PLOS ONE

RESEARCH ARTICLE

Prediction and visualization of Mergers and Acquisitions using Economic Complexity

Lorenzo Arsini[1,4], **Matteo Straccamore**[1,2]*, **Andrea Zaccaria**[3,2]

1 Dipartimento di Fisica, Università "Sapienza", Rome, Italy, **2** Centro Ricerche Enrico Fermi, Piazza del Viminale, Rome, Italy, **3** Istituto dei Sistemi Complessi (ISC) - CNR, UoS Sapienza, Rome, Italy, **4** INFN, Section of Rome, Rome, Italy

* matteo.straccamore@cref.it

Check for updates

OPEN ACCESS

Citation: Arsini L, Straccamore M, Zaccaria A (2023) Prediction and visualization of Mergers and Acquisitions using Economic Complexity. PLoS ONE 18(4): e0283217. https://doi.org/10.1371/journal.pone.0283217

Editor: Vincenzo Basile, University of Naples Federico II: Universita degli Studi di Napoli Federico II, ITALY

Received: October 13, 2022

Accepted: March 3, 2023

Published: April 3, 2023

Copyright: © 2023 Arsini et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: Data cannot be shared publicly because it contains sensitive information about mergers and acquisitions between companies and the patenting activity of such companies. The restrictions are imposed by the company that gathered this data, Bureau van Dijk, and the data are available on their website (www.bvdinfo.com) for researchers who meet the criteria for access to the confidential data.

Funding: This work was supported by the Centro Ricerche Enrico Fermi (CREF) project "Complessità

Abstract

Mergers and Acquisitions represent important forms of business deals, both because of the volumes involved in the transactions and because of the role of the innovation activity of companies. Nevertheless, Economic Complexity methods have not been applied to the study of this field. By considering the patent activity of about one thousand companies, we develop a method to predict future acquisitions by assuming that companies deal more frequently with technologically related ones. We address both the problem of predicting a pair of companies for a future deal and that of finding a target company given an acquirer. We compare different forecasting methodologies, including machine learning and network-based algorithms, showing that a simple angular distance with the addition of the industry sector information outperforms the other approaches. Finally, we present the Continuous Company Space, a two-dimensional representation of firms to visualize their technological proximity and possible deals. Companies and policymakers can use this approach to identify companies most likely to pursue deals or explore possible innovation strategies.

Introduction

Mergers & Acquisitions (M&A) are one of the most popular forms of business development and represent the subject of considerable research in financial economics [1]. Such operations are used extensively as a financial instrument by firms of any region and size and constitute a business that, only in 2019, has almost reached 4 trillion dollars (Source: Institute for Mergers, Acquisitions, and Alliances (IMAA) https://imaa-institute.org). Despite the huge spread of the phenomenon, from a statistical point of view it is recorded that, on average, a M&A does not bring significant economic benefit to the involved companies [2]. There is, however, a strong variance in the data, which includes both acquisitions of huge success as well as dramatic failures. Despite the various attempts, there is no agreement in existing academic research about the right variables that decisively influence the realization and the outcome of an acquisition [3, 4]. The complexity of the phenomenon under consideration, as an economic and social process, is reflected in the heterogeneity of the studies carried out in this field, which lack comprehensive theoretical models and common variables [5]. In this work, we study M&A using

in Economia" and by the Istituto Sistemi Complessi (ISC-CNR) project "A data-driven complexity approach for economic growth". The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

tools and data from the Economic Complexity framework [6–9] and, in particular, the concept of Relatedness [10–12], that we use to compare the patenting activity of companies.

Every year, hundreds of new technologies are developed, as processes and products become increasingly complex. In such a rapidly changing environment, it is crucial for a company to stay ahead and get a good position in the innovation race. Often, developing new technologies in internal R&D environments is not enough or convenient in terms of time and costs. Because of this, many companies seek to expand their technological horizon by undergoing an acquisition. Through a technological acquisition, the acquiring company can absorb the target's innovation capabilities, recombine technologies to produce further innovation, intensify internal research and skills development [13], or launch products into a new market. In these terms, an acquisition can be seen as an expansion of the acquirer's knowledge and capability base. This kind of subject has been widely addressed in the corporate strategy literature that focuses on firms' diversification, as discussed in the following. Here we restrict our study to deals in which it is possible to individuate an acquirer and a target company. We refer to this set of deals with the generic term "acquisitions".

Several efforts have been done in researching a comprehensive theory on the diversification of firms (both at productive and technological levels) since the early studies of Penrose and Teece [14, 15]. Many authors took up these ideas to explain the reasons for the diversification of firms, the way in which they expand, and the financial outcome (for a complete literature review we refer to [16]). Beyond the concept of diversification as the simple scope of (both productive and patenting) activities in which a firm is involved, recent research focused on the key concept of the *relatedness* between these activities. A seminal work on this subject is the one of Teece et al. [17]. The authors measure the similarity between activities by counting the excess of their co-occurrences with respect to a suitable null model. Teece et al. show that activities that are more related are also more frequently combined within the same company. Teece's relatedness metrics were taken up by several authors whose researches focus on the internal coherence of firms' technological portfolios, that is, the innovative sectors to which their patents belong. This stream of literature substantially confirms Teece's results also in the technological field [18–20]. The patent-based approach offers various advantages such as the wide geographical coverage and the richness of information that patents provide about inventions (e.g. bibliographic data, citations, claims, technological fields impacted by the patents, etc.) [21, 22]. Similar themes were also faced using the approach known as Economic Complexity. This approach stands out for the use of tools from complexity science, such as co-occurrence networks [11, 23] and machine learning algorithms [24, 25]. In [26] this approach is used to study the technological diversification of firms and the relatedness between their activities. The authors introduced the concept of *coherent diversification* and showed that firms that diversify their patenting activity in a coherent manner, i.e. expanding in related sectors, have on average higher levels of labor productivity. Relatedness measures between companies and technological sectors can also be used to forecast the technological and productive diversification of firms [27, 28]. In this paper, we argue that methods similar to the ones described in these papers can be used to build relatedness measures to study the phenomenon of Mergers and Acquisitions (M&A).

