Master's Thesis

Modelling portfolios of cyber-related emerging technologies: a complex-system approach

MFE Supervisor:
Prof. Negar Kiyavash

Author:
Anita Mezzetti

Company Supervisors:
Dr. Alain Mermoud
Dr. Dimitri Percia David

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

Financial Engineering

July 30, 2021

To Edoardo,
my North Star.

Acknowledgements

The completion of this thesis would have not been possible without the guide of Dr. Dimitri Percia David, who has followed me in this project since the day I arrived at the CYD Campus. He has been a great mentor: he is always open to give support and help, creating the best working conditions and, at the same time, he has always given me the freedom necessary to develop my ideas. He has always provided me with respectful and productive feedback. I would like to thank also Dr. Alain Mermoud, who has done a great work in organising the different projects at the CYD Campus and has actively participated to this project. I would like to pay my special regards to both of them for giving me the opportunity to present my project to other researchers from different institutions during the armasuisse Alp Retreat in Gstaad on July 1.

Besides my supervisors at the CYD Campus, Prof. Thomas Maillart was so kind to make available his expertise, participating in various meeting at crucial points and giving many pieces of advice. His ideas and his approach have been particularly useful because they are always fresh and non-trivial. Moreover, he is also one of the authors of the article on which this project relies on [28].

At EPFL, I am lucky to be followed by Prof. Negar Kiyavash, who was my professor during the Network Analytics class. Her lectures have inspired me to continue my studies in this subject and I am honoured she accepted to be part also of this journey.

Moreover, I am grateful to Chi Thang Duong, who has been always available and helpful. Especially at the beginning, when I was a bit lost, he helped me get oriented thanks to his advanced skills.

I am happy to have shared this experience with other interns, especially Santiago Antón Moreno who has started his project at the CYD Campus with me. I feel grateful I had the opportunity to meet some new people in person despite the current situation.

Last but not least, I would like to express my deep gratitude for my family and my friends. Even if there often are hundreds of kilometres that separate us, I always feel them really close.

Contents

List of Figures

List of Tables

Abstract

The cybersecurity technological landscape is a complex ecosystem in which entities – such as companies and technologies – influence each other in a non-trivial manner. Understanding influence measures of each entity is central when it comes to take informed technological investment decisions.

To recognise the mutual influence of companies and technologies in cybersecurity, we consider a bi-partite graph that links companies and technologies. Then, we weight nodes by applying a recursive algorithm based on the method of reflection. This endeavour helps to assign a measure of how an entity impacts the cybersecurity market. Our results help (i) to measure the magnitude of influence of each entity, (ii) decision-makers to address more informed investment strategies, according to their preferences.

Investors can customise the algorithm by indicating which external factors –such as previous investments and geographical positions– are relevant for them. They can select their interests among a list of properties about companies and technologies and weights them according to their needs. This preferences are automatically included in the algorithm and the TechRank's scores changes accordingly.

Chapter 1

Introduction

This thesis arises from a collaboration between EPFL and the cyberdefence Campus (CYD), a newly established research centre that is part of the Swiss Department of Defence (armasuisse).[1] This work is part of a wider project focused on technology forecasting and market monitoring in cyberdefence. Its main goal is investigating the innovation structures and the dynamics underlying the hype cycle of novel technologies. This hype cycle should be demystified and modelled by managing portfolios of emerging and/or disruptive technologies related to cyberdefence, and applying real-option models to capture the net-present value of uncertain innovation portfolios for cyberdefence. As a matter of fact, beyond the hype associated with the excitement of their development, emerging technologies are prone to uncertainties regarding the future benefits and risks they carry for society. Here, we recognise that there is a dearth of knowledge regarding the quantitative understanding of risks-adjusted benefits of novel technologies. Therefore, a quantitative framework for continuous monitoring of their perceived and achieved risk-adjusted benefits is needed. To sum up, the big and complex question underlying this broad work is:

Given an ever-accelerating flow of emerging and/or disruptive technologies, how should cyberdefence resources be invested through time (e.g., training, labour, as well as direct and indirect funding), and into which technologies should cyberdefence actors invest?

Specifically, this thesis proposes the TechRank algorithm: a methodology that assigns a score to each entity (i.e. technologies and companies) according to its influence in the technological ecosystem. The objective is to contribute to helping decision-makers to make more informed-decisions for investments. Our focus is the complex ecosystem of the cybersecurity market [21], but this work is extendable to many fields.

[1] armasuisse website: homepage
armasuisse website: CYD Campus

In particular, the cybersecurity technological landscape is crawling with start-ups and interesting innovations. As a matter of fact, due to the constantly increasing number of cyber-attacks –dangerous in terms of costs and leaks of sensitive data– and the increased usage of the Internet of Things (IoT); investments in cybersecurity are growing exponentially and we expect they will continue to increase for the next years.[2] According to Bloomberg, "the global cybersecurity market size is expected to reach USD $326.4 billion by 2027, registering a compound annual growth rate (CAGR) of 10.0% from 2020 to 2027".[3] In cybersecurity, the role of new technologies is extremely relevant, because the technology landscape is continuously evolving and there are always new risk factors. Therefore, our research is a right response to the main current challenges in this field. As an example, during the pandemic the number of cyberattacks has increased even more, especially because lock-downs have forced many wealthy people and office managers to work at home, increasing the opportunities for fraudsters to exploit communications links.[4]

To develop the TechRank algorithm, we first model and map the ecosystem of entities (i.e., technologies, companies) from *Crunchbase* (see Section 3.1) using a *bi-partite network*. The bi-partite network structure is suitable for describing this *complex system*, composed of heterogeneous entities interacting with each other, and whose behaviour is hard to capture due to the nodes relationships of dependencies and competitions.

Once we have created the bi-partite network, we evaluate the relative influence of its nodes in the whole ecosystem by adapting a recursive algorithm that returns a network-centrality measure.

This methodology should help decision-makers and investors to quantitatively assess the influence of the entities that constitute the cybersecurity ecosystem, reducing potential investment uncertainties. Indeed, the cybersecurity market is shaped by complex and fast-paced technological developments and spotting the best investments strategy is a non-trivial task.

As a matter of fact, if a certain research area is too much inflated, startups working on it are likely to face many issues. Around 90% of startups fail and the most common reason, 42% of cases, is due to misreading market demand. Secondly, 29% of times startups fail because they run out of funding and personal money.[5] These risks concern also established companies: some businesses fail to stay atop their industries when they face certain types of

[2] The New York Times: "As Cyberattacks Surge, Security Start-Ups Reap the Rewards" by Erin Woo (July 26, 2021).
Yahoo Finance: "Microsoft Securing its Position with Cybersecurity Investments" by TipRanks (July 20, 2021).

[3] Bloomberg: "Global Cybersecurity Market Could Exceed $320 Billion in Revenues by 2027" (July 29, 2020).

[4] Financial Times: "Cyber attacks multiply on wealthy investors" by Matthew Vincent (March 18, 2021).

[5] Findstack: "The Ultimate List of Startup Statistics for 2021" by Jack Steward.

market and technological changes [12]. Christensen [12], in his well-known *The Innovator's Dilemma* book, highlights that also well-managed companies may break down, for instance because they invest aggressively in new technologies. Therefore, our goal of selecting the right technologies to invest in and creating an optimal investment strategy is a necessary response to the main issues of entrepreneurship [10].

Our research takes inspiration from the well known Google's PageRank algorithm [35], whose goal is to rank Web pages according to readers interests. Using a similar approach applied to bi-partite networks, we aim to assign a score companies and technologies. In this way, investors can invest into the entities that are the most influential –according to our network-centrality measure–, and according to their preferences. We want to help decision-makers to make complex decisions with more awareness without compromising their preferences.

To achieve this last goal, part of the TechRank algorithm focuses on investors' preferences. Our idea is to provide them with a list of properties about the entities and let them choose which reflect their interests and how much. Their choices are given as input to the algorithm and the final result, the entities' score, is influenced by them. This enables investors to select a personalised portfolio strategy using a quantitative methodology.

Chapter 2

Background and Related Work

This chapter surveys previous work that has been done about applying network analytics to investigate nodes centrality, bi-partite networks, and applications for optimising investments. It also introduces the theoretical framework in which our work is located.

2.1 Centrality measures

The study of networks has deep roots, and the start of the search about the importance of their nodes was not late in coming. Centrality analysis helps to understand the importance and influence of different nodes within a network. In other words, ranking nodes to identify highly central or topologically important nodes is the main goal of centrality measures. In 1948, Bavelas [3] developed the first centrality measure: a structural centrality measure in the context of social graphs. It was 1972 when Bonacich [6] proposed the eigenvector centrality, and Freeman [18], four years later, summarised the existing literature to create a formal mathematical framework for centrality, including degree, closeness, and betweenness. As Freeman mentioned in his established article, the combination of different kinds of centrality measures might be appropriate in a given application. This is the basis from which our research arises: a work which aims to measure the centrality of each entity by combining and expanding some centrality measures, adding some novel ideas (described in Chapter 4).

Specifically, the simplest centrality measure is the *Degree* of a node, which counts the number of edges linked to the node in a network. It denotes the number of neighbours of a node. It is a valid metric, because it is easy and fast to calculate and is based only on the neighbourhood of each node: we do not have to analyse the whole network in case we are interested only in some vertices. For these reasons, we decide to use it as starting point in our algorithm. One of the drawbacks is that, if the degree tells us which nodes have a lot of connections, it does not show which is in the "middle" of the network: we may have two nodes with the same degree, but one very central and the other one peripheral. For instance, Figure 2.1 shows that node 0 and node 1 have both degree 3, but 1 is central (many shortest paths between two vertices include it), while 0 is peripheral. Therefore, the degree centrality can be considered as a local centrality measure; it cannot capture the influence across nodes within the graph.

Another important centrality measure is *Closeness*, which measures how long it will take for information to spread from a certain node to others in the network. Specifically, closeness is the defined as inverse of *fairness*, i.e. the sum of distances with all other nodes. In our case, this is not suitable because we are not interested in the distance between two entities: we want to capture the influence of a node, but for the technological landscape this does not depend on shortest distances. The same holds for *Betweenness centrality*, based on the number of shortest paths passing through the node. Betweenness centrality counts how many times a node acts as a bridge in the shortest path between a couple of other vertices [39, 19]. Considering that a company that works on many technologies is more likely to be influential and, simultaneously, a spread technology is probably relevant, we decide to use degree centrality as starting point. However, we need to refine this metric in order to capture the complex structure of influence of the whole technological landscape.

Recently, researchers have been also focusing on the task of identifying the top-k nodes of a complex network. This is a challenging task, especially if we do not possess knowledge of the entire network. However, our goal is different: we want to rank the nodes, not identify a group of them, and this is a much less explored field. The main issue that we meet when ranking a node is that we need to compute the centrality of all the nodes and compare them to extract the rank [38]. This is not always feasible due to the size of the network. In order to overcome this problem, Saxena and Iyengar [38] tried to estimate the global centrality of a node without analysing the whole network.

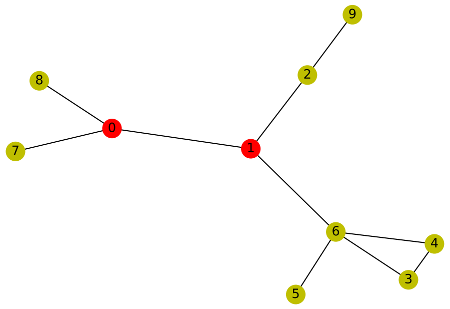


Figure 2.1: Difference between a central and a peripheral node.

2.1.1 Page Rank

As mentioned, since the '70s a large number of centrality measures has been introduced in the literature. An important step has been done with the development of the PageRank algorithm [35]. In 1997, Page and Brin, Google's founders, developed the PageRank algorithm, a method that attempts to rank Web pages objectively. The fast-growing Web was creating new challenges, so researchers needed to find a way to manage this large and diversified amount of information. They came up with a solution that produces a score for each page capturing its influence starting from its relations within the graph. Moreover, the uncontrolled growth of web pages can be easily compared to the start-ups boom of recent years. Both require an efficient ranking methodology, capable to detect the most influential entities.

Since, and thanks to, the development of the PageRank algorithm, a lot of work has been done to amplify its applications and improve it. Xing and Ghorbani [44], as an extension of the original PageRank method, proposed the weighted PageRank (WPR) algorithm. The main difference is that this algorithm assigns larger rank values to more important pages instead of dividing the rank value of a page evenly among its outlink pages.[1] In other words, each outlink page gets a value proportional to its popularity, considering the weights of the links. This approach is useful for our case because we want our algorithm to be customizable according to investors' preferences and we take inspiration from this work when including investors' choices (see Section 4.2).

However, the PageRank and the WPR algorithms both have a limitation: they do not enable to leverage n-partite structures. As a matter of fact, Web pages can all be linked to one another, while we work with two different kinds of entities, technologies and companies, whose characteristics are different and their ranks must stay separate. PageRank is not able to capture this important structural property. Considering we need a structure that keeps track of these two groups of nodes, we decide to use bi-partite networks.

2.1.2 Bi-partite networks

Networks are a fundamental tool for capturing relationships in many fields, from the nervous system to the Internet. Graphs (G) are composed by vertices (V) and edges (E) –$G = (V, E)$–, thanks to which we can build links and mathematically analyse many properties of the whole system, and of the singular entities as well. Usually, to graphically represent a real system, we model it to a high level of abstraction synthesising its information into simple graph frameworks. Thanks to this strategy, we can study the

[1] Given a Web page W, an inlink of W is a link (URL) of another Web page that includes a link pointing to W. An outlink of W is a link appearing in W which points to another Web page. From IGI Global: What is Inlink and Outlink.

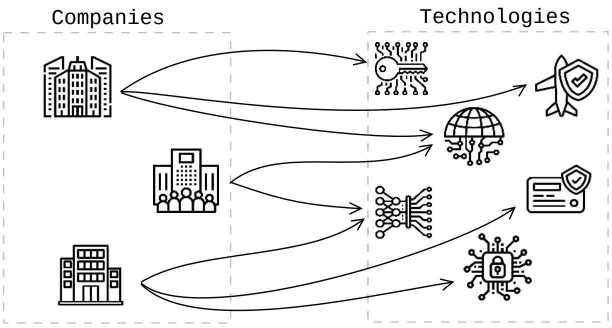


Figure 2.2: Scheme of a bi-partite structure composed by companies (layer 1) and technologies (layer 2).

same properties – as the small-world one– and use similar tools for networks belonging to completely different fields. The drawback is that this simplification contributes to an information loss in the modelling process. In other words, simple networks structures might in general discard important information about the structure and function of the original system [29]. As a consequence, the failure of a very small fraction of nodes in one network may lead to the complete fragmentation of a system and dramatic consequences [8]. To solve the problem, there have been introduced many amplifications to the simple structure $G = (V, E)$, resulting in graphs with more powerful features. For instance, in case vertices are connected by relationships of different kinds, it is better to work with *multiplex networks*: networks where each node appears in a set of different layers, and each layer describes all the edges of a given type [1].

