.1 (2019), e0210236.

[254] Rudolf P Rohr, Serguei Saavedra, and Jordi Bascompte. “On the structural stability of mutualistic systems”. In: *Science* 345.6195 (2014), p. 1253497.

[255] Paul M Romer. “Endogenous technological change”. In: *Journal of political Economy* 98.5, Part 2 (1990), S71–S102.

[256] Paul M Romer. “The origins of endogenous growth”. In: *Journal of Economic perspectives* 8.1 (1994), pp. 3–22.

[257] Nathan Rosenberg. *Inside the black box: technology and economics*. cambridge university press, 1982.

[258] Dylan Ross, Konstantinos Lagogiannis, and Barbara Webb. “A model of larval biomechanics reveals exploitable passive properties for efficient locomotion”. In: *Conference on Biomimetic and Biohybrid Systems*. Springer. 2015, pp. 1–12.

[259] Philip T Roundy, Mike Bradshaw, and Beverly K Brockman. “The emergence of entrepreneurial ecosystems: A complex adaptive systems approach”. In: *Journal of business research* 86 (2018), pp. 1–10.

[260] Serguei Saavedra, Felix Reed-Tsochas, and Brian Uzzi. “A simple model of bipartite cooperation for ecological and organizational networks”. In: *Nature* 457.7228 (2009), pp. 463–466.

[261] Serguei Saavedra et al. “Strong contributors to network persistence are the most vulnerable to extinction”. In: *Nature* 478.7368 (2011), pp. 233–235.

[262] Tim Sainburg, Leland McInnes, and Timothy Q Gentner. “Parametric UMAP embeddings for representation and semisupervised learning”. In: *Neural Computation* 33.11 (2021), pp. 2881–2907.

[263] Ulrich Schetter. “A Measure of Countries’ Distance to Frontier Based on Comparative Advantage”. In: *CID Research Fellow and Graduate Student Working Paper* 135 (2022).

[264] Axel Schneider et al. “HECTOR, a bio-inspired and compliant hexapod robot”. In: *Conference on Biomimetic and Biohybrid Systems*. Springer. 2014, pp. 427–429.

[265] Scott Shane and Sankaran Venkataraman. “The promise of entrepreneurship as a field of research”. In: *Academy of management review* 25.1 (2000), pp. 217–226.

[266] Scott Shane et al. “Founder passion, neural engagement and informal investor interest in startup pitches: An fMRI study”. In: *Journal of Business Venturing* 35.4 (2020), p. 105949.

[267] JK Shang et al. “Artificial insect wings of diverse morphology for flapping-wing micro air vehicles”. In: *Bioinspiration & biomimetics* 4.3 (2009), p. 036002.

[268] Prafull Sharma and Yingbo Li. “Self-supervised contextual keyword and keyphrase retrieval with self-labelling”. In: (2019).

[269] Yu She et al. “Modeling and validation of a novel bending actuator for soft robotics applications”. In: *Soft Robotics* 3.2 (2016), pp. 71–81.

[270] R Shepherd et al. *Perspective for Soft Robotics*. https://www.youtube.com/watch?v=0I_l5kPYiCU.

[271] Beth Silverstein and Carl Osborne. “Strategies for attracting healthcare venture capital.” In: *Journal of commercial biotechnology* 8.4 (2002).

[272] Benno I Simmons et al. “bmotif: A package for motif analyses of bipartite networks”. In: *Methods in Ecology and Evolution* 10.5 (2019), pp. 695–701.

[273] Sahar Sohangir and Dingding Wang. “Improved sqrt-cosine similarity measurement”. In: *Journal of Big Data* 4.1 (2017), pp. 1–13.

[274] Morten Sørensen. “How smart is smart money? A two-sided matching model of venture capital”. In: *The Journal of Finance* 62.6 (2007), pp. 2725–2762.

[275] Olav Sorenson and Toby E Stuart. “Syndication networks and the spatial distribution of venture capital investments”. In: *American journal of sociology* 106.6 (2001), pp. 1546–1588.