In the recent period, the literature concerning M&A has been expanding in various directions. IN a recent work, Liu et al. [29] provide guidance to companies in terms of the risks involved in M&As by presenting success and failure cases. In [30], the authors seek to assess efficiency indicators for the acquiring organization and determine the determinants of the acquiring firm by examining the impact of ten listed companies in the sample that underwent M&A between 2010 and 2017. In [31] the differential effects of foreign investor protection (FIP) and domestic investor protection (DIP) on cross-border M&As are investigated, finding

that the effects are stronger for M&As with larger shares, in target countries further from home countries, and in target countries with better financial development. The research paper [32] aims to investigate the impact of the pandemic COVID-19 on global M&A activity. Ogendo and Ariemba [33] suggest that firms in emerging markets can use mergers and acquisitions during a downturn to deliver superior value to shareholders. Novita and Rasyid [34] analyze the company's financial performance through financial ratios before and after M&A of companies listed on the Indonesia Stock Exchange in a period covering the years 2016–2019. This study indicates that there are differences before and after mergers and acquisitions in the Return on Assets and Price Earning Ratios, but the other financial ratios tested did not show any differences when studied. Finally, [35] review the study of M&A in the financial sector, and in [36] the authors apply technology relatedness to study the M&A between Japanese Electric Motor firms.

Despite some sort of heterogeneity in the M&A economic literature, it is still possible to identify some principles and recurring ideas. First of all, the concept of *absorptive capacity* [37], is the ability of the acquirer company to assimilate knowledge and competencies of the target firm. Extending this concept, good integration between acquirer and target companies is thought to be linked to the relatedness between the two [38]. The concept of *relatedness*, and in general the interaction between resources and capabilities of the companies involved in a M&A process, is definitely the most widespread in the literature and is applied in many different contexts. For example, it has been shown how geographical distance negatively influences the probability of a M&A to occur [39, 40], and also how similarities in terms of the ownership or the industrial sector can, on the contrary, increase such probability [41, 42]. In [43], analyzing a large set of acquisitions and employing a similarity measure introduced by Teece et al. in [17], a statistically significant correlation between the occurrence of a M&A and the *industry relatedness* is found. As in the economic literature that focuses on diversification, in the last two decades, a section of the M&A studies started to analyze acquisitions from the point of view of the patenting activity of involved firms. This stream of literature focused on the similarity between the companies' *knowledge bases*, or in other words, their *technological relatedness*. One of the first studies of this kind is the one of Ahuja-Katilia [44]. The authors computed a measure of technological similarity between the acquirer and the target firms as the overlap of the set of all patents produced and cited by those firms. This measure is found to have an inverse parabolic behavior with the innovation performance after the acquisition. In other words, the optimal post-acquisition performances are observed for an intermediate level of technological similarity. In a similar fashion, many successive authors built different measures of technological relatedness and applied them to companies involved in M&A processes and tried to link them to post-acquisition performances. Examples of this stream of literature can be found in the work of Cloodt et al. [45], Cassiman et al. [46] and many others [47–51]. Although several studies recur the idea of this inverse parabolic behavior between relatedness and performances, results are not yet conclusive. Indeed, for now, there is not yet a standard, recognized, and effective method to build robust performances [49] or relatedness measures [52], and as shown in [49], results may vary on metrics definitions. In general, the majority of the M&A literature that builds relatedness measures between acquirers and target firms focuses on correlating such measures with successive performances and not on using them for predictions. However, as pointed out in [25], we believe that a forecast constitutes an important test to compare the goodness of relatedness assessments. Notable forecast exercise includes [53], in which an ensemble learning algorithm is trained on a set of relative features between companies, built using patent data, to predict future acquisitions, and the attempt to M&A prediction in [54]. In this latter work, a very large set of M&A features is built employing

financial, geographical, industrial, and patent data of firms. Then a tree-based algorithm is trained for M&A prediction.

In short, although the importance of relatedness between acquirer and target in an acquisition is now recognised, there is not yet a standard method for calculating and, most importantly, evaluating the goodness of such relatedness measures. There is, also, a fundamental lack of studies that compare different relatedness measures in a systematic way. In this heterogeneous context, we propose our M&A prediction study. We build upon a capability view of firms and follow the methods and ideas of the Economic Complexity stream of literature. Starting from patent data we define different relatedness measures between firms to exploit their technological affinity. These metrics are then used to make predictions on possible M&A pairs of companies and results are evaluated as in a machine learning classification problem. In this way, we can study in a systematical way the degree to which the considered relatedness measure is able to discriminate between pair of companies that complete an acquisition and randomly assembled pairs. We find that both the technological and the industrial affinity play a role and that cosine similarities outperform the others. Finally, we believe that it is also fundamental to have a simple visual representation of results for a straightforward interpretation, an effort lacking in the present literature. To fill this gap, we adapt the concept of Continuous Projection Space [24, 27] to our case, building a 2D space in which related companies are close to each other.

Results

Our investigation consists of four steps: i) we compute different measures that quantify how much a company, a possible target of an acquisition, is related to the present technological activity of the acquirer company; ii) we use these relatedness measures to forecast whether a deal will happen or not; iii) we quantify our ability to forecast the deals using different relatedness and performance measures; iv) we represent the deals in a two-dimensional plane. Before discussing our findings, we briefly introduce our methodology and data. More details are provided in the Methods section.

Data and testing procedure

The results of this work are based on the patent data coming from the PATSTAT database (www.epo.org/searching-for-patents/business/patstat) that contains information about over 40000 patents and their technology sectors of belonging. These are classified by the use of technology codes that are encoded using IPC's 6 digits classification (https://www.wipo.int/classifications/ipc/en/). From now on we will refer to these technology codes as "technologies". Subsequently, this information is matched with the AMADEUS database (https://amadeus.bvdinfo.com), which covers over 20 million firms with European registered offices. Finally, the M&A information comes from two different databases: Crunchbase (https://www.crunchbase.com) and Zephyr (https://login.bvdinfo.com/R0/zephyrneo). The final set of deals used for the analyses presented in this paper is made up of 8737 companies of which 913 are involved in 547 M&A deals. We select these companies because we can assign to them a univocal industrial sector, based on Crunchbase data on industrial sectors. Complete information on data processing and Crunchbase name matching, and industrial sectors classification can be found in the Methods section and in the S1 File. With these data, we can associate patents, and more specifically the technologies, with the companies involved in M&A; in particular, we build temporal bipartite networks connecting patenting companies with their technologies.