When it is possible to distinguish the nature of the edges, multiplex networks are an effective approach, which starts from embedding the edges in different layers according to their type. However, even if we have two kinds of nodes, companies and technologies, the nature of the edges is only one. Therefore, a more suitable approach is bi-partite networks. Bi-partite networks are good representation for describing the technological landscape (see Figure 2.2): there are two sets of nodes, companies and technologies, which are interconnected but do not present edges within the same set.

In the light of this, we search for applications of the PageRank algorithm to bi-partite structures [5, 16, 41, 28]. In particular, the extension by Klein, Maillart, and Chuang [28] suits our needs. Let us further explain it in the next section.

2.1.3 The method of reflection: an application to Wikipedia

Nowadays, network theory, thanks to its adaptability, is useful in many fields. Therefore, it often happens that a certain study, which focuses on a specific application, can be relevant also in other areas. The work by Klein, Maillart, and Chuang [28], which forms the basis of our research, analyses the influence of editors and articles on Wikipedia, leveraging a bi-partite structure.

This article is relevant because Klein, Maillart, and Chuang [28] create an extension of PageRank for a scheme (editors and articles) that can be easily applied to the technological landscape (companies and technologies). Of course, our analysis is different in the sense that the relationships are very complicated and involve many more external factors, even if we decide not to consider the evolution throughout time. A major benefit of their analysis is that they start from an unweighted graph: the links between authors and articles do not have weights. This reflects our condition, in which start-ups and technologies are not linked by weighted edges. In particular, Klein, Maillart, and Chuang [28] develop a recursive algorithm by which the two entities (editors and articles) both contribute to the quality - for the articles – or the expertise - for the authors - of the other one. It is a cooperation and coordination scheme which aims at solving common interaction problems that emerge with such structures. Even if the technological landscape is a more complex scenario, their goals reflect ours.

Apart from this article, to the best of our knowledge, there is not much work about how to measure nodes' influence using an unweighted network scheme of coordination and cooperation. Other work [5, 16, 41] which tries to extend the PageRank algorithm to multiplex networks is not useful for our case. As a matter of fact, they assume that only some clusters of the graph are multiplex networks and they extend the PageRank algorithm only to analyse sub-graph centrality. Bi-partite networks are a pretext for transforming directed networks into undirected ones with twice the number of vertexes. Moreover, they all study directed graphs, while we are searching a suitable solution for an undirected and unweighted network.

More in general, the main objective to define centrality measures is to rank a node and, to the best of our knowledge, little work has been done when it comes to assess the importance/influence of entities for optimising investments. In this work, we focus on the cybersecurity market, which is particularly complex [21]. The idea of measuring the global rank of a node starting from local information and other centrality measure is still an open research question in many sectors [39].

The bi-partied random walker method developed by Klein, Maillart, and Chuang [28] for Wikipedia starts by building the adjacency matrix $M_{e,a}$ that takes value 1 if editor e has edited article a and 0 otherwise. In other words, it takes track of all the editors' contributions. We get that $M_{e,a} \in \mathbb{R}^{n_e, n_a}$, where n_e is the total number of editors and n_a the total number of articles. If we sort editors in the matrix by the number of articles they have contributed to, we understand that $M_{e,a}$ can often have a triangular structure. Afterwards, they assign a contribution value (expertise) to each editor and a quality value to each article using the degree metrics. In particular, they measure the expertise of an editor (w_e^0) by summing the number of articles he has worked on and the quality of an article (w_a^0) by summing the number of editors who worked on that:

$$\begin{cases} w_e^0 & = \sum_{a=1}^{n_a} M_{e,a} = k_e \\ w_a^0 & = \sum_{e=1}^{n_e} M_{e,a} = k_a \end{cases}. \tag{2.1}$$

They represent the starting point of the recursive algorithm.

In order to describe how the algorithm moves to the next step, please note the algorithm is a Markov process: the step w^n (where $w^n = w^n(\alpha, \beta)$) depends only to information available at w^{n-1}. At each step, the algorithm incorporates information about the expertise of editors and the quality of articles, leveraging the bi-partite network structure. The process can be seen as a random walker with jumps, whose transition probability is zero in case $M_{e,a} = 0$. The intuition is similar to the PageRank one. At this point, the authors define two variables for the transition probability, $G_{e,a}(\beta)$ and $G_{e,a}(\alpha)$), that explain how to move to one step to the next one. In particular, $G_{e,a}(\beta)$ represents the probability of jumping from article a to editor e and depends on the initial conditions. We get that

$$\begin{cases} G_{e,a}(\beta) & = \frac{M_{e,a}k_e^{-\beta}}{\sum_{e'=1}^{n_e} M_{e',a}k_e'^{-\beta}} \\ G_{a,e}(\alpha) & = \frac{M_{e,a}k_a^{-\alpha}}{\sum_{a'=1}^{n_a} M_{e,a'}k_a'^{-\alpha}} \end{cases}. \tag{2.2}$$

Thanks to the transition probabilities, we get the recursive step:

$$\begin{cases} w_e^{n+1}(\alpha, \beta) & = \sum_{a=1}^{n_a} G_{e,a}(\beta) w_a^n(\alpha, \beta) \\ w_a^{n+1}(\alpha, \beta) & = \sum_{e=1}^{n_e} G_{e,a}(\alpha) w_e^n(\alpha, \beta) \end{cases}. \tag{2.3}$$

In Equation (2.2), Klein, Maillart, and Chuang [28] introduce two parameters, α and β, that inform how coordination generates value.

Parameters Analysis Let us have a look on how β, for instance, influences the transition probability in Equation (2.2):

- If $\beta = 0$: we recover the starting step.
- If $\beta > 0$: the probability $G_{e,a}(\beta)$ to jump from an article to an editor is a *power law function.* Hence, to a larger k_e corresponds a lower probability to jump.
- If $-1 < \beta < 0$: the function is concave.
- If $\beta < -1$: the function is convex, so more articles edited by an editor have a positive influence on the probability.

The same considerations hold for α in relation to $G_{a,e}(\alpha)$, the probability to jump from an editor to an article.

In order to choose the best parameters α^* and β^*, we need to calibrate them. Klein, Maillart, and Chuang [28] perform a *grid search* to maximise the Spearman rank-correlation between the rank given by the model and a ground-truth metrics obtained independently. Grid search simply consists in choosing the best parameter from a list of options that we provide, hence automating the 'trial-and-error' method. Spearman rank-correlation, denoted by ρ, is a non-parametric measure of the strength and direction of association between two variables through a monotonic function [45].

α^* and β^* must optimise both ρ_e and ρ_a:

$$\begin{cases} (\alpha^*, \beta^*) = \arg\max_{\alpha,\beta} \rho_e(\alpha, \beta) \\ (\alpha^*, \beta^*) = \arg\max_{\alpha,\beta} \rho_a(\alpha, \beta) \end{cases} . \tag{2.4}$$

Optimal parameters maximise the correlation because the higher is the correlation between the resulting model and the exogenous ground-truth, the better the collaboration structure is captured by the algorithm. The selection of the proper ground-truth for our case is widely discussed in Section 4.2.

Results Klein, Maillart, and Chuang [28] observe a *less-is-more* situation: "the number of articles ever touched by an editor better reflects the structure of collaboration and value creation, compared to edit counts, a much richer information input"[28]. They observe that if too many editors work on a article, this can create disvalue. Regarding the parameters, they work with different categories of Wikipedia articles and they find out that, while α remains quite constant, β significantly changes among categories. Therefore, the two, even if very related, may considerably change in terms of optimisation. We discuss if the same applies for our results in Chapter 5.

Chapter 3

Data

This chapter gives an overview of Crunchbase and it describes why it is a valid data-source for this research. We conclude by introducing also another platform, TMM, that will come into play once its new version will be released.

3.1 Crunchbase

To build the networks, we mainly rely on Crunchbase(CB) as data-source.[1] The CB platform, developed by a US-based company, is a source of information about start-up activities and their financing within and across countries. It started covering companies in the US, but in recent years it has expanded to the rest of the world. Data are sourced thanks to an investors network and a community of contributors. CB also aims to leverage big data and open-source information, but in a semi-automated fashion (there are employers who constantly work to clean and improve the dataset). Crunchbase information includes investments and funding information, mergers and acquisitions, news, and industry trends.

Usability of Crunchbase Data can be accessed in two ways: using an API or downloading a .csv file directly from the CB website. Data are divided in different databases, according to what they are related to. It is possible to find a complete description in Table 3.1.[2]

Another useful feature of the CB databases is that we can also download only a sample of them. In this way researchers can understand how data are structured and start doing some tests before downloading a large amount of data.

[1] Crunchbase website: https://www.crunchbase.com/

[2] Crunchbase daily CSV export from https://data.crunchbase.com/docs/daily-csv-export; data downloaded on April 28, 2021.

.CSV file name	*Description*
organizations	Organisation profiles available on CB platform.
organization_desc	Long descriptions for organisation profiles.
acquisitions	List of all acquisitions available on CB platform.
org_parents	Map between parent organisations and subsidiaries.
ipos	Detail for each IPO in the dataset.
people	People profiles available on CB platform.
people_desc	Long descriptions for people profiles.
degrees	Detail for people's education background.
jobs	List of all job and advisory roles.
investors	Active investors, both organisations and people.
investments	All investments made by investors.
investment_partners	Partners responsible for their firm's investments.
funds	Details for investors' investment funds.
funding_rounds	Details for each funding round in the dataset.
events	Event details.
event_appearances	Event participation details.

Table 3.1: Description of .CSV files of the CB platform.

3.1.1 Related work about Crunchbase

In their work, Dalle, Besten, and Menon [13] give a big picture about this platform and about why it stands out from other data sources. They mention many papers that state that Crunchbase is a valuable and reliable source, both for the quality of the data and the usability of the structure. Their analysis has been useful to confirm that CB is a suitable choice, also considering it includes many investors' and companies' properties that reflect what we need. Using the Crunchbase data, we may also expand our analysis for studying, for instance, the impact of incubators and accelerators on start-ups or a social analysis about how also the gender may be an influential factors [14]. Moreover, many works have benefited from the *linkability* of CB, which can be easily matched with other relevant data sources, such as PATSTAT and Seed, in order to extend the analysis [13].[3,4]

Another remarkable work, which relies on Crunchbase data, is the one conducted by Zhong et al. [46]. Their article highlights that the recent venture capital boom requires quantitative methodologies of screening and evaluation. They recognise that the birth of recent companies, for the majority tech companies, needs methodological decision making, which goes beyond personal choices. However, the well-known data driven decision making approach is particularly difficult for start-ups: a simple data analysis is often

[3] PATSTAT: https://www.epo.org/searching-for-patents/business/patstat.html.

[4] Seed-DB: https://www.seed-db.com/about/view?page=definition.

not sufficient, because usually there are not enough information for new technologies. This work does not leverage network theory, but proposes a two steps methodology which a investor should follow. First, the identification of the best start-ups to invest in, and then the identification of the best investment strategy. This structure shall apply also in our work, with the essential difference that we use network theory. In addition, this article develops a probabilistic latent factor.[5] Even if it is a good solution to estimate investment preferences in a collaborative way, it does not sufficiently focus on the technologies on which the start-ups work. Moreover, it is run only on US data. This made sense in 2016, when there were not many data available outside US, while now we need a global methodology, applicable everywhere. From the work by Dalle, Besten, and Menon [13], we also borrow the idea of including many properties of start-ups into the analysis, such as their proximity to a certain investor, the number of funding rounds and frequency of news on social media. Start-ups' properties contribute to the regression in their case, while they are included in a way suitable with graph theory in ours.

Also Liang and Yuan [31] use CB data to study the social impact feature when predicting investor funding using network theory. As expected, they show that investors are more inclined to invest in a company if they have a social relationship in terms of closeness. Interestingly, if investors and companies share too many common neighbours, investors are less likely to invest, because a form of competitiveness is established. Also in this case, the authors suggest many potentially influential factors, including the influence of Social Media. This study is possible thanks to the information contained in the CB platform, which includes also social features. This is useful also for including impactful factors –as we believe social aspects may be– to our algorithm.

3.2 TMM

The *Technology & Market Monitoring* (TMM) platform is an information system, developed by armasuisse S+T, that aims to exploit big data and open-source information in an automated way for the purpose of getting a better overview of the Swiss markets, products and trends of technologies.[6,7] The TMM system crawls and aggregates information from different online

[5] A latent variable model is a statistical model of observed data that incorporates latent variables (inferred by other variables, not directly observed). A probabilistic latent factor model works in a probabilistic framework. This may raise some problems: for instance, in probability a variable is observable or unobservable, but it is not always the case that non-observable variables are latent outside the probabilistic framework [17].

[6] Technology & Market Monitoring 1.0: https://tmm.dslab.ch/home

[7] armasuisse: Science and Technology S+T: https://www.ar.admin.ch/en/armasuisse-wissenschaft-und-technologie-w-t/home.html

resources as patent offices (*Patentsview*), commercial registers (*Zefix*) and websites (*Wikipedia* and *Indeed*).[8,9,10]

While the CB platform is more focused on delivering information about public and private companies on a global scale, TMM mainly covers the Swiss market [15]. After an evaluation of the usability of TMM, we have decided to focus on CB. First, it is more easily accessible and the data are much more structured (the TMM project is a work in progress, still under development and expansion). Moreover, CB extensively covers micro- and new-born enterprises, the main actors we are interested in. It also captures much additional information about the companies and the investors, that is not available yet on TMM.

However, TMM contains time series and soon it will include also the information that is currently available only on CB. It is in the interest of armasuisse and the CYD campus to use TMM as the main data source when the platform will be ready: time series will allow to extend this work and make predictions (see Section 6.3).

3.3 Investors analysis

As we have mentioned Crunchbase provides some information about investors. In order to have a clearer picture of the beneficiaries of our algorithm, we analyse the CB dataset dedicated to investors (called *investors* in Table 3.1). There are a total of 185,784 investors at the moment and they are divided into 78,001 (41.98%) organisations and 107,783 (58.02%) people.[11] Therefore, there are more people who are investors than organisations. However, interestingly, the same does not apply if we focus on Swiss investors only: the split is almost fifty-fifty (706 organisations and 732 people).

If we analyse the roles, Figure 3.1 shows that the majority of the investors are investors only (87.11%). However, some organisations are both companies and investors (12,65%). In this case, the partition is similar if we focus on Switzerland only: 79.90% are investors only and 16,27% are also companies. What remains are usually universities. Even if schools are very few compared to other investors (0.23% globally), their role is really important.[12] Moreover, in Switzerland they account for almost 4% of the investors, which is significantly larger than the global average. As a matter of fact, universities are not only centres of education and research, but community pillars, supporting the local economy and driving change [2]. Bautista

[8] Patentsview: https://patentsview.org.