[276] Didier Sornette and Anders Johansen. “Large financial crashes”. In: *Physica A: Statistical Mechanics and its Applications* 245.3-4 (1997), pp. 411–422.

[277] Aswathy Sreenivasan and M Suresh. “Agility adaptability and alignment in startups”. In: *Journal of Science and Technology Policy Management* (2023).

[278] FC Stam, Ben Spigel, et al. “Entrepreneurial ecosystems”. In: *USE Discussion paper series* 16.13 (2016).

[279] Phillip PA Staniczenko, Jason C Kopp, and Stefano Allesina. “The ghost of nestedness in ecological networks”. In: *Nature communications* 4.1 (2013), pp. 1–6.

[280] Alexander G Steele, Alexander Hunt, and Appolinaire C Etoundi. “Development of a bio-inspired knee joint mechanism for a bipedal robot”. In: *Conference on Biomimetic and Biohybrid Systems*. Springer. 2017, pp. 418–427.

[281] Howard H Stevenson. “Why entrepreneurship has won”. In: *Coleman White Paper* 2.4 (2000), p. 483.

[282] Agostino Stilli, Helge A Wurdemann, and Kaspar Althoefer. “A novel concept for safe, stiffness-controllable robot links”. In: *Soft robotics* 4.1 (2017), pp. 16–22.

[283] Lewi Stone. “The stability of mutualism”. In: *Nature communications* 11.1 (2020), p. 2648.

[284] Arho Suominen, Hannes Toivanen, and Marko Seppänen. “Firms’ knowledge profiles: Mapping patent data with unsupervised learning”. In: *Technological Forecasting and Social Change* 115 (2017), pp. 131–142.

[285] Samir Suweis et al. “Emergence of structural and dynamical properties of ecological mutualistic networks”. In: *Nature* 500.7463 (2013), pp. 449–452.

[286] Shura Suzuki et al. “Quadruped gait transition from walk to pace to rotary gallop by exploiting head movement”. In: *Conference on biomimetic and biohybrid systems*. Springer. 2016, pp. 532–539.

[287] Nicholas S Szczecinski and Roger D Quinn. “MantisBot changes stepping speed by entraining CPGs to positive velocity feedback”. In: *Conference on Biomimetic and Biohybrid Systems*. Springer. 2017, pp. 440–452.

[288] Nicholas S Szczecinski et al. “Modeling the dynamic sensory discharges of insect campaniform sensilla”. In: *Conference on Biomimetic and Biohybrid Systems*. Springer. 2020, pp. 342–353.

[289] Alexander V Terekhov, Guglielmo Montone, and J Kevin O’Regan. “Knowledge transfer in deep block-modular neural networks”. In: *Conference on Biomimetic and Biohybrid Systems*. Springer. 2015, pp. 268–279.

[290] Emanuele Teti, Alberto Dell’Acqua, and Ada Bovsunovsky. “Diversification and size in venture capital investing”. In: *Eurasian Business Review* (2024), pp. 1–26.

[291] *The Economist. (2022, July 13). Which covid-19 vaccine saved the most lives in 2021?* Retrieved from https://www.economist.com/graphic-detail/2022/07/13/which-covid-19-vaccine-saved-the-most-lives-in-2021.

[292] Elisa Thébault and Colin Fontaine. “Stability of ecological communities and the architecture of mutualistic and trophic networks”. In: *Science* 329.5993 (2010), pp. 853–856.

[293] Patricia H Thornton, Domingo Ribeiro-Soriano, and David Urbano. “Socio-cultural factors and entrepreneurial activity: An overview”. In: *International small business journal* 29.2 (2011), pp. 105–118.

[294] Sarah Tiba, Frank J van Rijnsoever, and Marko P Hekkert. "Sustainability startups and where to find them: Investigating the share of sustainability startups across entrepreneurial ecosystems and the causal drivers of differences". In: *Journal of Cleaner Production* 306 (2021), p. 127054.