This temporal network is represented by 13 yearly adjacency matrices $\mathbf{M}^y$, one for each year y from 2000 to 2012 that link 8737 companies to 7132 technologies. The matrix element M^y_{ft}

represents how much a technology t is present in the patenting activity of firm f in year y: specifically, given a year y, we assign to each patent one unit of weight; this is then divided into equal shares between all the observed (firm f-technology t) pairs and, finally, the matrix is built by summing element-wise these contributions. This procedure takes into account that, usually, more than one code is present in each patent, and rarely a single patent is submitted by more than one applicant firm.

Moreover, assuming that a patent filed in a certain year is representative of the firm's innovation capabilities also in the following years, we will also consider the matrices M_{ft}^{Y}, each defined as the sum of M_{ft}^{y} over the years from 2000 to Y.

The $\mathbf{M}^{Y}$ matrices can be used to train different algorithms to calculate our predictions about possible M&A. In particular, we use $\mathbf{M}^{Y}$ to calculate the similarity between each pair of companies; such similarity is assumed to be related to the probability that two companies will have a M&A in the year Y.

Measures of similarity between companies

Our predictions of M&A events are based on various measures of business affinity, based using only patent data with adding in some cases information related to the company's industry sector. We give here a brief description of these metrics, referring the interested reader to the Methods section for a more detailed explanation. We will call a metric of *similarity* if it is computed between elements of the same type, for example between two companies, or two technologies. Instead, we will refer to metrics of *relatedness* if they are computed between different elements, such as a company and a technology. We divide our metrics into three different categories:

- Direct measures. These metrics are based on the construction of a similarity measure between firms. In this case, we can think of firms as vectors in a technology codes' space with coordinates $(M_{f1}^{Y}, \ldots, M_{ft}^{Y}, \ldots, M_{fn}^{Y})$, where $n = 7132$ is the number of technologies, that is the dimensionality of the space in which the firm-vectors are defined. We use as *direct* measures:
  - **Common Tech**, the scalar product between two firm-vectors, which provides the number of technologies that co-occur in both companies, possibly fractional, since **M** elements can be fractional;
  - **Jaffe**, the cosine of the angle between two firm-vectors, introduced by Adam Jaffe in [55] and adopted in this context in [48];
  - **Euclidean Distance (EU)**: the inverse euclidean distance between two firms;
  - **Jaffe + Sectors (J+S)**: we incorporate the information on technological portfolios (given by the Jaffe measure) with the information regarding the companies' industrial sector. In formula:

$$P_{ff'} = \alpha S_{ff'} + (1 - \alpha) J_{ff'},$$

    where J_{ff} represents the Jaffe measure between the firms' technological portfolio, S_{ff} is equal to 1 if both firms f and f' belong to the same sector and equal to 0 otherwise, and finally the parameter α can be tuned according to the goodness of the prediction on M&A processes. We find that the optimal value of $\alpha = \hat{\alpha}$ depends on the class imbalance of the prediction exercise; a detailed investigation is presented in the Methods section.
    All these measures can be also interpreted as a projection of the company-technology

bipartite network onto the company layer; this can be accomplished by using the method of co-occurrences and different normalizations [52].

- Indirect measures. Here we assess the distance between firms by performing two steps of computation. First, we build relatedness metrics between technologies and firms and second we evaluate the relatedness between firms. We employed two different ways to build the relatedness between firms and technologies: one is based on co-occurrences networks, and the other on Machine Learning.
  - Networks: Following the standard co-occurrence approach [17], we projected the bipartite network onto the technology layer using two different normalizations, obtaining two symmetric matrices $B_{tt'}$. We refer to these two normalizations as **Technology Space** (TS) [11] and **Micro-Partial** (MP), based on the work of Teece et al. [17]. Finally, we compute the prediction scores, called *coherence*, as $\gamma_{ft}^{Y} = \sum_{t'} M_{ft'}^{Y} B_{tt'}$.
  - **Random Forest**: The second approach to the construction of a relatedness measure between firms and technologies is based on the use of a machine learning algorithm. Following [27], we employ the Random Forest classifier [56], trained on patent data, and tested in [27] to predict the technologies that will be patented by firms in the future. The output of this classifier is a score RF_{ft}^{Y} which represents the likelihood that the link M_{ft}^{Y} is 1. These scores represent an optimal measure of the relatedness [24, 25], in this case, between a company and a technology. The functioning of RF, including training and testing, is explained in more detail in the Methods section.
    Given these different methodologies to assess the relatedness between firms and technologies, we compute the similarity between acquirer and target firms in a M&A deal as the mean relatedness between the acquirer and all the technologies in the target's portfolio. In the following, we will globally refer to these indirect measures as **Mean Coherence**: $\bar{\gamma}_{f,f'}$ and $\bar{RF}_{f,f'}$. So the measures we test are Mean Coherence Technology Space (**MC TS**), Mean Coherence Micro Partial (**MC MP**), and Mean Coherence Random Forest (**MC RF**). Due to their definition, $\bar{\gamma}_{f,f'}$ and $\bar{RF}_{f,f'}$ are, in general, highly correlated with the diversification of the firm *f*, i.e. the number of technologies to which *f* is linked, which, in our case, is the acquirer firm in the deal. Acquirers are often highly diversified firms and thus we want to asses if such a correlation can be leveraged to improve the results of our forecast exercise. In order to do so, we rescaled these two measures between 0 and 1. In this way, we will have rescaled Mean Coherence Technology Space (**MC TS resc**), rescaled Mean Coherence Micro Partial (**MC MP resc**), and rescaled Mean Coherence Random Forest (**MC RF resc**). Thanks to this rescaling the correlation between our relatedness measures and the diversification of firms significantly reduces. For example, the Spearman correlation between MC RF and diversification, after the rescaling, decreases from 0.65 to 0.42. More details on the rescaling effect and the correlation between Mean Coherence and diversification are given in the S1 File.