[9] Zefix: https://www.zefix.ch.

[10] Indeed: https://indeed.com.

[11] Data downloaded on July 21, 2021.

[12] Universities as social impact investors: the snowball effect of value, Efficiency Exchange (December 9, 2020).

Puig, Mauleón, and Casado [2] claim that universities play a key role in both technology development and economic growth, impacting both local and global economy. Therefore, when we develop solutions for investors, we should keep in mind also the impact of universities, even if they account for a small portion of investors. Crunchbase has developed an algorithm for ranking investors and among the first 1000 globally, we find seven universities: Harvard University (344), Northwestern University (366), University of Michigan (414), Carnegie Mellon University (630), University of Cambridge (652), Duke University (819), and Yale University (1000).[13]

If we analyse where investors are located, from Figure 3.3 we can see that the majority of investors are located in the USA (29.62%). In particular, there is a wide gap between the first one and the second one, China, where 7.04% of investors are located. Figure 3.4 shows that in Switzerland 35.40% of investors are in Zurich, followed by Geneva, Zug, Basel and Lausanne.

[13] Crunchbase: Investors visited on July 21, 2021.

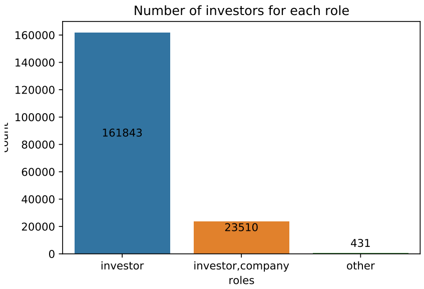


Figure 3.1: Investors' roles.

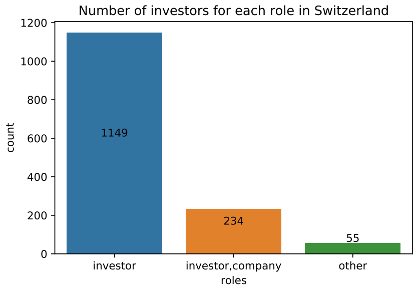


Figure 3.2: Investors' roles in CH.

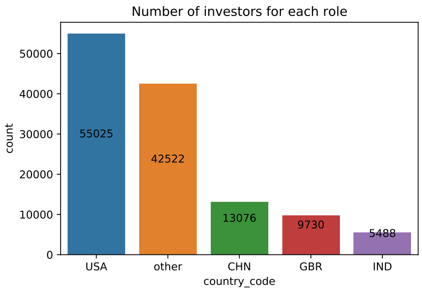


Figure 3.3: Investors' roles.

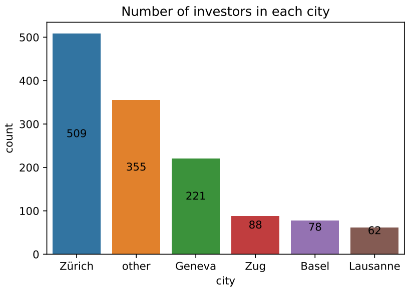


Figure 3.4: Investors' roles in CH.

Chapter 4

Methodology

To obtain a customizable rank for determining the influence of each entity, we develop an algorithm that consists of two main parts. The first, explained in Section 4.1, adapts the work by Klein, Maillart, and Chuang [28], thoroughly introduced in Section 2.1.3. The second part, Section 4.2, focuses on parameters optimisation and the inclusion of exogenous factors. We conclude this chapter with a financial analysis about how decision makers may use the results of the algorithm for investing in an informed way (see Section 4.3).

4.1 Adaptation of the work by Klein et al.

The main tool of this research is a bi-partite network that describes the relations among companies (C) and technologies (T): it summarises on which technologies each company is working. Figure 4.1 gives an idea of the structure of this bi-partite network. This structure enables us to benefit from techniques from different fields, such as network theory, Markov chains and machine learning.

As a matter of fact, the evaluation of startups and new technologies still largely depends on investors' personal choice and, as we mentioned in Chapter 1, this may lead to misreading market demand. This work aims to lead to more methodical decision making for investors.

We adapt the recursive algorithm developed by Klein, Maillart, and Chuang [28], based on the method proposed by Hidalgo and Hausmann [24].

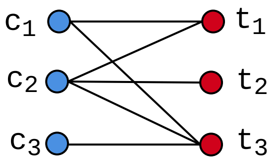


Figure 4.1: Example of bipartite network.

This methodology should encompass the complex structure of cooperation and competition of the technological landscape. Moreover, the resulting rank should condense the positive influence of well-established companies on technologies and, at the same time, the positive impact of newborn companies on unexplored fields. In the same way, a company is going to receive a higher rank thanks to the positive influence of important technologies. Klein, Maillart, and Chuang [28] claim that too many editors working on a Wikipedia article can create disvalue and we investigate if the same applies for cybersecurity technologies: if too many companies work on the same field and the business gap will narrow, will companies lose market share?

As Klein, Maillart, and Chuang [28], we build the adjacency matrix $M^{CT}_{c,t} \in \mathbb{R}^{n^c,n^t}$, which takes value 1 if a company c works on a technology t and 0 otherwise. n^c represents the total number of companies we are considering and n^t the total number of technologies. The aim of the algorithm is to assign to every node a weight that captures its relevance within the graph.

Under the assumption that a relevant entity has a relatively high number of neighbours, we initialise the algorithm with the degree of each entity

$$\begin{cases} w^0_c &= \sum_{t=1}^{n^t} M^{CT}_{c,t} = k_c \\ w^0_t &= \sum_{c=1}^{n^c} M^{CT}_{c,t} = k_t \end{cases}. \tag{4.1}$$

The idea underlying this assumption is that well-established companies have more means to diversify their expertise [11, 20]. Therefore, initialising the algorithm counting the neighbours of each entity is a good strategy that reflects the technological landscape. We investigate the limitations of this assumption in Section 6.3.

The algorithm is a *random walk* that, at each step, incorporates information about the expertise of companies and the relevance of technologies. The transition probabilities, $G_{c,t}$ and $G_{t,c}$, describe how much the entities weights change step-by-step. They can be both positive or negative depending on whether they represents a jump up or down in the score respectively: if the relationship between c and t brings value, the weight of the entities increases and the amount of the increase is driven by the transition probabilities. On the other hand, a negative transition probability suggests that this edge does not bring benefit to the entities in question. $G_{c,t}$ and $G_{t,c}$ are given by:

$$\begin{cases} G_{c,t}(\beta) &= \frac{M^{CT}_{c,t} k_c^{-\beta}}{\sum_{c'=1}^{n^c} M^{CT}_{c',t} k_c'^{-\beta}} \\ G_{t,c}(\alpha) &= \frac{M^{CT}_{c,t} k_t^{-\alpha}}{\sum_{t'=1}^{n^t} M^{CT}_{c,t'} k_t'^{-\alpha}} \end{cases}. \tag{4.2}$$

α and β inform how coordination generates value or disvalue. Then, we finally get the recursive step:

$$\begin{cases} w_c^{n+1} & = \sum_{t=1}^{n^t} G_{c,t}(\beta) w_t^n \\ w_t^{n+1} & = \sum_{c=1}^{n^c} G_{t,c}(\alpha) w_c^n \end{cases} . \tag{4.3}$$

Similarly to PageRank, the recursion ends when the rank stabilises and reaches convergence.

However, the TechRank algorithm is much more: the market ecosystem is extremely complex and a reliable influence factor needs to take into account other exogenous variables. A possible example is the investments' impact on companies, directly, and on technologies, indirectly via the investments on companies. We discuss how to include the influence of external factors and optimise the parameters α and β in the following section.

4.2 Inclusion of exogenous factors

We now conduct an analysis about what is the best way to take into account the influence of exogenous factors. First, we try to directly change the weights in Equation (4.3), but we believe that the simplicity of the random walker is one of the main strengths of the methodology and, thus, that this system should stay like this. Then, we investigate whether changing the transition probabilities, defined in Equation (4.2), is a good idea. It turns out it is not because the main role of these probabilities is capturing the coordination scheme between companies and technologies. As a matter of fact, by adding elements to Equation (4.2), we risk to change the goal of these transition probabilities, lowering their capabilities of fulfilling their main task. We suggest that the best strategy consists in including them as ground-truth in the parameters' calibration part. This choice allows to maintain the original simple form of the work by Klein, Maillart, and Chuang [28] and it does not risk to undermine the capability of the algorithm of capturing the coordination structure of the technological landscape. At the same time, it solves the issue of finding a proper external ground-truth during the parameters optimisation step.

As mentioned in Section 2.1.3, the parameter optimisation step requires a ground-truth metric, obtained independently. Once we have this ground-truth, we can find the Spearman correlation, ρ_c for companies and ρ_t for technologies, among this independent metric and the score obtained from the bi-partite random walker technique described in Section 4.1. Because ρ_c and ρ_t depend on α and β though Equation (4.2), the parameters are optimised by maximising these correlations:

$$\begin{cases} (\alpha^*, \beta^*) = \arg\max_{\alpha,\beta} \rho_c(\alpha, \beta) \\ (\alpha^*, \beta^*) = \arg\max_{\alpha,\beta} \rho_t(\alpha, \beta) \end{cases} . \tag{4.4}$$

Variable	$\in$	Description
$n^{(C)}$	$\mathbb{N}$	Number of external features available for companies.
n^{c}	$\mathbb{N}$	Number of companies.
$n^{(T)}$	$\mathbb{N}$	Number of external features available for technologies.
n^{t}	$\mathbb{N}$	Number of technologies.
$p_i^{(C)}$	$[0, 1]$	Percentage of interest in the company preference number i.
$p_j^{(T)}$	$[0, 1]$	Percentage of interest in the technology preference number j.
$f_i^{(C)}$	$\mathbb{R}^{n^c}$	Vector of factors associated to the company preference number i. One value for each company.
$f_j^{(T)}$	$\mathbb{R}^{n^t}$	Vector of factors associated to the technology preference number j. One value for each technology.

Table 4.1: Description of variables (part 1).

We solve this optimisation problem with a grid search as explained in Section 2.1.3.

Looking at the previous formula, it is clear that both parameters depend on both companies and technologies. Therefore, it is not the case that one parameter is paired with technologies and one with companies. This codependency is the key that allows to create the structure of cooperation and coordination within the bi-partite graph.

In order to calculate the correlation between the TechRank score –which assigns a weight w_c to each company and w_t to each technology– and a ground truth evaluation –which assigns $\hat{w_c}$ to each company and $\hat{w_t}$ to each technology–, they need to have the same form. Therefore, we normalise the TechRank results in the range $[0, 1]$, where 1 represents the most relevant entity, and we build the exogenous metric to be in the same range.

As mentioned above, the algorithm should be customisable in order to reflect the investors' preferences. Therefore, given a list of features about the entities, investors can choose in which ones they are interested and how much. For simplicity, let us work with companies for now. Table 4.1 contains more information about the notation. We suppose that the decision-makers have $n^{(C)}$ features to pick from, denoted as $f_1^{(C)}, ..., f_{n^{(C)}}^{(C)}$, where C represents the fact they are associated with the companies (T will denote the association to technologies instead). To each feature $f_i^{(C)}$ is associated a percentage of interest $p_i^{(C)}$. Of course, the percentages must sum up to 100%: $\sum_{i=0}^{n^{(C)}} p_i^{(C)} = 1$. For example, if the companies' available characteristics are the amount of previous investments and their geographical proximity to the investors, $n^{(C)} = 2$. Decision-makers may then decide to be 80% interested in the first feature and 20% in the second one by selecting $p_1^{(C)} = 0.8$ and $p_2^{(C)} = 0.2$.

It may also happen that, not only investors are not interested in a factor, but they believe it may negatively impact their decisions. In this case, the feature must be multiplied for -1. For instance, considering the last example, investors may think that companies with a big amount of previous investments are not a good choice and, thus, assign a negative impact to that factor. Therefore, they can keep the same percentages of interest in previous investments and geographical position, 80% and 20% respectively, but the sign of $f_1^{(C)}$ is negative.

The biggest challenge is converting quantitative and qualitative properties into a number $f_i^{(C)} \in [0, 1]$, related to the feature number i. For some features, the transition is quite straightforward, while for some others –especially for qualitative factors– a more complex approach is needed. Once we have created all the factors $f^{(C)} = f_1^{(C)}, ..., f_{n^{(C)}}^{(C)}$, the exogenous evaluation $\hat{w_c}$ is given by

$$\hat{w_c} = \sum_{i=1}^{n^{(C)}} p_i^{(C)} f_i^{(C)} = p^{(C)} \cdot f^{(C)}. \tag{4.5}$$

Considering $\sum_{i=0}^{n^{(C)}} p_i^{(C)} = 1$ and that $f_i^{(C)} \in [0, 1]$ for each company i, we have that $\hat{w_c} \in [0, 1]$. The same holds for $\hat{w_t}$.

In conclusion, the final formula is given by

$$\begin{cases} \hat{w_c} = p^{(C)} \cdot f^{(C)} \\ \hat{w_t} = p^{(T)} \cdot f^{(T)} \\ \sum_{i=0}^{n^{(C)}} p_i^{(C)} = 1 \\ \sum_{i=0}^{n^{(T)}} p_i^{(T)} = 1 \end{cases}, \tag{4.6}$$

where $n^{(T)}$ is the number of the technology-related features: $f^{(T)} = (f_1^{(T)}, ..., f_{n^{(T)}}^{(T)})$. As for companies, investors set their preferences choosing $p_1^{(T)}, ..., p_{n^{(T)}}^{(T)}$.

The properties to pick from depend on the data available. In our case, CB provides quite a wide range of information about companies and investors (see Table 3.1). We choose some of the attributes and transform them into potential relevant features for decision-makers. As discussed, each requires a different strategy and we describe some of them in the following paragraphs. These methodologies may be affected if data changes, for instance if we move to TMM data.

4.2.1 Previous Investments

One of the most important factors for evaluating companies depends on the amount of previous investments. Investors may rather invest in companies that have been already supported in the past or, vice versa, they may rather prioritise less acknowledged businesses.

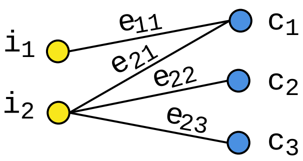


Figure 4.2: Example of a I-C bi-partite network.

For the purpose of calculating this factor, we use a CB dataset (called *funding_ rounds* in Table 3.1) that keeps track of the amount of all the funding rounds from each investor i to each company c. This structure can be captured with another bi-partite network that describes the links among investors (I) and companies (C). In contrast to the C-T graph, in this case edges are weighted: the weight is given by the sum of all the previous (until the current day denoted as $\mathcal{T}$) investments from a certain investor i to a certain company c. As for the other bipartite graph, we can calculate the adjacency matrix M^{IC}. Supposing the amount γ of a single investment from i to c at time t is represented by $\gamma_t^{i,c}$, the weight of the edge i-c is given by $e_{i,c} = \sum_{t=0}^{\mathcal{T}} \gamma_t^{i,c}$. Table 4.2 presents more information about the notation.