[295] Alexandra Gabriela Ţiţan. "The efficient market hypothesis: Review of specialized literature and empirical research". In: *Procedia Economics and Finance* 32 (2015), pp. 442–449.

[296] Alice Tonazzini et al. "Plant root strategies for robotic soil penetration". In: *Conference on Biomimetic and Biohybrid Systems*. Springer. 2013, pp. 447–449.

[297] RS Trask, Hugo R Williams, and IP Bond. "Self-healing polymer composites: mimicking nature to enhance performance". In: *Bioinspiration & Biomimetics* 2.1 (2007), P1.

[298] Chengyi Tu, Joel Carr, and Samir Suweis. "A data driven network approach to rank countries production diversity and food specialization". In: *Plos one* 11.11 (2016), e0165941.

[299] Werner Ulrich, Mário Almeida-Neto, and Nicholas J Gotelli. "A consumer's guide to nestedness analysis". In: *Oikos* 118.1 (2009), pp. 3–17.

[300] Brian Uzzi. "The sources and consequences of embeddedness for the economic performance of organizations: The network effect". In: *American sociological review* (1996), pp. 674–698.

[301] Seyed Mohammad Mirkhalaf Valashani and Francois Barthelat. "A laser-engraved glass duplicating the structure, mechanics and performance of natural nacre". In: *Bioinspiration & biomimetics* 10.2 (2015), p. 026005.

[302] Fernanda S Valdovinos. "Mutualistic networks: moving closer to a predictive theory". In: *Ecology letters* 22.9 (2019), pp. 1517–1534.

[303] Lorenzo Vannucci et al. "Eye-head stabilization mechanism for a humanoid robot tested on human inertial data". In: *Conference on Biomimetic and Biohybrid Systems*. Springer. 2016, pp. 341–352.

[304] Ashish Vaswani et al. "Attention is all you need". In: *Advances in neural information processing systems* 30 (2017).

[305] Conceicao Vedovello. "Firms' R&D activity and intensity and the university–enterprise partnerships". In: *Technological forecasting and social change* 58.3 (1998), pp. 215–226.

[306] Subhashini Venugopalan and Varun Rai. "Topic based classification and pattern identification in patents". In: *Technological Forecasting and Social Change* 94 (2015), pp. 236–250.

[307] Tamás Vicsek and Anna Zafeiris. "Collective motion". In: *Physics reports* 517.3-4 (2012), pp. 71–140.

[308] Alex Villanueva, Colin Smith, and Shashank Priya. "A biomimetic robotic jellyfish (Robojelly) actuated by shape memory alloy composite actuators". In: *Bioinspiration & biomimetics* 6.3 (2011), p. 036004.

[309] Nguyen Xuan Vinh, Julien Epps, and James Bailey. "Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance". In: *The Journal of Machine Learning Research* 11 (2010), pp. 2837–2854.

[310] Csaba Virágh et al. "Flocking algorithm for autonomous flying robots". In: *Bioinspiration & biomimetics* 9.2 (2014), p. 025012.

[311] Elena Voita et al. "Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned". In: *arXiv preprint arXiv:1905.09418* (2019).

[312] Ulrike Von Luxburg. "A tutorial on spectral clustering". In: *Statistics and computing* 17 (2007), pp. 395–416.

[313] Sazali Abdul Wahab, Raduan Che Rose, and Suzana Idayu Wati Osman. "Defining the concepts of technology and technology transfer: A literature analysis". In: *International business research* 5.1 (2012), pp. 61–71.

[314] Hao Wang et al. "A study of graph-based system for multi-view clustering". In: *Knowledge-Based Systems* 163 (2019), pp. 1009–1019.

[315] Yingfan Wang et al. "Understanding how dimension reduction tools work: an empirical approach to deciphering t-SNE, UMAP, TriMAP, and PaCMAP for data visualization". In: *The Journal of Machine Learning Research* 22.1 (2021), pp. 9129–9201.