- **Continuous Company Space** (CCS): These two measures of similarity between firms refer to the construction of the Continuous Projection Space (CPS) [24, 27]. CPS represents a way to visualize the similarity between the nodes of one layer of a temporal bipartite network. In this case, we will represent companies in a two-dimensional space. In particular, we build the Continuous Company Space (CCS) starting from two measures of distance between companies, based respectively on Jaffe (we will refer to it as **CCS Jaffe**) and the Jaffe + Sectors model (we will refer to it as **CCS J+S**). CPS is instead usually built starting from the

prediction scores of a machine learning model [24]; here we build the CPS only using the best-performing measures. As we will show in the following, CCS has a minor predictive power than the original distances due to the loss of information in the dimensionality reduction process. Nevertheless, it represents an optimal way to visualize similarities between companies.

Visualization on CCS

In Fig 1, we present the CCS computed starting from the Jaffe + Sectors model with the parameter α fixed to 0.1. We decided to show the CCS computed also with the industrial sector information to give an easy visualization of the effects of the last. Here each point represents a company, and the relative distances are a low-dimensional representation of the distances provided by the Jaffe + Sectors model, obtained using the t-SNE algorithm [57]. On the upper panel, we colored the companies according to the respective industrial sector. As expected, since we are using also the information on companies' industrial sectors to build the similarity measure, we find a clear clustering among the firms that belong to the same sector. However, a relatively high number of companies end up in a cluster different from the one they should belong to according to the exogenous classification. This is the case, for instance, of many Integrated Control Technology (ICT) companies (light blue), which are spread into different communities on the leftmost side of the plot. These companies have a patenting activity much

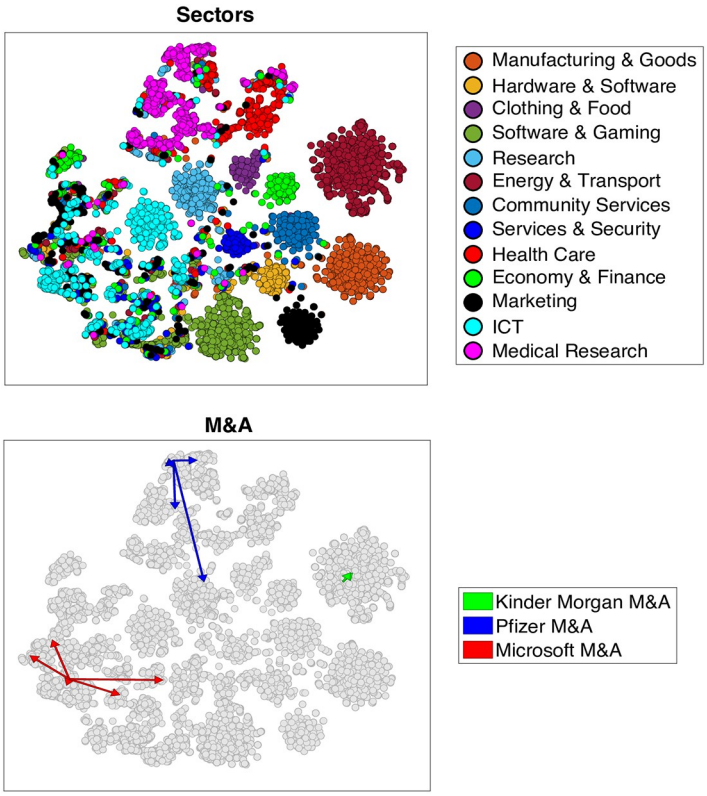

Fig 1. CCS Jaffe + Sectors ($\alpha = 0.1$). In the CCS (Continuous Company Space) representations above each point represent a company. On the upper panel, points are colored according to companies' industrial sectors. The effects of the information on the industry are visible from the clustering of firms of the same color. On the lower panel, we represent on the CCS some acquisitions performed by three large companies as arrows from the acquirer firms to the targets. This is an example of how acquisitions are likely to be done locally in this space.

https://doi.org/10.1371/journal.pone.0283217.g001

similar to Marketing companies (black) and Software and Gaming (dark green). We recall that in the CCS the similarity between the two companies is given by their spatial closeness in this space. As a consequence, our testing assumption is that the closer two companies are in the CCS, the greater the likelihood of a deal between them.

A direct consequence of the company's disposition can be observed in the M&A behavior. On the lower panel, we show what M&A processes look like in this space. We draw arrows from 3 big acquirer companies (Pfizer, Microsoft, and Kinder Morgan) in three different sectors to their target firms. As can be seen, acquisitions are likely done locally in this space, meaning that the similarity measure provided by the CCS can be connected with firms undergoing a merger or acquisition process. Obviously, a company may also decide to perform a deal related to a target that is not close in this space. This is a strategic choice: companies may want to enter into a different area of the CCS and drastically diversify their technological activity. Also, in this case, the CCS provides a map to navigate the innovation space.

Predictions

In this section, we present the results of our forecast exercise. We employ all the measures of relatedness between companies described before to predict which companies will take part in a M&A process. The idea is that it is on average more probable that a deal will happen between closer companies. We will use this assumption to compare the different possible measures of relatedness.