In order to find an attribute $f_c^C \in [0,1]$ for each company c, first we need to sum the contribution of all its investors.[1] Then, we find the maximum among all the previous investments received by each company and we divide all the investments by this maximum.

We present an example to clarify: let us suppose there are two investors $-i_1$, i_2- and three companies $-c_1$, c_2, c_3- and that the structure which synthesises the history of their investments is given by Figure 4.2. We compute the maximum $e_{\max}$ as $\max\{e_{11}+e_{21}, e_{22}, e_{23}\}$ and the features related to the investments for each company as:

$$\begin{cases} f_1^C &= \frac{e_{11}+e_{21}}{e_{\max}} \\ f_2^C &= \frac{e_{22}}{e_{\max}} \\ f_3^C &= \frac{e_{23}}{e_{\max}} \end{cases} . \tag{4.7}$$

If n^i is the total number of investors and n^c the total number of companies, the methodology described above can be generalised to the following

[1] The notation f_c^C can be confusing if compared to $f^{(C)}$ in Equation (4.5). Here, considering we are working on an exogenous factor only –previous investments–, f_c^C represents the factor related to a specific company c ($f_d^{(C)}$ is related to another company d for instance). On the contrary, in Equation (4.5), the subscript identifies a specific factor (See Table 4.1).

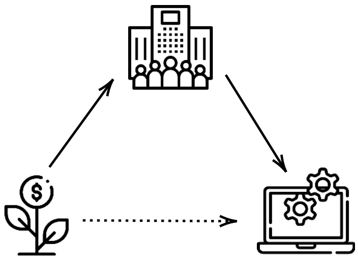


Figure 4.3: Tripartite structure of investors, companies and technologies.

formula:

$$\begin{cases} e^{IC}_{i,c} = \sum_{t=0}^{\mathcal{T}} \gamma_t^{i,c} & \forall i, c \\ e^{C}_{c} = \sum_{i=1}^{n^i} e_{i,c} M^{IC}_{i,c} & \forall c \\ e_{\max} = \max_c e^C_c \\ f^{(C)}_c = e^C_c / e_{\max} \end{cases} \tag{4.8}$$

for each $c \in 1, ..., n^c$. This methodology is also described in Algorithm 1. Thanks to Equation (4.7), for each company we have a factor between 0 and 1 that summarises the amount of previous investments.

Algorithm 1 Previous investments factor (companies)

```
$e^C \leftarrow [0] \cdot len(c\_names)$
for $c \in range(c\_names)$ do
    for $i \in range(i\_names)$ do
        for $c \in range(i\_names)$ do
            $e^{IC}_{i,c} \leftarrow \sum_{t=0}^{\mathcal{T}} \gamma_t^{i,c}$   ▷ $\gamma^t_{i,c}$ is the amount of the investment from $i$ to $c$ at time $t$
            $e^C[c] \leftarrow e^C[c] + e^{IC}_{i,c}$
        end for
    end for
end for
$e^C_{max} \leftarrow \max(e^C)$
$f^C \leftarrow e^C / e_{max}$   ▷ $f^C$: list of previous investments for each technology
return $f^C$
```

Tripartite structure Analysing the relations among investors and companies through a bi-partite graph, we can create a link with the C-T bi-partite network obtaining a tripartite structure I-C-T, showed in Figure 4.4. Thanks to this configuration, we can assign some features to technologies starting from the companies –direct link– or investors –indirect link. For instance, we can find the amount of previous investments on a technology by looking

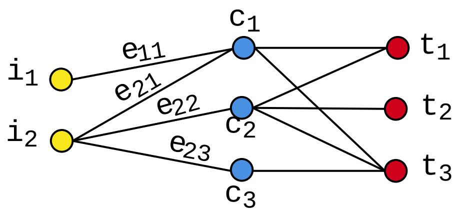


Figure 4.4: Example of tripartite network.

Variable	$\in$	Description
n^i	$\mathbb{N}$	Number of investors.
M^{CT}	$\mathbb{R}^{n^c \cdot n^t}$	Adjacency matrix of the C-T bipartite network.
M^{IC}	$\mathbb{R}^{n^i \cdot n^c}$	Adjacency matrix of the I-C bipartite network.
$\gamma_t^{i,c}$	$\mathbb{R}$	Amount of the funding round between c and i at time t.
e^{IC}	$\mathbb{R}^{n^i \cdot n^c}$	Total amount of investment between each investor to each company.
e^C	$\mathbb{R}^{n^c}$	Total amount of investments toward each company.
e^T	$\mathbb{R}^{n^t}$	Total amount of investments toward each technology.
$e^C_{\max}$	$\mathbb{R}$	Maximum amount of total investments among all the companies.
$e^T_{\max}$	$\mathbb{R}$	Maximum amount of total investments among all the technologies.
f^C_c	$[0, 1]$	Factor related to previous investments into the company number c.
f^T_t	$[0, 1]$	Factor related to previous investments into the technology number t.

Table 4.2: Description of variables (part 2).

at the funding rounds of the companies which are working on it. There is an evident limitation with this approach: we cannot know, using the data currently available, how much a company is spending for the development a specific technology given all the funding received by the company. We better describe this weak point in Section 6.3. For now, this idea of evaluating the previous investments in a technology is a fictitious example for describing how information can flow from investors to technologies. Supposing we keep the notation of the previous sections, we can find the previous investments' factor for technology using

$$\begin{cases} e^{(I,C)}_{i,c} = \sum_{t=0}^{(T)} \gamma^t_{i,c} & \forall i, c \\ e^C_c = \sum_{i=1}^{n^i} e_{i,c} & \forall c \\ e^T_t = \sum_{c=1}^{n^c} e_c M^{CT}_{c,t} \\ e_{\max} = \max_t e^T_t \\ f^{(T)}_t = e^T_t / e_{\max} \end{cases} . \tag{4.9}$$

Table 4.2 contains a detailed description of the variables involved. We can also work with matrix multiplications:

$$\begin{cases} e^{IC}_{i,c} = \sum_{t=0}^{(T)} \gamma_t^{i,c} & \forall i, c \\ e^{C}_{c} = \sum_{i=1}^{n^i} e^{IC}_{i,c} & \forall c \\ e^{T} = e^{C} \cdot M^{CT} & \\ e^{T}_{\max} = \max_{t} \left(e^{T}\right) & \\ f^{(T)} = e^{T} / e_{\max} & \end{cases} . \tag{4.10}$$

The algorithm of this methodology is proposed in Algorithm 2.

Algorithm 2 Previous investments factor (technologies)

```
$e^C \leftarrow [0] \cdot len(c\_names)$
for $c \in range(c\_names)$ do
    for $i \in range(i\_names)$ do
        $e^{IC}_{i,c} \leftarrow \sum_{t=0}^{(T)} \gamma_t^{IC}$   ▷ $\gamma^t_{i,c}$ is the amount of the investment from $i$ to $c$ at time $t$
        $e^C[c] \leftarrow e^C[c] + e^{IC}_{i,c}$
    end for
end for
$e^T \leftarrow e^C \cdot M^{CT}$   ▷ Matrix multiplication
$e_{max} \leftarrow \max(e^T)$
$f^T \leftarrow e^T / e_{max}$   ▷ $f^T$: list of previous investments for each technology
return $f^T$
```

4.2.2 Geographical position

Decision-makers may want to base their decisions also on the geographical position of the companies. In this case, we need a factor that accounts for the distance between the favoured area and the companies. Crunchbase provides us with the address of enterprises and investors. From the address, we can find the geographic coordinates and we can use the *Haversine* function to measure the distance.

For instance, investors may prefer short-distance investments or they may select some places with high potential –such as hubs near prestigious universities or areas rich of resources. In case they have a strong opinion and they are not open to take into consideration other regions apart from the ones they have selected, we can directly filter the companies' list before running the algorithm. If this holds and if they have no further preferences inside the area they have selected, then no external factor related to the geographical position is needed. On the other hand, if decision-makers are open to invest in different areas, even if they have some preferences about the location of the companies, we do not need to filter the data at the beginning: we can add an external factor that favours companies situated in the area chosen by the investors and disfavour the furthermost companies. In this way, the latter can still be taken into consideration, but they need outstanding results to overpass the score of well-situated companies. The idea is not to lose good opportunities due to the geographical position keeping the focus on the selected area.

As a matter of fact, there are many reasons that may lead decision-makers to prefer a specific area. We can think, for example, of *tech hubs*: areas where startups grow quickly and there are good investments opportunities. The San Francisco Bay Area has been the world's leading technological innovation hub since the '70s and investments' titans are likely to continue to be attracted by this innovation centre, a concentration of wealth and bright minds.[2] However, some investors have been deciding to search for opportunities somewhere else. For instance, Singapore is a good example of an important *tech city* outside the US: it ensures high quality thanks to its reputable universities, it is a safe city with good infrastructures and it has been capable to develop young entrepreneurs.[3]

Some companies have even decided to open HQs in less prolific business environments. For example, Twitter has recently done a big step towards globalisation, announcing that they will establish an HQ in Ghana.[4] It is likely that this decision has been driven not only by humanitarian reasons, but also by practical motivations that played a central role: Africa is a free trade area and Ghana supports online freedom and the open internet. Moreover, Ghana's Human Development Index has increased by 31.4% between 1990 and 2019.[5,6]

In general, it may be assumed that investments' decisions are driven by humanitarian reasons. Big investments may have a significant impact on poorer regions and influence the income inequality among countries, but this is not always the case: some work [30, 9] shows that the amount of foreign direct investments is unrelated to the distribution of income. Even if one of the main drivers of globalisation are multinational enterprises (MNEs), a relatively small number of MNEs accounts for most of the world's trade and foreign investment [30, 37]. Moreover, it has been shown that both in developing and developed countries, income inequality is unaffected by the presence of multinational corporations [30, 9]. These analyses suggest that investors' decisions about the location are not likely to be driven by humanitarian reasons only but, as in the Twitter case, supported by concrete

[2] The New Yorker: How "Silicon Valley" Nails Silicon Valley by Andrew Marantz (June 9, 2016).

[3] Financial times: Welcome to Silicon Allee: the new global tech hubs by Lisa Freedman (October 12, 2020).

[4] BBC: Ghana basks in Twitter's surprise choice as Africa HQ by Ijeoma Ndukwe (April 25, 2021).

[5] The Next Frontier: Human Development and the Anthropocene by the United Nations Development Programme (2020).

[6] Unesco: Ghana - Communication Indicator - Freedom of Expression.

and tangible motivations. Now, we focus on how to create an exogenous factor regarding the companies geographical position. We start by describing the Haversine distance.

Haversine distance The angular distance between two points on the Earth's surface can be measured thanks to the Haversine distance ($\text{hav}(\theta)$), using the latitude and the longitude of the geographical points [26]. Let (λ_1, ϕ_1) and (λ_2, ϕ_2) be the longitude and the latitude in radiance of two points on a sphere, the *central angle* θ among them is given by the spherical law of cosines:

$$\theta = \arccos\left(\sin\phi_1 \sin\phi_2 + \cos\phi_1 \cos\phi_2 \cos\left(\lambda_2 - \lambda_1\right)\right). \tag{4.11}$$

From this, we can easily find the *great-circle distance*:

$$d = r \cdot \theta, \tag{4.12}$$

where r is the sphere radius [27]. The Haversine function of a general angle α is

$$\text{hav}(\alpha) = \sin^2 \frac{\alpha}{2} = \frac{1 - \cos\alpha}{2}. \tag{4.13}$$

From this, we get the Haversine formula of θ, directly computed from (λ_1, ϕ_1) and (λ_2, ϕ_2):

$$h = \text{hav}(\theta) = \text{hav}(\phi_2 - \phi_1) + \cos\phi_1 \cos\phi_2 \, \text{hav}(\lambda_2 - \lambda_1). \tag{4.14}$$

Now that we have defined how to calculate the Haversine distance h, we can describe how to use it to obtain a factor $f_c^{(C)} \in [0, 1]$ for each company. Please note that we do not discuss how to shift this property to technologies here: there is no purpose in discussing the position of a technology.

We are interested in the Haversine distance $h_{i,c}$ between the company c and the investor i. If we assume that the factor is the closeness between the two, we get that $f_c^{(C)}$ tends to 1 when the distance is low:

$$f_c^{(C)} \to 1 \quad \text{when} \quad h_{i,c} \to 0. \tag{4.15}$$

To compute $f_c^{(C)}$, we first need to find $h_{i,c}$ for each company and identify the maximum distance $h_{\max}$ among all the companies. Also for the previous investments' factor we find the maximum. The reason is that we are comparing different entities, we are not assigning a global score: we are searching the best ones among the nodes available. Calculating the maximum mathematically allows to obtain a factor $f_c^{(C)} \in [0, 1]$ that is on the same order of magnitude of the other factors. Without it, different factors may have very different scales, also outside $[0, 1]$. Consequently, when we weight them according to investors' preferences, we would obtain a $\hat{w}_c$ composed by factors of different magnitudes and potentially situated outside the interval $[0, 1]$.

Algorithm 3 Geographic coordinates factor

```
$h\_dict \leftarrow \{\}$
for $c\_name, c\_address \in c\_locations$ do
    $lat \leftarrow c\_address.latitude$
    $lon \leftarrow c\_address.longitude$
    $h \leftarrow haver\_dist(lat, lon, lat\_inv, lon\_in)$    ▷ haver_dist is a function we have created
    $h\_dict[c\_name] \leftarrow 1/h$
end for
$h\_max \leftarrow \max(h\_dict)$
for $c\_name, h \in h\_dict$ do
    $h\_dict[c\_name] \leftarrow 1 - h/h\_max$
end for
return $h\_dict$
```

Once we have $h_{\max}$, the factor for each company c is given by $f_c^{(C)} = 1 - h_{i,c}/h_{\max}$: the formula is given by $x = h_{i,c}/h_{\max}$ in order to have a number between 0 and one 1, then, $1 - x$ because the lower x is the better it is and we have set that to the best companies (zero distance) is assigned $f_c^{(C)} = 1$.

Algorithm 3 synthesises the main steps of the code. Please note that we suppose that the address of each company, $c\_address$, is already in a suitable for finding the geographical coordinates of the investors (lat_inv, lon_in). In case decision-makers prefer long-distance investments, the algorithm does not change but we set $f_c^{(C)} = h_{i,c}/h_{\max}$.

4.2.3 Other potential factors

Apart from the aforementioned exogenous factors, many more might be included in TechRank, keeping in mind that the rule according to which each feature $f_i^{(C)}$ must be in $[0, 1]$ should always apply. In particular, there are some quantitative and qualitative properties about the companies that may play a key role for some investors. In order to transform them to a quantitative number, we must find the proper strategy case by case, as we have done in the previous sections for the previous investments (quantitative) and the geographical position (qualitative).