[316] Duncan J Watts and Steven H Strogatz. "Collective dynamics of 'small-world'networks". In: *nature* 393.6684 (1998), pp. 440–442.

[317] Quentin Waymel et al. "Impact of the rise of artificial intelligence in radiology: what do radiologists think?" In: *Diagnostic and interventional imaging* 100.6 (2019), pp. 327–336.

[318] Victoria A Webster et al. "Aplysia californica as a novel source of material for biohybrid robots and organic machines". In: *Conference on Biomimetic and Biohybrid Systems*. Springer. 2016, pp. 365–374.

[319] Ying Wei et al. "A novel, variable stiffness robotic gripper based on integrated soft actuating and particle jamming". In: *Soft Robotics* 3.3 (2016), pp. 134–143.

[320] Anthony Westphal, Daniel Blustein, and Joseph Ayers. "A biomimetic neuronal network-based controller for guided helicopter flight". In: *Conference on Biomimetic and Biohybrid Systems*. Springer. 2013, pp. 299–310.

[321] Johan Wiklund et al. "The future of entrepreneurship research". In: *Entrepreneurship Theory and Practice* 35.1 (2011), pp. 1–9.

[322] Wilton Wilton and William Toh. "Determinants of entrepreneurship: A framework for successful entrepreneurship". In: *World Review of Entrepreneurship, Management and Sustainable Development* 8.3 (2012), pp. 285–296.

[323] Svante Wold, Kim Esbensen, and Paul Geladi. “Principal component analysis”. In: *Chemometrics and intelligent laboratory systems* 2.1-3 (1987), pp. 37–52.

[324] Thomas Wolf et al. “Transformers: State-of-the-art natural language processing”. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. 2020, pp. 38–45.

[325] David H Wolpert and William G Macready. “No free lunch theorems for optimization”. In: *IEEE transactions on evolutionary computation* 1.1 (1997), pp. 67–82.

[326] Hong Xiong and Ying Fan. “How to Better Identify Venture Capital Network Communities: Exploration of A Semi-Supervised Community Detection Method”. In: *Journal of Social Computing* 2.1 (2021), pp. 27–42.

[327] Feiyu Xu et al. “Explainable AI: A brief survey on history, research areas, approaches and challenges”. In: *Natural Language Processing and Chinese Computing: 8th CCF International Conference, NLPCC 2019, Dunhuang, China, October 9–14, 2019, Proceedings, Part II 8*. Springer. 2019, pp. 563–574.

[328] Yan Yang and Hao Wang. “Multi-view clustering: A survey”. In: *Big Data Mining and Analytics* 1.2 (2018), pp. 83–107.

[329] Yang Yang et al. “Bioinspired robotic fingers based on pneumatic actuator and 3D printing of smart material”. In: *Soft robotics* 4.2 (2017), pp. 147–162.

[330] Akira Yoshino. “The birth of the lithium-ion battery”. In: *Angewandte Chemie International Edition* 51.24 (2012), pp. 5798–5800.

[331] Davide Zappetti et al. “Phase changing materials-based variable-stiffness tensegrity structures”. In: *Soft robotics* 7.3 (2020), pp. 362–369.

[332] Xin Zhang et al. “Modeling risk contagion in the venture capital market: a multilayer network approach”. In: *Complexity* 2019 (2019), pp. 1–11.

[333] Zihao Zhang and Bradley Cantrell. “Cultivated wildness: Technodiversity and wildness in machines”. In: *arXiv preprint arXiv:2305.02328* (2023).

[334] Shunzhi Zhu, Lizhao Liu, and Yan Wang. “Information retrieval using Hellinger distance and sqrt-cos similarity”. In: *2012 7th International Conference on Computer Science & Education (ICCSE)*. IEEE. 2012, pp. 925–929.

[335] Bob Zider. “How venture capital works”. In: *Harvard business review* 76.6 (1998), pp. 131–139.