We adopted the Best-F1 score to compare the different methods in terms of the goodness of their predictions. The Best-F1 represents the maximum value that can be obtained by finding the optimal threshold used to compute the F1 score [58, 59] (this metric is discussed in detail in the Methods section). The F1 score is defined as the harmonic mean between precision and recall. The highest possible value of F1 is 1, which indicates that both precision and recall are equal to 1, and the lowest possible value is 0 if one of the precision or recall is zero. Other performance metrics are discussed in the S1 File. For each acquirer-target pair, we calculate our similarity measures and then we rearrange them to an array of predictions $\mathbf{s}$, whose elements and size will be discussed in the following. Then we compare these predictions with the array $\bar{\mathbf{s}}$ whose elements are 1 or 0, if the pair is a positive example, i.e. the pair completes an acquisition, or not. To evaluate the goodness of our predictions, we compare $\mathbf{s}$ and $\bar{\mathbf{s}}$ computing the Best-F1. Finally, to show more easily interpretable results, we compute a normalized version of the Best-F1 metric, based on the method used in [54]. For each prediction, we divide the actual Best-F1 value with the one computed shuffling the $\bar{\mathbf{s}}$ array. In this way, a normalised Best-F1 value equal to 1 will correspond to a random prediction, while values greater than 1 will increasingly indicate more significant predictions.

The forecast can be formulated in two different exercises, that lead to different prediction arrays even if the measure is the same:

- **Pair Prediction**: Given a set of companies, we want to predict which pairs of firms will undergo a M&A process. This is the prediction task investigated in [54]; here the point of view is of an external observer who compares all the possible pairs.
- **Target Prediction**: Given an acquirer company, we want to predict which firm is likely to be its target. Here the point of view is of the acquirer: this firm compares all possible targets with its technological portfolio.

Since our prediction exercise is a classification problem, we need both positive and negative samples. Therefore, for each true acquisition, we extract N random examples, with the parameter N controlling the class imbalance of the problem. Because of this the shape of $\mathbf{s}$ and $\bar{\mathbf{s}}$ will

depend on the class imbalance and will be equal to $N_{\text{true}} \times (1 + N)$. Specifically, in the Pair Prediction case, we extract random pairs of companies, while in the Target Prediction case, we extract only the target firms. We repeat this exercise for each class imbalance 20 times, so extracting different sets of negative samples, and calculating the Best-F1 mean and standard deviation. Notice that, as a robustness check, we evaluated our predictions also using other performance measures than the Best F1 score. Since the results were fully compatible with the ones shown by the Best F1, we decided to present these in the main text and the others in the S1 File.

In Fig 2, we show the Best F1 of the prediction tasks for 3 different values of negative samples, and so of class imbalance (1 VS 1, 1 VS 20, and 1 VS 200) both for the Target and Pair prediction. With each of the three colors, we refer to the three different types of categories of metrics: direct, indirect, and CCS measures.

We also report all the Best-F1 values in Table 1 for a more quantitative comparison. Finally, in the S1 File, the interested reader can find the results of the Target Prediction task obtained by dividing the M&As on the basis of the industrial sector of the acquirer company. We find that *Healthcare*, *Services and security*, and *Medical Research* are the sector in which M&A deals are mostly connected to our measures of technological relatedness. On the contrary, our methods show a relatively low prediction performance for *Economy and Finance*. This is clearly a consequence of the different importance of patenting activity in these sectors.

General considerations. One first consideration is that, in general, the Best F1 values of the Pair Prediction task are higher than the ones in the Target Prediction case. This suggests that it is easier to predict which pair of companies are more likely to complete an acquisition than which target will be chosen from a hypothetical acquirer firm. This effect could be a direct consequence of how the tasks are defined. The two tasks differ mainly in the choice of negative samples. In the Target Prediction task, we are comparing different possible targets for the same acquirer and some of them can be similar to the acquirer company even if they don't undergo an acquisition process with it. On the other hand, in the Pair Prediction task, we are comparing the M&A pairs with randomly assembled pairs of companies that, in most cases, are quite different from each other. Moreover, as seen from the left panels of the Figures, if we consider a problem with only one negative example for each positive one, the performances of all measures are quite similar to each other and not so far from random predictions. In this case, due to the low-class imbalance, even a random prediction achieves non-trivial results, with a 50% chance of getting the prediction right. Increasing the class imbalance, differences between metrics become more evident, since the prediction task becomes harder. In both cases, the highest values of Best F1 are reached by the measure Jaffe + Sectors with the α values chosen a posteriori to maximize the performances at the given class imbalance. As one can see, this measure represents a positive correction to the Jaffe measure, which itself provides the second-best predictive performance among the similarity metrics we studied. Note that in a high dimensional sparse space as the technologies' one, with 7132 dimensions, it is often recognised that metrics based on cosine similarity or scalar product are a good choice [60].

The metric with the worst performance turns out to be the inverse euclidean distance *EU*, which is computed in a technology codes' space with 7132 dimensions. It is known that such a metric loses much of its descriptive power when the dimensionality of the space increases [61]. This behavior is usually referred to as the "curse of dimensionality": when the number of dimensions is high, data become sparser and some distance measures (like for example the l_k ones, with $k > 1$) lose informative power. As suggested by the authors of [61], we tried different measures with fractional k, however, we did not find any sensible improvement in the results.