Among the potentially relevant properties about companies, investors may be interested in when the company has been funded, its number of employee, the number of women in the board, its social networks activity, if the company has gone public etc. For the sake of brevity, we do not explore them in this work, but they can be a good starting point for further research (see Section 6.3).

4.3 Financial Analysis

At this stage, we have obtained the TechRank scores for companies and technologies. However, many open questions for creating the best investment strategy still need an answer. For instance, investors should choose whether to look at the technologies' scores and investing in companies working on them; or they prefer to focus on enterprises directly. Another open question regards how to choose the number of entities in which to invest. Before that, let us devote Section 4.3.1 to describe the potential users of TechRank, the investors; with a particular focus on venture capital investors. As a matter of fact, studying the recipients of our methodology is an important step for understanding what would be the best strategy for them.

4.3.1 Investors

When a company develops a new technology and needs to fund its projects, it can look for different kinds of investors, which we now describe.

Banks and personal investors

Capital can be raised by asking business loans to banks. A loan has a strict repayment schedule to adhere to and its interest rates may rise if the loan includes floating rates. On the other side, companies do not have to give the lender a percentage of the profits or shares. One of the reasons why banks may not be a suitable choice for the development of new technologies is that loans often need a collateral to justify lending. Moreover, loans are not very flexible: they are often bounded by strict regulations to avoid high-risk debt and they have to respect strict conditions. these may be obstacles for the development of *risky* projects including new technologies.

While business loans must be repaid regardless of if the project is successful or not, personal investors should understand and accept the risk that if the project is inconclusive and are aware that they can lose their money. Nevertheless, investors expect a share of the profits in exchange for their capital and sharing equity may become a problem on the long run.

Venture Capitalists

Venture capitalists, or venture capital investors (VC) have four principal properties:

- They have a *financial intermediary* which takes investors' capital and invest it into the companies (See Figure 4.5).
- They only invest in *private companies* in order to fund the internal growth of companies.

- They have a role in helping companies to manage their portfolio and skills.
- Their primary goal is to maximise their returns by finding promising investments opportunities through sales or *initial public offerings* (IPO).

In particular, venture capitalists seek investments with a potential high-growth rate, as well as high-risk and high-return. IPO refers to the process in which shares of a company are sold to the general public (institutional investors). This allows the company to raise funds from public investors, which is considered a true sign of success: the company has grown to a point in which its revenue volume and profitability are large enough to allow public ownership. After the IPO, companies go public: a public company is a company whose ownership is organised through shares which can be traded on a stock exchange or in over-the-counter markets [7].

Moreover, VCs often are highly specialised, so they can participate actively: for instance they get board seats and voting rights [33].

Figure 4.5 shows how VCs work in relationship with investors and companies. VCs raise capital mainly from institutional fund investors, professional asset managers, insurance companies, pension funds or fund-of-funds (FOF)[7]. Beyond that, there are also wealthy families or endowment funds who can be interested in VCs as they promise potentially higher investment profits compared to regular asset classes [4].

Some VCs establish *business incubators* as an investment opportunity. Business incubators play a key role for startups: they accelerate the growth of companies and help in maintaining the business focus. For startups, usually the first focus is surviving, due to the many challenges they have to face at the beginning. Incubators are essential also because they provide funding opportunities, through investor relations and partners. They may invest in startups in exchange for equity or offer funding further during the program. Moreover, they help startups in creating a good network with industry leaders and mentors.

[7] A FOF is an investment strategy that consists of holding a portfolio of other private-equity funds rather than investing directly [33].

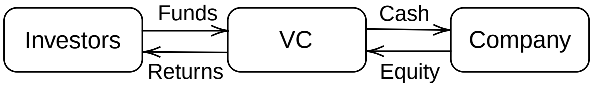


Figure 4.5: VC's role as financial intermediary.

Angel Investors

Angel investors are individuals who provide capital for start-ups, usually in exchange for ownership equity or convertible bonds. VC can be confused with angel investors, but the difference is given by the fact that angels own the capital and, thus, there is not any intermediary. Angels directly invest in the company. As VCs, angel investors provides strategic and operational expertise [43].

Peer-to-Peer lenders

Peer-to-peer (P2P) lending allows to obtain loans directly from borrowers without an official financial institution as intermediary. P2P lending is usually done using online platforms that match lenders with the potential borrowers and, among the benefits, we find lower fees and interest rates, higher returns for the investors, and more accessible funds. However, credit risks are generally high and there is no protection or insurance to the lenders in case of the borrower's default.[8]

4.3.2 Investment strategy

Now that we have a better idea of the different profiles that may be interested in TechRank, we investigate how investors can find the investment strategy that better reflects their needs and preferences. As aforementioned, the first choice to be done is whether focusing on companies or technologies. In the latter case, investors have to check which companies are working on the best technologies –those with the higher TechRank score– and this leads us back to the first case. The step we have just outlined implies many questions, such as: on how many technologies do they want to invest? If one, do they prefer to split their capital among one or more companies working on that technology? If they want to invest in many technologies, how they want to split their capital among them? Equally? In addition, is it better to invest in more companies working on a technology or to choose the best company for each technology?

This scenario is complex and cannot be solved by a predefined procedure. It is important to remember that, even if they focus on technologies, funds go to companies, not to technologies directly. The flowchart in Figure 4.6 presents a general guideline of the decision process that investors may follow. Of course, it should be adapted on a case-by-case basis. In particular, the step "Decide how much invest in each company" is complex and hides many further questions, which depend on the many different factors and preferences.

[8] Corporate Financial Institute: Peer-to-Peer Lending

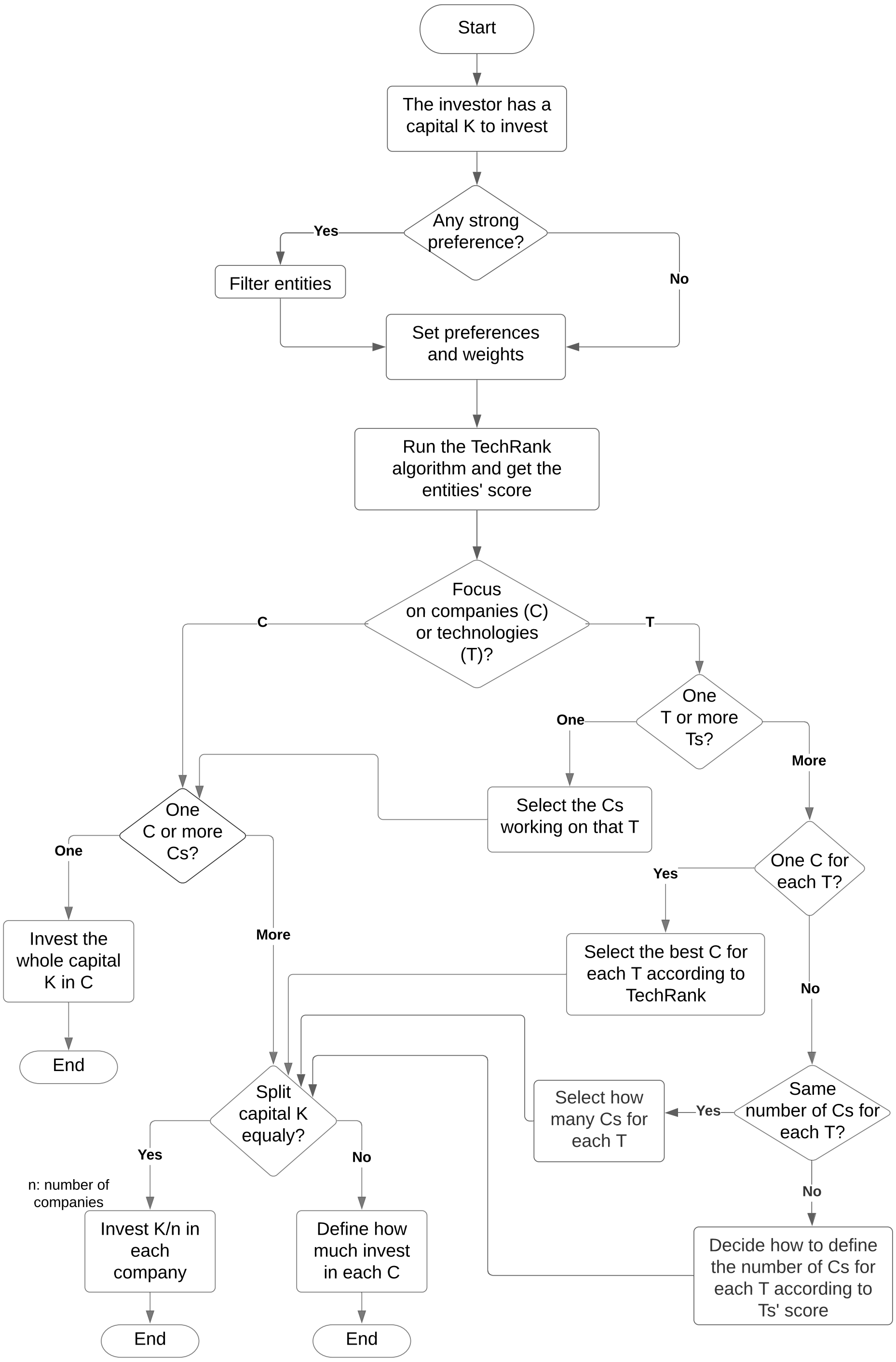

Start
The investor has a capital K to invest
Any strong preference?
Yes
No
Filter entities
Set preferences and weights
Run the TechRank algorithm and get the entities' score
Focus on companies (C) or technologies (T)?
C
T
One T or more Ts?
One
More
Select the Cs working on that T
One C or more Cs?
One
More
One C for each T?
Yes
No
Invest the whole capital K in C
Select the best C for each T according to TechRank
End
Split capital K equaly?
Select how many Cs for each T
Yes
Same number of Cs for each T?
Yes
No
No
n: number of companies
Invest K/n in each company
Define how much invest in each C
Decide how to define the number of Cs for each T according to Ts' score
End
End

Figure 4.6: Flowchart that describes a guideline for the investment process.

Chapter 5

Results

In this chapter, we show and explain our results. We start by the outcomes of the TechRank algorithm in the cybersecurity field, followed by a comparison with the Crunchbase rank. We also study the run-time of TechRank. Then, we focus on the the role of exogenous factors. We also explore how results change when running TechRank applied to another field, the medical one. To conclude, we also explore some characteristics of the targets of this methodology, the investors.

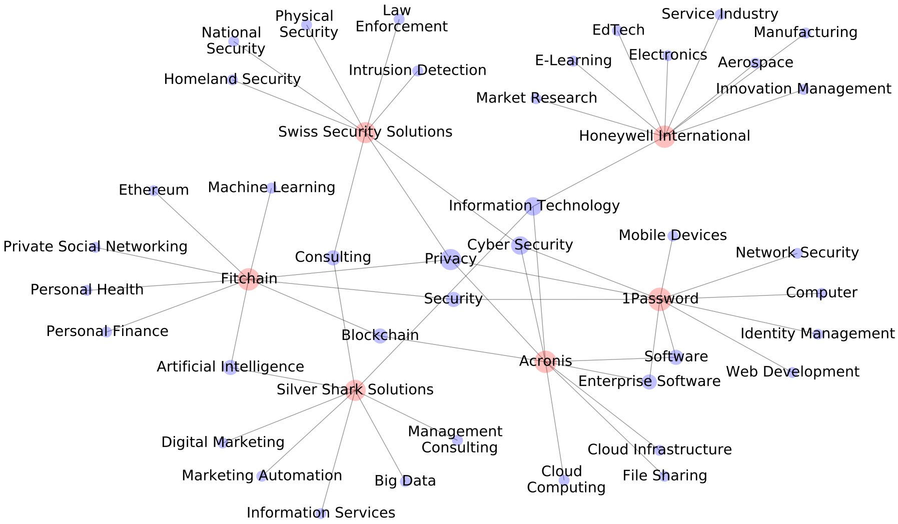


Figure 5.1: Bi-partite network for selected cybersecurity companies (red nodes) and technologies (blue nodes) they are working on. The nodes' size represents the number of neighbours.

5.1 Cybersecurity field

From the CB platform, we select all the companies whose description contains at least two words from a list of words related to the cybersecurity field and we get 2429 companies and 477 technologies.[1] Figure 5.1 gives an idea of the structure of the bi-partite network that describes on which technologies each company is working on.

The first results regard the parameter calibration step. For now, we suppose that investors are only interested in previous investments, both for technologies and companies. We examine how results change with a change of the preferences using a smaller sample of companies in Section 5.1.3. Figure 5.2 shows the optimisation landscape in which the correlations ρ_c and ρ_t change according to α and β. We identify α^* and β^* (0.04 and -1.88 for companies and 0.48 and -2.00 for technologies respectively) and we use them in the recursive algorithm.

The evolution of the TechRank random walker, explained in Section 4.1, is illustrated in Figure A.6. The results of the evolution show that while the positions of the entities change significantly during the first steps, then they gradually tend to stabilise. Considering 2429 companies and 477 technologies, the algorithm stabilises after 723 iterations for companies and after 1120 for technologies. Interestingly, entities which have a high score at the beginning (recall that we initialise the score of each node with its degree) usually do not change their position significantly: they tend to remain among the best ones. Therefore, the algorithm assigns the best scores to entities with many neighbours. However, the opposite does not apply for entities with a low degree: they can significantly change the score, especially the technologies. Therefore, even if TechRank is able to recognise the importance of the most established entities (companies that work on many technologies and technologies used by many companies), it also enhances new technologies. Figure A.2 shows the first classified entities in cybersecurity.

We also study how the TechRank results change with a change in the number of companies and technologies. To change the number of entities, we set only the number of companies n^c; the number of technologies n^t is instead a consequence of n^c: n^t comes automatically by investigating in which technologies each company is working on. For instance, in the cybersecurity field, by selecting 10 companies randomly, we get 26 technologies. Considering that the total number of companies on CB related to the cybersecurity field are 2429, we study the run-time running the algorithm for 10, 100, 499, 997, 1494, 1990, and 2429 companies and 26, 134, 306, 372, 431, 456, and 477 technologies respectively.[2]

[1] The list of the words can be found in Appendix A2.

[2] The reason why we have 499 companies and not 500 is because we delete the companies which do not include a list of their technologies. So, even if we select a certain number of companies, the final n^c may be a bit lower.

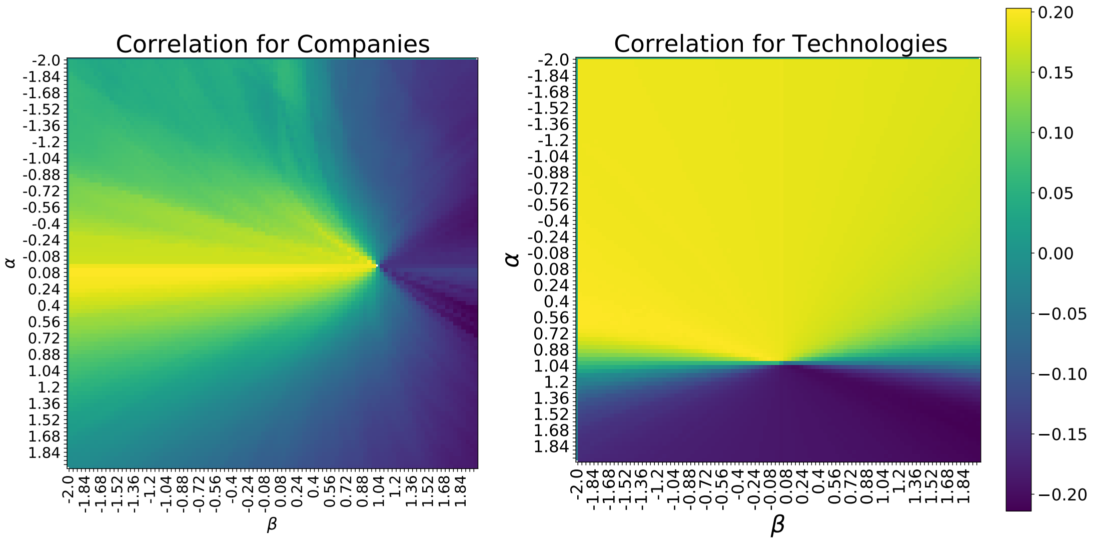


Figure 5.2: Grid search for 2429 companies and 477 technologies. 100% preferences in previous investments.

In order to better understand how the algorithm works, we conduct a detailed analysis considering 10 companies only.[3] Figure 5.4 shows that AppOmni does not change its position, but two of its technologies (SaaS and cloud management) increase their scores.[4] If we consider only this environment composed of 10 companies (see Figure 5.5), SaaS and cloud management do not have other links. The strength of this company lies on its capability of combining important technologies (software, cyber security and cloud security) with less explored fields (SaaS and cloud management). Also the Integrity Market Group company works alone on some fields: marketing, digital marketing and advertising. However, this company does not employ other more established technologies and thus we do not register any improvement in its score. As a matter of fact, it is completely separated from the rest of the network (its technologies are not linked to other companies). If we look at Lacework and Acronis instead, we see that the two companies follow an opposite trend: Lacework significantly increases its score, while Acronis drastically drops from the first position (due to its high number of links) to the last one. We suggest that the reason of this decline is given by the fact that Acronis works on too many technologies, most of which are not explored by other companies. Therefore, even if it works also on some very well established technologies, this is not enough for compensating the

[3] Please note that we are aware that the number of entities is too low to consider the results reliable for developing an investment strategy. However, the goal here is to explain how the cooperation and competition structure of TechRank works.

[4] SaaS: Software as a Service.

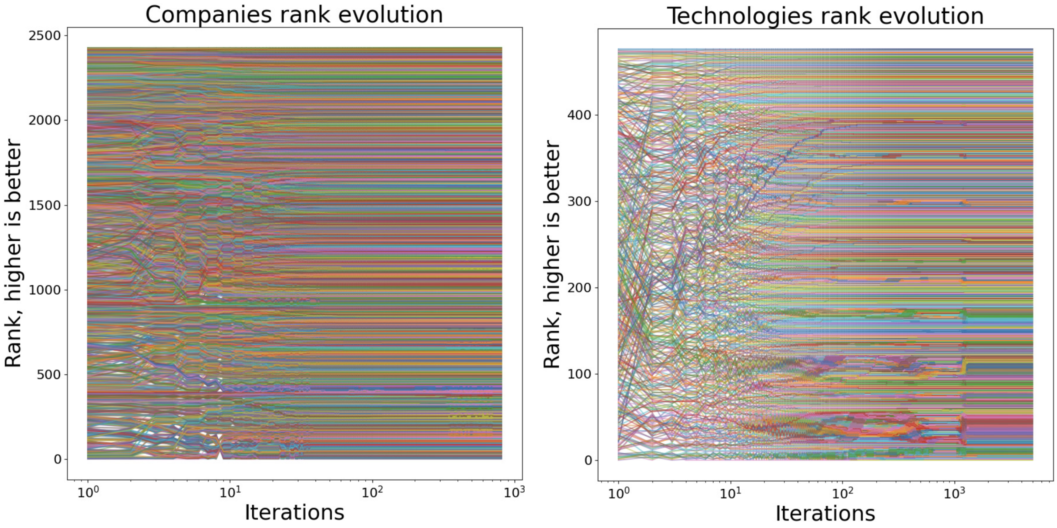


Figure 5.3: TechRank scores evolution for 2429 companies and 477 technologies in cybersecurity.

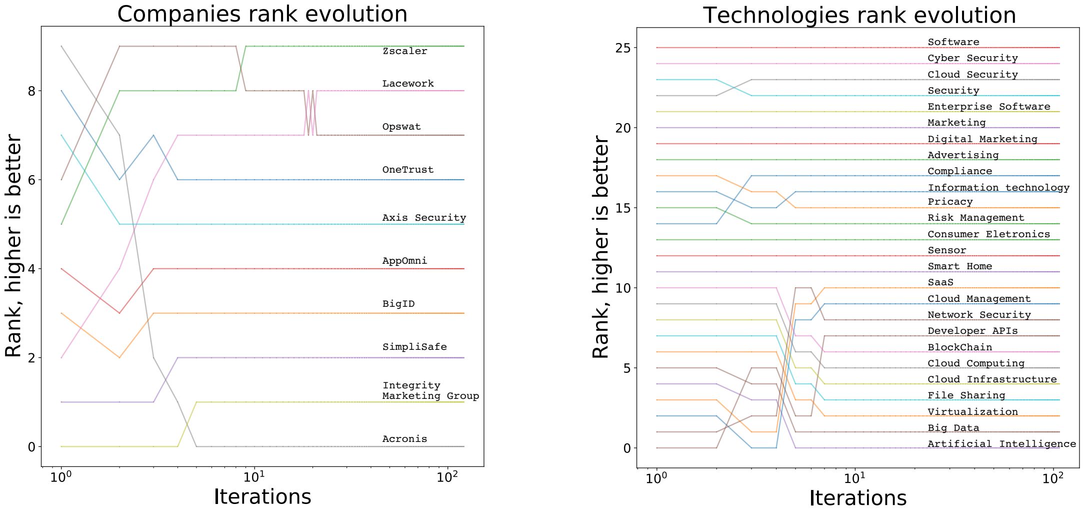


Figure 5.4: Rank Evolution considering 10 companies and 26 technologies in cybersecurity.

uncertainty of the fields in which it has invested alone. On the other hand, Lacework has a good balance: it relies on some acknowledged technologies (security, cloud security and software) and expands its horizon working also on compliance (only two companies work on it). We can see that also the compliance technology benefits from this balance, increasing its rank by 3

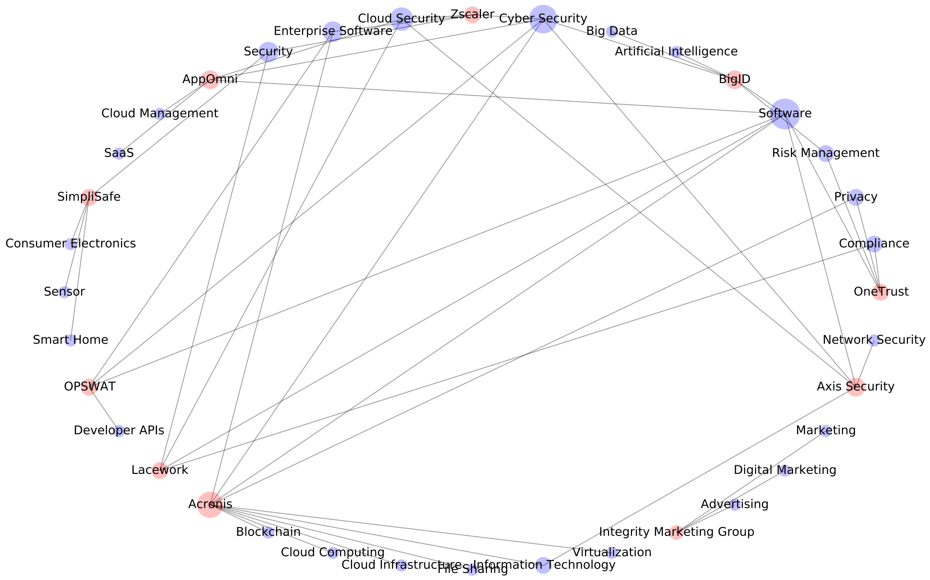


Figure 5.5: Circular network representation considering a landscape of 10 companies and 26 technologies in cybersecurity.

positions.

From our analysis involving different numbers of entities, it arises that the number of iterations needed to achieve convergence does not depend on the number of entities: Table A.2 shows that increasing the number of entities, sometimes the number of iteration decreases and sometimes increases. In general, we observe that technologies need more iterations. Considering that there are many more companies than technologies and that each company has at least one edge, the technologies nodes have, on average, a higher degree than the one companies have; thus, we expect the structure and the dynamics related to technologies to be more complex. From these results, we conclude that the complexity of the algorithm does not depend on the number of entities only, but the structure of the network plays a key role as well.

5.1.1 Comparison with the Crunchbase rank

The CB platform assigns a rank to the top companies – according to their algorithm – in each industry. The CB rank takes into account the entity's strength of relationships, funding events, news articles, acquisitions, etc..[5] We compare our results in cybersecurity with the CB rank, and we investigate the strength and direction of the association between the two scores using

[5] https://about.crunchbase.com/blog/influential-companies/

the *Spearman's correlation* coefficient.[6]

To make the ranks comparable, we convert our algorithm's output into a ranking. The resulting Spearman's correlation (0.014) shows that the two ranks are not correlated: even if the goals of TechRank and the CB rank are similar, their substantial differences are reflected in this outcome. First, the CB rank is fixed, while TechRank is customisable according to the needs of the investors, which naturally leads to different results that depend on the preferences of the investors. Moreover, the CB rank focuses more on the level of activity of the company, rather than its influence on the market. Among all the factors that influence it, nothing can be found about if and how technologies contribute to the results; which, on the contrary, is essential in our case. Furthermore, the CB rank results from an algorithm that involves all the companies, while we focus only on some of them. Therefore, the number of companies involved influences the inner dynamics of the ranking algorithms and this leads to different results. We try to change the investors' preferences and we see that the correlation always stays below 0.02.

Some of the differences we have just presented are also benefits of TechRank: it is a customisable methodology and it focuses on the market influence of companies, which is a more relevant feature for investors than their level of social activity. Another big advantage of the TechRank algorithm is its transparency, while, on the other hand, we do not know the exact mechanism by which CB (which is not open source) ranks the entities, and thus potential investors cannot understand how much the CB rank reflects their preferences. Moreover, CB simply ranks companies, while our algorithm assigns a weight, which allows to identify not only the order, but a quantitative idea of the distance between one entity and the next one. Another good feature of the TechRank algorithm is the opportunity for decision-makers to set a threshold before running the algorithm. For instance, as already mentioned, to include only companies located in certain areas. Finally, another benefit is that we assign a score also to technologies: some investors may be interested in a certain field and not in companies only. Also for this reason, the TechRank methodology enables them to create the portfolio that better reflects their preferences.

5.1.2 Run-time

All the code related to the TechRank algorithm is explained in Appendix A1. We run it on a machine with a 16-cores Intel Xeon CPU E5-2620 v4 @ 2.10GHz and with 126GB of memory.

In order to study the run-time of TechRank, we need to investigate how the time spent for running the algorithm changes with a change of the number of companies and technologies.

[6] Please note that we are still considering that investors are interested only in previous investments for now.

Concerning the parameters' calibration, Figure 5.7a provides an idea of how timing changes by increasing the number of entities in cybersecurity, for both companies and technologies. Interestingly, the run-time growth does not depend on the number of entities only: for technologies the curve is much steeper than that for companies. However, considering that the number of technologies is a direct consequence of the number of companies, we can consider as x-axis the number of companies and technologies together[7] (e.g. 10 companies work on 26 technologies). Therefore, we can plot companies and technologies together: Figure 5.7a shows that the run-time of the parameters' calibration phase for the two entities is much more similar if we consider them together.

The random walk phase lines in Figure 5.7b represent the amount of time needed to run the algorithm to come to convergence. The similarity between the run-time for companies and technologies is interesting because, considering the numbers of the entities are very different, we would have expected significant different run-times. This is another factor that shows how much the two are correlated and supports the capability of TechRank to capture the complex coordination scheme of the cybersecurity technological landscape. It is possible to find in Appendix A2 a precise table including all the run-times.

5.1.3 Exogenous factors

Section 5.1 presents our results for the cybersecurity field, supposing we focus on previous investments only (100% for both companies and technologies). Here, we present how much results change by setting different preferences. Due to long run-time reasons, we analyse 1000 companies only.

For instance, we suppose that the investors are interested in the geographical position of the company only. To study how the results change

[7] Please note that the number of technologies increases with the number of companies.

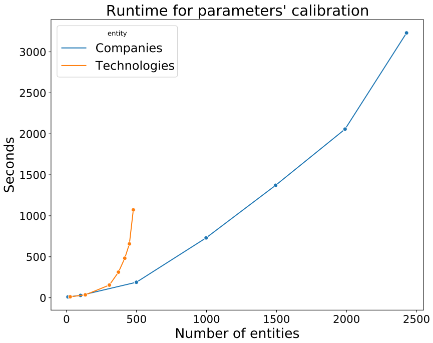
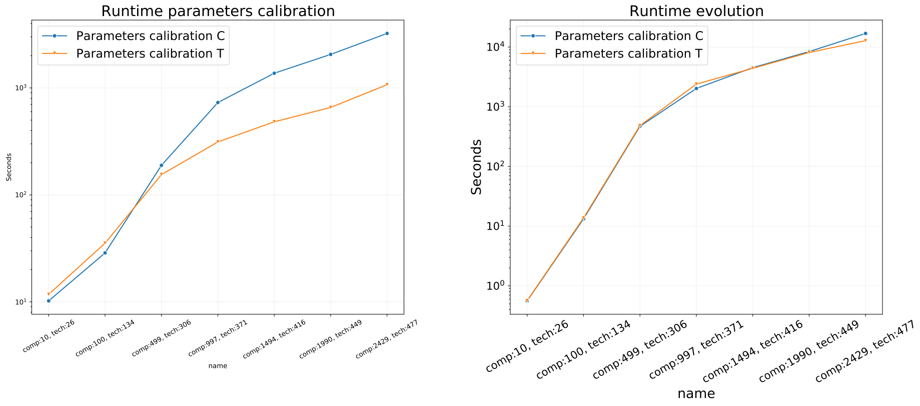


Figure 5.6: Run-time for parameters' calibration for the cybersecurity field.