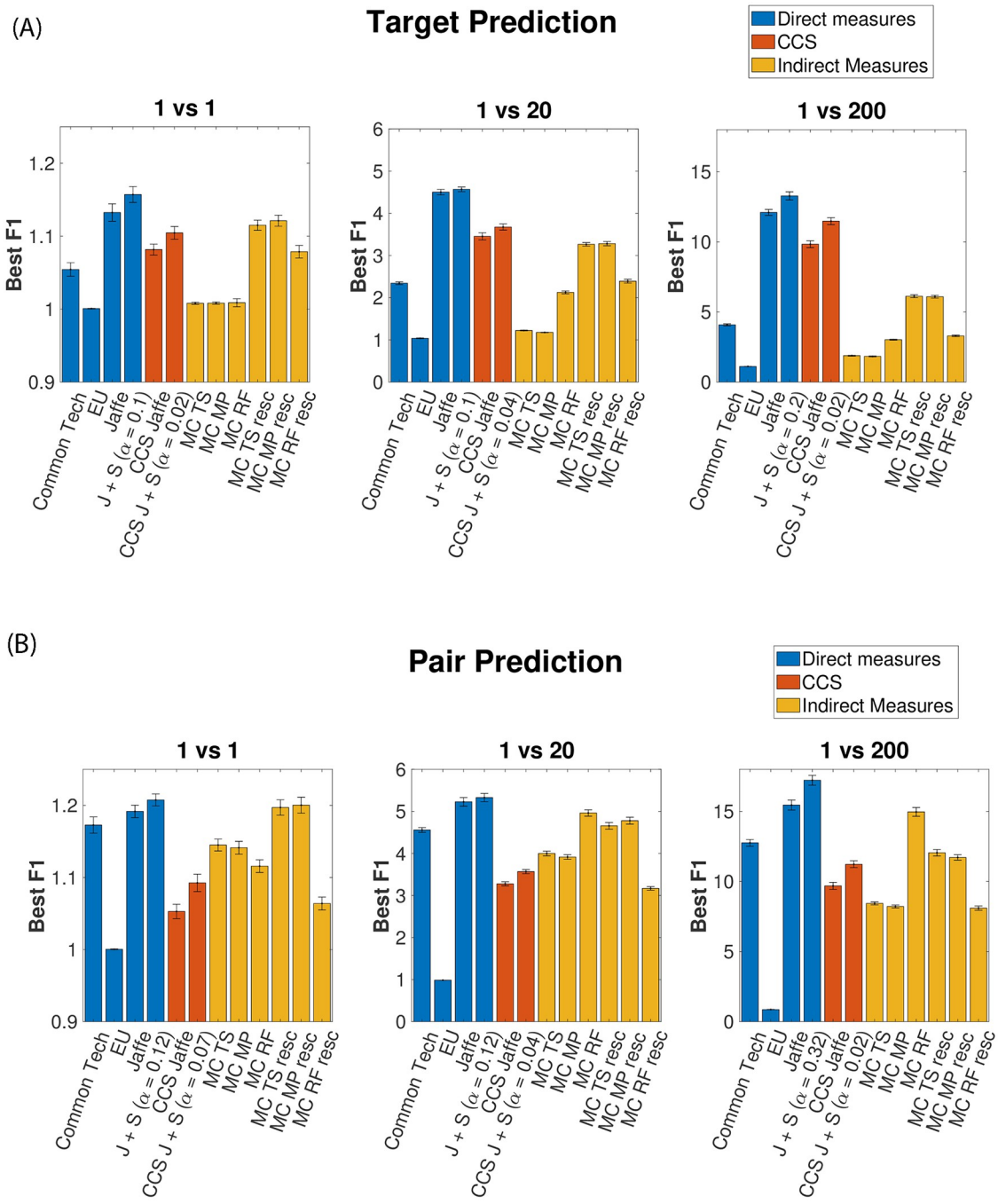

Fig 2. Predictions evaluation of Target and Pair forecasts. Comparison between the prediction performances of the different similarity metrics. We investigated different values of class imbalance, that is for each true M&A we extract 1, 20, and 200 negative cases (random couples with no deals). In both figures, we use three different colors to distinguish the typologies of metrics. Blue represents direct metrics, yellow indirect ones, and orange is the CCS ones. The error bars are computed by repeating the extraction of random Target companies and/or Acquirer companies 20 times. **A**: Comparison among the Best F1 scores on Target Prediction (fixed acquirer). **B**: Comparison among the Best F1 scores on Pair Prediction (both target and acquirer are not fixed). In both cases, Best-F1 values are normalised to ones of random predictions. With this normalisation, a Best-F1 value around 1 will indicate a random-like prediction, while values greater than 1 will correspond to more significant predictions. The Jaffe + Sectors metrics outperforms all other approaches.

https://doi.org/10.1371/journal.pone.0283217.g002

Table 1. Best-F1 values table of Fig 2. Jaffe + Sectors (J + S) constantly outperforms all the other measures on both Target and Pair Prediction and for all class imbalances. On the other hand, Euclidean Distance (EU) is always the worst performing. Due to the normalisation with respect to the random case, the Best-F1 values increase with the class imbalance, but the ranking of the different measures remains nearly the same.

	Target Prediction			Pair Prediction		
	1 vs 1	**1 vs 20**	**1 vs 200**	**1 vs 1**	**1 vs 20**	**1 vs 200**
Common Tech	1,05	2,34	4,09	1,17	4,56	12,76
EU	1,00	1,04	1,12	1,00	0,99	0,86
Jaffe	1,13	4,50	12,10	1,19	5,23	15,45
J + S	**1,16**	**4,57**	**13,28**	**1,21**	**5,33**	**17,22**
CCS Jaffe	1,08	3,45	9,84	1,05	3,28	9,69
CCS J + S	1,10	3,68	11,48	1,09	3,57	11,23
MC TS	1,01	1,22	1,89	1,15	4,00	8,43
MC MP	1,01	1,18	1,83	1,14	3,91	8,21
MC RF	1,01	2,13	3,02	1,12	4,96	14,96
MC TS resc	1,12	3,26	6,12	1,20	4,66	12,05
MC MP resc	1,12	3,28	6,09	1,20	4,78	11,71
MC RF resc	1,08	2,39	3,30	1,06	3,17	8,10

https://doi.org/10.1371/journal.pone.0283217.t001

CCS results. Regarding the results relative to the CCSs, directly using these metrics for predictions leads to worse results compared to the ones obtained employing Jaffe and Jaffe + Sectors. This is, in general, expected, since the construction of CCSs implies a dimensionality reduction that usually implies a loss of information. Nevertheless, the predictive performance of the CCS remains comparable with the other measures, especially in the 1 vs 1 and in the target prediction exercise.

Note that regarding the Jaffe + Sector measure, we performed an optimisation of the Best-F1 with respect to α for the relative CCS metric. In this case, the maximum of the Best F1 is less dependent on α and class imbalance than in the original measure. Moreover, best values of α are in most cases around or less than 0.05: in the CCS only information on industrial sectors is needed for optimal predictions. In Fig 2, among the other measures, we present the results relative to the CCS Jaffe + Sector with the optimised value of α for each class imbalance. The interested reader can find more details about the optimisation on α in the CCS's case in the S1 File. An observation regards the difference between Target Prediction and Pair Prediction in the CCS. While in the former the difference between the Best F1 of Jaffe and CCS Jaffe is only 25%, in the latter the difference rises to nearly 40%. This can be easily explained by saying that in CCS it is easier to make a prediction if we fix the acquirer, i.e. if we use knowledge of who will do the deal. In other words, if we do not fix the acquirer, CCS should be able to evaluate all possible pairs of points (companies) on it; fixed the acquirer, on the other hand, it will go to evaluate the proximity with a fixed point (acquirer companies) and the other points. Mathematically if we suppose to have N companies/points on the CCS, in the pair prediction we should evaluate $N \times (N-1)$ pairs of companies, instead of the target one only $N-1$ ones.

Rescaling effects. The performance of indirect measurements (yellow bars in the figure) depends strongly on whether or not we rescale them between 0 and 1. Remember that we use the rescaling to take into account the diversification of companies. On average, $\bar{\gamma}_{f,f'}$ and $\bar{RF}_{f,f'}$ are strongly correlated with the diversification of the firm f. Here we refer to the diversification of acquirers because, by construction, these measures refer to those companies.

With the rescaling, the $\bar{\gamma}_{f,f'}$ (i.e. "MC TS" → "MC TS resc" and "MC MP" → "MC MP resc" columns in Fig 2) predictions clearly improve; the prediction performance of $\bar{RF}_{f,f'}$ (i.e. "MC RF" → "MC RF resc" columns in Fig 2), instead, increases on the Target Prediction exercise,

while decreases in the case of the Pair Prediction. When rescaling, we observe two different effects on the performances that depend on the diversification of the acquirer companies.

After the rescaling, $\bar{\gamma}_{f,f'}$ loses most of its correlation with the diversification, while $\bar{RF}_{f,f'}$ maintains it only for highly diversified firms (the interested reader can find the relative plots of average *coherence* versus diversification in the S1 File).

If the diversification of the acquirers is the same in both positive and negative examples (i.e. real and not real M&A), rescaling, especially for $\bar{\gamma}_{f,f'}$, the predictive power of the metrics increases. In fact, if the metrics are not correlated with the diversification we can avoid both parts of False Positives and False Negatives. The first ones come from highly diversified random extracted acquirers that have, on average, high values of γ, while the second ones come from low diversified true acquirers that have, on average, low values of γ.

If the diversification of the negative cases' acquirers is lower than the positive cases' ones, the best predictions are made with the metrics without rescaling. In fact, in these cases, the correlation between the metrics and the diversification helps the prediction: not rescaling, we can have more True Positives and True Negatives samples, with respect to the rescaling one. The first ones come from highly diversified true acquirers that have, on average, high values of $\bar{\gamma}_{f,f'}$ and $\bar{RF}_{f,f'}$, while the second ones come from low diversified random extracted acquirers that have, on average, low values of $\bar{\gamma}_{f,f'}$ and $\bar{RF}_{f,f'}$. These two effects have two different consequences if we are treating Target or Pair prediction.

In the Target Prediction case, acquirers are always the same and so, their diversification doesn't change. We see from Fig 2**a** that rescaling the network-based indirect measures $\bar{\gamma}_{f,f'}$ leads to better performances, while the $\bar{RF}_{f,f'}$ one doesn't experience any strong change. In the Pair Prediction case, we can see in Fig 2**b** a combination of the two described effects on prediction. In fact, the diversification of random companies is lower than the one of acquirers, but this difference is not so wide. As a consequence, for the network-based measures, it is still convenient to rescale the metrics, even if the difference between the Best F1 level in the rescaled and not-rescaled case reduces, in comparison with the Target Prediction task. On the contrary, leaving the $\bar{RF}_{f,f'}$ measure without rescaling gives better results in predictions. For the sake of completeness, in the S1 File, we also report the analysis done extracting the negative examples among all companies in our dataset, and not extracting them from only the 913 final companies involved in the M&A.

Comparison between Jaffe and Common Tech. We remind the reader that, Jaffe is the cosine angle between two firms in the space of technologies, while Common Tech is the scalar product between them. In other words, Common Tech can be interpreted as the module of the magnitude that determines the technological proximity between companies, while Jaffe represents the versor. For a definition, it is clear that Common Tech is correlated with the diversification of the acquirers while Jaffe is not, and this is because the second is the normalized quantity (versor) of the first, i.e. we can see it as a rescaling between 0 and 1. This reflects in the fact that the difference in Best F1 values between the two measures is wider in the Target Prediction than in the Pair Prediction. Jaffe's better performance than Common Tech is telling us that, in the space of technologies, information about the module is not as important as information about the direction.

Discussion

Mergers and Acquisitions (M&A) represent a huge market, in which the innovative activity of companies plays a major role. Building on the studies of Penrose and Teece [14, 15], which concern the analysis of the diversification profiles of companies, we apply the Economic

Complexity approach focusing on technological diversification [18–20] to determine future M&A deals between companies. In particular, we compare different types of relatedness measures to assess the similarity between the acquiring companies and the target firms in terms of their technology portfolios, determined by their patenting activity. Even if various authors investigated how different similarity measures may be correlated to possible M&A deals [43–47, 50, 51, 53, 54], there is still no standard method for comparing such measures by evaluating their goodness. Our approach is to recast this issue as a classification problem, in which we forecast future deals and we measure our prediction performance using the indicators typical of machine learning tests.

To the best of our knowledge, this is the first attempt to assess the similarity between firms using Economic Complexity methods. In order to compare the prediction performance of various measures of similarity, we analyze a database consisting of 8737 firms, of which 913 are involved in 547 M&A deals, and 7132 technology sectors. We develop a forecasting exercise using the assumption that, on average, a pair of firms will more likely sign a deal if they are similar from a technological point of view. We find that the best performing metric uses the Jaffe cosine similarity between the two technological portfolios combined with the information about the industrial sector. This metric clearly outperforms the standard methodologies usually adopted in Economic Complexity, that is, networks of co-occurrences. Our results are robust with respect to two different types of forecasting exercises: the Pair Prediction, in which we want to find the most probable pair of firms; and the Target Prediction, in which we identify the "best" (in terms of similarity) target firm for a specific acquirer. Note that here we are using the forecast exercise as a test to compare different possible measures of similarity between firms; we are not claiming that a deal with a distant company is, in some sense, better or worse. We are simply assuming, as a test hypothesis, that is more probable for a firm to negotiate a deal with a close firm. A firm could perform a strategic jump and make a deal with a distant company to attack, for instance, a new market. In this context, we expect the Continuous Company Space (CCS) to be of practical help. The CCS is a visualization tool to represent the proximity between firms in a two-dimensional plane. It can be used to inform strategic M&A policies; for instance, to plan the expansion towards a relatively distant sector by acquiring a target company specialized therein.