(a) Parameters' calibration.

(b) Ranks' evolution.

Figure 5.7: Run-time for the cybersecurity field (y log scale).

according to this preference, we first assume an investor to be based in New York City and, then, to be based in San Francisco. Table 5.1 shows the state/region and the country of the five top ranked companies. We find a positive change in the position of the companies: the first one is in the state of New York if the investor is based in New York City, while the top company is in California if the investor is based in San Francisco. Also the other top companies show a good correlation between the HQ of the company and the investor' preference, even if we find some exceptions, like Singapore and Beijing. These exceptions show that, even if these companies are disadvantaged due to their long distance, their rank overcomes this flaw placing them among the most attractive companies. Therefore, even if their locations is not ideal, it might make sense to take them into account.

Investors' HQ	1	2	3	4	5
New York	'New York' 'USA'	'Massachusetts' 'USA'	'Quebec' 'CAN'	'California' 'USA'	'Singapore'
San Francisco	'California' 'USA'	'Illinois' 'USA'	'California' 'USA'	'Beijing' 'CHN'	'Arizona' 'USA'

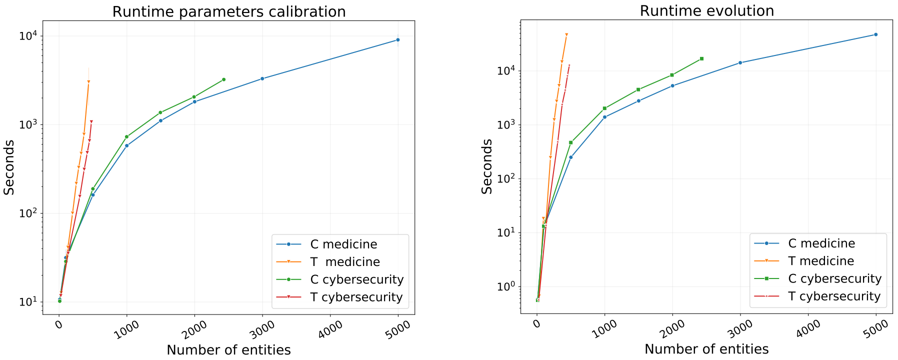
Table 5.1: Position of the first five companies when preferences are 100% in the HQ of the investors (New York and San Francisco).

5.2 Other fields

In order to expand our research and understand how results change working in other fields, we test TechRank for companies belonging to the *medical* field. We select the companies with the same methodology employed for the cybersecurity sector and we get a total of 4996 companies and 437 technologies.[8] We choose this sector because there are many companies working on it and this allows us to test TechRank with more than twice the number of companies working in cybersecurity.

Figure A.5 shows the run-time of TechRank in the medicine sector, but we believe that Figure 5.8 is more useful, because it compares the run-times of the cybersecurity and medical sectors. In order to make the two fields comparable, we set as x-label the number of entities, for both companies and technologies. The results reveal that the run-time of the two fields, for both the parameter calibration and the random walker steps, follow the same behaviour, for both companies and technologies. Increasing the number of entities, we do not detect any unexpected change in terms of run-time: the technology evolution lines are very steep, while the companies ones have a linear form (the y-label is logarithmic).

We also note that, while in the work by Klein, Maillart, and Chuang [28] α remains almost constant and β changes significantly, Table A.3 for cybersecurity and Table A.5 for medicine show that changing the field both parameters significantly change.



(a) Parameters' calibration.

(b) Ranks' evolution.

Figure 5.8: Run-time comparison between the medical and cybersecurity sector.

[8] Please find the complete list of words related to medicine in Appendix A2.2

Chapter 6

Conclusion

This last chapter summarises the methodology and the results of TechRank. We then investigate the limitations of our research and suggests further steps that may be taken in the future.

6.1 Conclusion

In this thesis we introduce TechRank, a complex-system approach for modelling portfolios using an algorithm that assigns a score to companies and technologies. This methodology constitutes the first step towards a new data-driven investment strategy, which enables investors to follow their preferences while benefiting from a quantitative and data-driven approach.

We start by adapting an article by Klein, Maillart, and Chuang [28] and we then include the preferences of investors in the parameters' calibration step. In particular, we develop a ground truth score that summarises investors preferences and whose components (one for each preference) are the result of a case-by-case study.

The results show that the algorithm stops after a certain number of iterations, which depends on the number of entities and the complexity of the relationships within the bi-partite network. Using a restricted number of companies in cybersecurity, we analyse the TechRank scores and we try to explain why some entities increase and some decrease their scores during the evolution. Moreover, we explore how results change moving the investors' focus from previous investments to the geographical position of the company. We also study how TechRank behaves when applied to the medical field and we observe that there are not any unexpected changes.

We believe that our approach, which directly depends on the capabilities of the companies, is a good way for investors to analyse if an entity has good development prospects. This, together with the inclusion of exogenous factors, leads to a personalizable and see-through approach, new compared to the common un-adaptable rankings currently available, like the Crunchbase rank.

Thanks to the interdisciplinary core of the TechRank solution, investors can undertake a transparent decision-making process when dealing with highly complex scenarios, as the cybersecurity market. Our TechRank is the kickoff for a complementary (or even alternative) modern technological portfolio analysis.

In conclusion, even if we develop the methodology in the cybersecurity domain, we believe that the algorithm is extendable to every large organisation dealing with a high level of uncertainty. Therefore, we expect that the TechRank algorithm will be applied and further developed in other fields.

6.2 Limitations

The limitations of the present study naturally embrace how we select technologies related to cybersecurity. We choose them according to their short description provided by CB: we search if it includes one or more words related to cybersecurity, e.g. security, privacy, confidentiality, defence. Considering these words can be used also in other fields, we require that the description contains at least two of them for classifying the companies as related to cybersecurity. This strategy may be improved by employing more sophisticated techniques, such as *natural language processing* (NLP) [34].

This work is also limited by the lack of information about how each company allocates its resources among all the technologies it is working on. As a matter of fact, we only have a list of technologies for each company without any further detail. It would be very helpful to understand how many funds are reserved for each technology. Moreover, if we have access to time series, it would be interesting to study how the bipartite network changes.

Regarding the last point, our availability of time series is limited by the Crunchbase platform. As mentioned in Section 6.3, we aim to solve this constraint employing data from TMM.

We also note that the introduction of exogenous variables may be biased due to the potential presence of outliers. As a matter of fact, in order to normalise the factors, in Section 4.2 we divide the factors by their maximum. However, this may lead to unproportionate results if the maximum is an outlier. For instance, if the maximum has a different order of magnitude compared to the other factors, the results will be one company with factor one (the maximum in fact) and all the others will be significantly smaller. On the other hand, for another preference, all the factors may be in the same order of magnitude as the maximum and, thus, the results will be all near to one. When it is time to put two preferences together, even if the first one is preferred by the investor and it is assigned a larger percentage, overall it will have a much smaller weight compared to the second one, because the magnitude of the factors is significantly smaller. However, we do not think that removing outliers is a viable solution, because this would lead to overlook potentially profitable opportunities.

Notwithstanding these limitations, this work suggests a new methodology for scoring entities, including external preferences. As all new-born ideas, it has some shortcomings.

6.3 Further research

When the TMM platform will include the information needed for the customisation the algorithm, this research can be expanded in order to include *time series*. We may investigate what happens when a company stops to work on a technology and what is the impact of new technologies. Further work needs to be done to establish whether all new technologies are successful –this is unlikely– or not. For the latter case, we may analyse after how much time a new technology is set aside in case it is not worth it. Understating this is crucial in order not to invest too hastily on exciting new ideas. A greater focus on *percolation theory* can produce interesting findings that account more for studying the effects of a node disappearance on the overall network structure [36]. To the best of our knowledge, no work focusing on predicting the "birth" and "survivability" of links between entities that compose the cybersecurity technological landscape has been done. This is another relevant study that can be carried out. For this purpose, *technology-forecasting Machine Learning models* should lead to interesting results[1]. Including time series into the algorithm may also mitigate limitations due to the lack of information about how companies split their sources.

As mentioned in Section 4.2.3, further research should be devoted to investigate how including other exogenous factors in TechRank, in order to give to investors the widest range of features to pick from possible.

Further research might explore the robustness of parameters: we may analyse if results remain unaffected by small variations in α and β. Moreover, further investigation should confirm the pertinence of the TechRank algorithm in other fields with much more entities. When the number of nodes increase, coordination problems may arise.

In addition, further work should be devoted for extra validation, especially when time series will be included. We check in Chapter 5 if our results are relevant with a small number of nodes, but this becomes much harder to test increasing the number of companies, because the coordination structure becomes extremely complex.

[1] Another student, who has been working with us at the CYD campus, is studying percolation theory and using supervised learning in order to make link-investigations predictions.

A further study could assess the long-term effects of the TechRank algorithm on investments returns and technologies' development. In order to do that, we need time series and a detailed analysis of effects.

In conclusion, this research has introduced many interesting research questions that need further investigation.

Bibliography

[1] Federico Battiston, Vincenzo Nicosia, and Vito Latora. "Structural measures for multiplex networks". In: *Physical Review E* 89.3 (Mar. 2014). ISSN: 1550-2376. DOI: 10.1103/physreve.89.032804. URL: http://dx.doi.org/10.1103/PhysRevE.89.032804.

[2] Núria Bautista Puig, Elba Mauleón, and Elías Casado. "The role of universities in sustainable development (SD)". In: Aug. 2019, pp. 91–116. ISBN: 9781315150161. DOI: 10.1201/b22452-5.

[3] Alex Bavelas. "A mathematical model for group structures". In: *Applied Anthropology* 7.3 (1948), pp. 16–30. ISSN: 00932914. URL: http://www.jstor.org/stable/44135428.

[4] Marko Bender. *Spatial proximity in venture capital financing: A theoretical and empirical analysis of Germany*. Jan. 2011, pp. 1–358. ISBN: 978-3-8349-2684-5. DOI: 10.1007/978-3-8349-6172-3.

[5] Michele Benzi, Ernesto Estrada, and Christine Klymko. "Ranking hubs and authorities using matrix functions". In: arXiv:1201.3120 (Jan. 2012), arXiv:1201.3120. arXiv: 1201.3120.

[6] Phillip Bonacich. "Factoring and weighting approaches to status scores and clique identification". In: *The Journal of Mathematical Sociology* 2.1 (1972), pp. 113–120. DOI: 10.1080/0022250X.1972.9989806. URL: https://doi.org/10.1080/0022250X.1972.9989806.

[7] Steven M. Bragg. "Initial Public Offering". In: (2011), pp. 263–287. DOI: https://doi.org/10.1002/9781118268360.ch15.

[8] Sergey V. Buldyrev et al. "Catastrophic cascade of failures in interdependent networks". In: *Nature* 464.7291 (Apr. 2010), pp. 1025–1028. ISSN: 1476-4687. DOI: 10.1038/nature08932. URL: http://dx.doi.org/10.1038/nature08932.

[9] Margit Bussmann, Indra de Soysa, and John R. Oneal. "The Effect Of Foreign Investment On Economic Development And Income Inequality". In: 18718 (2002). DOI: 10.22004/ag.econ.18718.

[10] Tony Byrne and Jarrod Gingras. *The Right Way to Select Technology*. Rosenfeld, 2017. ISBN: 1-933820-93-4.

[11] João Canito et al. "Unfolding the relations between companies and technologies under the Big Data umbrella". English. In: *Computers in Industry* 99 (), pp. 1–8. ISSN: 0166-3615. DOI: 10.1016/j.compind.2018.03.018.

[12] Clayton M. Christensen. *The Innovators Dilemma: When New Technologies Cause Great Firms to Fail.* Harvard Business Review Press, 2016. ISBN: 9781633691780.

[13] Jean-Michel Dalle, Matthijs den Besten, and Carlo Menon. "Using Crunchbase for economic and managerial research". In: 2017/08 (Nov. 2017). DOI: 10.1787/6c418d60-en. URL: https://ideas.repec.org/p/oec/stiaaa/2017-08-en.html.

[14] Jean-Michel Dalle, Matthijs den Besten, and Carlo Menon. "Using Crunchbase for economic and managerial research". In: 2017/08 (Nov. 2017). DOI: 10.1787/6c418d60-en. URL: https://ideas.repec.org/p/oec/stiaaa/2017-08-en.html.

[15] Dimitri Percia David. "Data Management Plan (DMP) of the 'Technology Forecastingand Market Monitoring for Cyber-Defence' project". In: (2020).

[16] Donato, D. et al. "Large scale properties of the Webgraph*". In: *Eur. Phys. J. B* 38.2 (2004), pp. 239–243. DOI: 10.1140/epjb/e2004-00056-6. URL: https://doi.org/10.1140/epjb/e2004-00056-6.

[17] Jiansheng Fang et al. "Probabilistic Latent Factor Model for Collaborative Filtering with Bayesian Inference". In: *2020 25th International Conference on Pattern Recognition (ICPR)* (Jan. 2021). DOI: 10.1109/icpr48806.2021.9412376. URL: http://dx.doi.org/10.1109/ICPR48806.2021.9412376.

[18] Linton C. Freeman. "Centrality in social networks conceptual clarification". In: *Social Networks* 1.3 (1978), pp. 215–239. ISSN: 0378-8733. DOI: https://doi.org/10.1016/0378-8733(78)90021-7. URL: https://www.sciencedirect.com/science/article/pii/0378873378900217.

[19] Linton C. Freeman. "Centrality in social networks conceptual clarification". In: *Social Networks* 1.3 (1978), pp. 215–239. ISSN: 0378-8733. DOI: https://doi.org/10.1016/0378-8733(78)90021-7. URL: https://www.sciencedirect.com/science/article/pii/0378873378900217.

[20] Andrew H. Gold, Arvind Malhotra, and Albert H. Segars. "Knowledge Management: An Organizational Capabilities Perspective". In: *J. Manage. Inf. Syst.* 18.1 (May 2001), pp. 185–214. ISSN: 0742-1222. DOI: 10.1080/07421222.2001.11045669. URL: https://doi.org/10.1080/07421222.2001.11045669.

[21] Lawrence A. Gordon et al. "Empirical Evidence on the Determinants of Cybersecurity Investments in Private Sector Firms". In: *Journal of Information Security* 9.2 (2018), pp. 133–153. DOI: 10.4236/jis.2018.92010.

[22] Aric A. Hagberg, Daniel A. Schult, and Pieter J. Swart. "Exploring Network Structure, Dynamics, and Function using NetworkX". In: *Proceedings of the 7th Python in Science Conference.* Ed. by Gaël Varoquaux, Travis Vaught, and Jarrod Millman. Pasadena, CA USA, 2008, pp. 11–15.

[23] Charles R. Harris et al. "Array programming with NumPy". In: *Nature* 585.7825 (Sept. 2020), pp. 357–362. DOI: 10.1038/s41586-020-2649-2. URL: https://doi.org/10.1038/s41586-020-2649-2.

[24] 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. ISSN: 0027-8424. DOI: 10.1073/pnas.0900943106. eprint: https://www.pnas.org/content/106/26/10570.full.pdf.

[25] J. D. Hunter. “Matplotlib: A 2D graphics environment”. In: *Computing in Science & Engineering* 9.3 (2007), pp. 90–95. DOI: 10.1109/MCSE.2007.55.

[26] Dr P. V. Ingole, Mr Mangesh, and K. Nichat. “Landmark based shortest path detection by using Dijkestra Algorithm and Haversine Formula”. In: *International Journal of Engineering Research and Applications(IJERA)* (May 2013). ISSN: 2248-9622.

[27] Lyman M. Kells, Willis F. Kern, and James R. Bland. *Plane And Spherical Trigonometry*. McGraw Hill Book Company, 2018. ISBN: 978-9353868109.

[28] Maximilian Klein, Thomas Maillart, and John Chuang. “The Virtuous Circle of Wikipedia: Recursive Measures of Collaboration Structures”. In: CSCW ’15 (2015), pp. 1106–1115. DOI: 10.1145/2675133.2675286. URL: https://doi.org/10.1145/2675133.2675286.

[29] Maciej Kurant and Patrick Thiran. “Layered Complex Networks”. In: *Phys. Rev. Lett.* 96 (13 Apr. 2006), p. 138701. DOI: 10.1103/PhysRevLett.96.138701. URL: https://link.aps.org/doi/10.1103/PhysRevLett.96.138701.

[30] Justine Kyove et al. “Globalization Impact on Multinational Enterprises”. In: *World* 2.2 (2021), pp. 216–230. ISSN: 2673-4060. DOI: 10.3390/world2020014. URL: https://www.mdpi.com/2673-4060/2/2/14.

[31] Yuxian Liang and Daphne Yuan. “Predicting investor funding behavior using crunchbase social network features”. In: *Internet Research* 26 (Feb. 2016), pp. 74–100. DOI: 10.1108/IntR-09-2014-0231.

[32] Wes McKinney. “Data Structures for Statistical Computing in Python”. In: *Proceedings of the 9th Python in Science Conference*. Ed. by Stéfan van der Walt and Jarrod Millman. 2010, pp. 56–61. DOI: 10.25080/Majora-92bf1922-00a.

[33] Andrew Metrick. *Venture Capital and the Finance of Innovation*. 2006.

[34] Prakash M. Nadkarni, Lucila Ohno-Machado, and Wendy Webber Chapman. “Natural language processing: an introduction.” In: *Journal of the American Medical Informatics Association* 18.5 (2011), pp. 544–551.

[35] Lawrence Page et al. “The PageRank Citation Ranking: Bringing Order to the Web”. In: 1999-66 (Nov. 1999). URL: http://ilpubs.stanford.edu:8090/422/.

[36] Mahendra Piraveenan, Mikhail Prokopenko, and Liaquat Hossain. “Percolation centrality: quantifying graph-theoretic impact of nodes during percolation in networks”. en. In: *PLOS ONE* 8.1 (2013). ISSN: 1932-6203. (Visited on 03/02/2021).

[37] Alan Rugman and Alain Verbeke. “A Perspective on Regional and Global Strategies of Multinational Enterprises”. In: 2004-19 (2004). DOI: https://doi.org/10.1057/palgrave.jibs.8400073.

[38] Akrati Saxena and S. R. S. Iyengar. “Global Rank Estimation”. In: (2017). URL: https://arxiv.org/abs/1710.11341.

[39] Akrati Saxena and Sudarshan Iyengar. “Centrality Measures in Complex Networks: A Survey”. In: *arXiv e-prints* (Nov. 2020). URL: https://ui.adsabs.harvard.edu/abs/2020arXiv201107190S.

[40] The pandas development team. *pandas-dev/pandas: Pandas*. Version latest. Feb. 2020. DOI: 10.5281/zenodo.3509134. URL: https://doi.org/10.5281/zenodo.3509134.

[41] Xiao Tu et al. "Novel Multiplex PageRank in Multilayer Networks". In: *IEEE Access* 6 (2018). DOI: 10.1109/ACCESS.2018.2807778.

[42] Michael L. Waskom. "seaborn: statistical data visualization". In: *Journal of Open Source Software* 6.60 (2021), p. 3021. DOI: 10.21105/joss.03021. URL: https://doi.org/10.21105/joss.03021.

[43] Karen E. Wilson. "Financing High-Growth Firms: The Role of Angel Investors". In: *Social Science Research Network* (2011). DOI: http://dx.doi.org/10.2139/ssrn.1983115.

[44] Wenpu Xing and Ali Ghorbani. "Weighted PageRank Algorithm". In: Jan. 2004, pp. 305–314. DOI: 10.1109/DNSR.2004.1344743.

[45] Jerrold H. Zar. "Spearman Rank Correlation". In: *Encyclopedia of Biostatistics* (2005). DOI: https://doi.org/10.1002/0470011815.b2a15150.

[46] Hao Zhong et al. "Which startup to invest in: a personalized portfolio strategy". In: *Annals of Operations Research* 263 (Apr. 2018). DOI: 10.1007/s10479-016-2316-z.

Appendices

A1 Code

All the code is saved on a repository on GitHub into the Technometrics Lab (EPFL CYD Campus) page.[1] The repository is currently private, but it will become public once we have published an article about TechRank.

We implement the algorithm and run the experiments using the Python programming language.[2]. We used the following libraries: NumPy [23], Pandas [40, 32], NetworkX [22], Matplotlib [25], Seaborn [42].

It is also possible to find it as a ZIP file located on Google Drive.[3] The data we use are not available because they are released by Crunchbase to the CYD Campus with a proprietary licence that does not allow us to share them.

We create a complete documentation in HTML using Sphinx.[4] To open it, please open the zip file containing the code, go in *docs/build/html* and open the *index.html* file.

A2 Tables of results

A2.1 Cybersecurity field

List of words related to cybersecurity: *cybersecurity*, *confidentiality*, *integrity*, *availability*, *secure*, *security*, *safe*, *reliability*, *dependability*, *confidential*, *confidentiality*, *integrity*, *availability*, *defence*, *defensive*, *privacy.*

A2.2 Medical field

List of words related to medicine: *cure, medicine, surgery, doctors, nurses, hospital, medication, prescription, pill, health, cancer, antibiotic, HIV, cancers, disease, resonance, rays, CAT, blood, blood transfusion, accident, injuries, emergency, poison, transplant, biotechnology, health care, healthcare,*

[1] Technometrics Lab Organisation on GitHub URL: https://github.com/technometrics-lab/

[2] Python homepage: https://python.org

[3] URL: https://go.epfl.ch/thesis_techrank

[4] Sphynx homepage: https://www.sphinx-doc.org/en/master/

health-tech, genetics, DNA, RNA, lab, heart, lung, lungs, kidneys, brain, gynaecologist, cholesterol, diabetes, stroke, infections, infection, ECG, sonogram.

Number of C	**Number of T**	**Parameters' calibration C**	**Parameters' calibration T**	**Convergence C**	**Convergence T**
10	26	10.21	11.75	0.56	0.57
100	134	28.69	35.37	13.24	13.72
499	306	189.03	154.79	470.10	483.25
997	371	730.43	312.65	2,023.18	2,392.46
1494	416	1,372.17	482.18	4,514.11	4,404.48
1990	449	2,057.42	656.95	8,396.26	8,096.69
2429	477	3,230.99	1,071.84	16,890.26	12,779.62

Table A.1: Run-time in seconds for the cybersecurity field.

Number of C	**Number of T**	**Number iterations C**	**Number iterations T**
10	26	32	18
100	134	100	155
499	306	134	2469
997	371	196	194
1494	416	180	871
1990	449	240	5000
2429	477	723	1120

Table A.2: Number of iterations to reach stability for the cybersecurity field.

Number of C	**Number of T**	α^* **C**	β^* **C**	α^* **T**	β^* **T**
10	26	-0.36	1.92	-2.00	0.00
100	134	-0.04	0.92	0.52	-1.04
499	306	-0.08	0.88	0.68	-1.36
997	371	-0.12	0.80	-2.00	0.00
1494	416	-0.12	0.80	0.92	-0.12
1990	449	-0.04	0.92	0.56	-2.00
2429	477	0.04	-1.88	0.48	-2.00

Table A.3: Optimal parameters for the cybersecurity field.

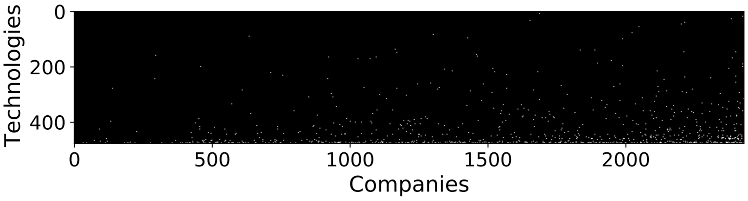
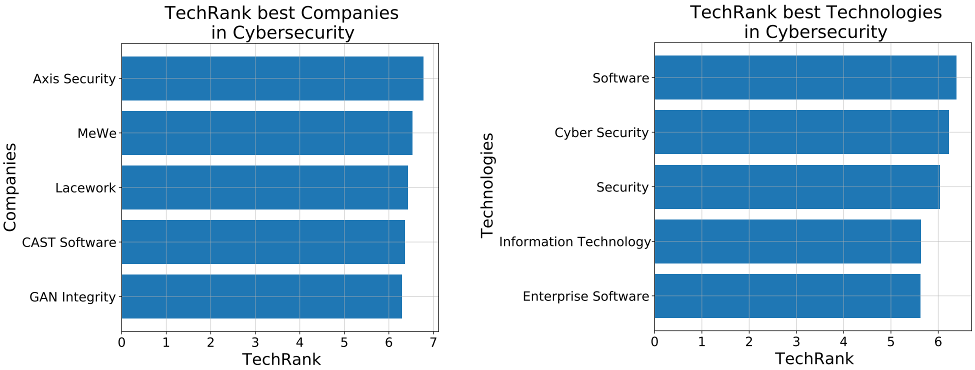


Figure A.1: Triangular structure of M^{CT} for 2429 companies and 477 technologies for the cybersecurity field.



(a) Companies.

(b) Technologies.

Figure A.2: Run-time for the cybersecurity field (y log scale).

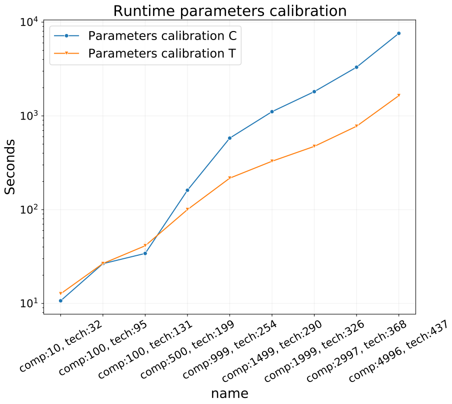


Figure A.3: Parameters' calibration.

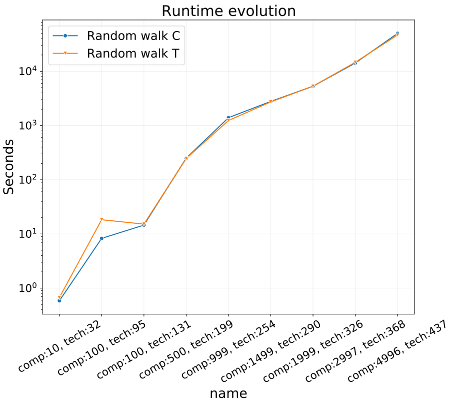


Figure A.4: Rank evolution.

Figure A.5: Run-time analysis for the medical field.

Number of C	**Number of T**	**Parameters' calibration C**	**Parameters' calibration T**	**Convergence C**	**Convergence T**
10	32	10.68	12.68	0.59	0.67
100	131	39.08	45.90	14.24	15.93
500	199	161.22	100.23	250.36	246.28
999	254	579.64	216.89	1,393.63	1,237.27
1499	290	1,109.84	328.63	2,783.47	2,724.17
1999	326	1,814.79	473.28	5,310.48	5,311.53
2997	368	3,306.56	776.12	14,188.12	14,652.66
4996	437	10,524.18	4,405.33	44,477.80	45,887.16

Table A.4: Run-time in seconds for the medical field.

Number of C	**Number of T**	α^* **C**	β^* **C**	α^* **T**	β^* **T**
10	32	1.48	0.32	0.64	-2.00
100	131	-0.24	0.52	0.80	-0.56
500	199	-0.28	0.32	-2.00	0.00
999	254	-0.04	1.04	1.48	-0.92
1499	290	-0.16	0.56	-2.00	0.20
1999	326	-0.12	0.64	-2.00	0.00
2997	368	-0.12	0.64	-2.00	0.00
4996	437	0.00	-2.00	0.52	1.08

Table A.5: Optimal parameters for the medical field.

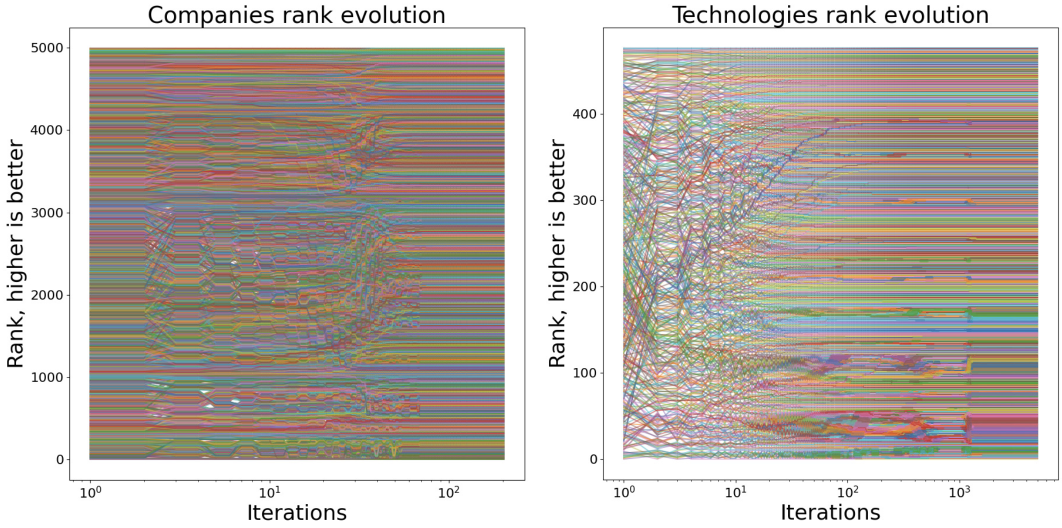


Figure A.6: TechRank scores evolution for 4996 companies and 437 technologies in medicine.