Conclusion

This paper applies Economic Complexity methods to the investigation of M&A deals from the perspective of the technological activity (patenting) of companies. We illustrate various measures of the similarity between firms and we check if they are able to predict M&A deals. We assume that on average similar firms make more deals than firms that are different in terms of the sector and the technological scope. In this way we are able to compare the different measures by checking their prediction performance. We classified these measures into three types: direct, that is measures by which the similarity between firms can be measured directly; indirect, measures that start from the relatedness between firms and technologies, and then assess the similarity between firms; and the CCS, a two-dimensional representation used to visualize the similarity between firms as the vicinity of the points in a plane. To evaluate the effectiveness of these measures, we presented two forecast exercises: the Target Prediction and the Pair Prediction. In the first case, we fix the acquirer and we try to predict which company it will do a deal with; in the second case, we try to predict the pair (both acquirer and target) of the deal. The results presented show that our similarity measures are able to capture a M&A signal (i.e., perform better than a random prediction) only by using information about the technology portfolio of companies. In this perspective, the best measure turns out to be a cosine similarity

between the two technological portfolios combined with the information about the industrial sector. Moreover, we introduced the CCS, a visualization tool in which companies are located in a two dimensional plane according to the relative distances in terms of their technological portfolios. The CCS may be a useful tool to design strategic policies of M&A.

The results presented in this paper may pave the way for several future works. Among them, discriminating between successful and unsuccessful M&A. In particular, by leveraging data on the financial performance of companies, the CCS or the other measures discussed here can be used to study the outcome of deals as a function of the distance between the two firms. Another issue may be to investigate the overlap between the geography and the technology effect on M&A. This could be done using an "effective distance" (see [62, 63]) that can be measured by calculating the similarity between countries; the hypothesis being that the deals occur more frequently between similar countries.

Materials and methods

In this Section, we describe in more detail the databases, the proximity measures, and the metrics used in the analysis.

Data

The information used to perform the analysis of the present paper can be obtained from four databases. The two databases AMADEUS and PATSTAT contain the information used for the construction of the bipartite company-technology networks. Zephyr and Crunchbase contain information on the M&A. The companies' industrial sectors are obtained from the Crunchbase database.

Companies. The information regarding the companies was obtained from the AMADEUS database (https://login.bvdinfo.com/R0/amadeusneo). This database contains information about over 20M companies located mainly in Europe. AMADEUS is managed by Bureau van Dijk Electronic Publishing (BvD), which specializes in providing financial, administrative, and budget information relating to companies. The BvD includes the same patent identifiers as the European Patent Office and this makes the AMADEUS and PATSTAT databases compatible with each other [26]. Although one of the most well-known problems of AMADEUS is that large companies are fully covered while those with fewer than 20 employees are underrepresented [64], for the purposes of this paper this is not a relevant problem.

Technology codes. The source of data on technologies and patents is the Worldwide Patent Statistical Database (PATSTAT, https://www.epo.org/searching-for-patents/business/patstat.html) of the European Patent Office (EPO), which aggregates and organizes data from regional and national patent offices. The most important element of this database is the presence of a standardized code defined within the International Patent Classification (IPC), a hierarchical classification system, internationally recognized, maintained, and updated by the World International Patent Organization (WIPO). The codes are organized by levels of increasing aggregation: the lowest level includes over 70,000 groups, while the highest includes only 8 sections. This coding is used to classify each patent from a technological point of view. For example, the code Axxxxx corresponds to the macro category "Human Needs" and Cxxxxx to the macro category "Chemistry"; considering the following figures we have, for example, with A01xxx the sector "Agriculture; Hunt", and with A43xxx the "Footwear" sector. It is important to note that classes "99" and subclasses "Z" are not considered in this work, as they represent technologies classified in "other classes or subclasses", and therefore are not well defined. The interested reader can find more details about this data in the work of [65].

Zephyr and Crunchbase. Merger & Acquisition data were acquired from two different databases: Zephyr and Crunchbase. Zephyr (https://www.bvdinfo.com/en-us/our-products/data/greenfield-investment-and-ma/zephyr) is a commercial database, maintained by the Bureau van Dijk Electronic Publishing (bvd), that contains information on the operations of M&A, IPO, Private Equity, Venture Capital and related Rumour worldwide. Specifically, in this work, we used the section of the database that concerns companies operating in the biopharmaceutical sector. This section includes information on nearly 4000 deals between 1997 and 2016 and over 3700 companies. Crunchbase (https://www.crunchbase.com) is another commercial database, originally created to track start-ups, containing information on public and private companies and related acquisitions, mergers, and investments, globally. Crunchbase's dataset is much larger than the Zephyr one, and contains information on over 100 thousand acquisitions, from 1922, and over a million companies.

Data processing

In this Section, we briefly describe how we combined M&A data with the information on technological portfolios and the construction of our Industrial sectors classification.

M&A data processing. To study M&A processes in relation to the technological portfolios of the companies involved, we linked the Zephyr and Crunchbase datasets to the AMADEUS-Patstat one. As fully described in [26], the AMADEUS-Patstat data can be seen as a bipartite network where each company, identified by its Bureau van Dijk ID (BVDID), is linked to the technology codes of its patents. The weight of the link between a company and a technology code is proportional to the share of patents, deposited by the company, that contain that technology code. The linking process between the M&A data to AMADEUS-Patstat is different for each of the two datasets. In the Zephyr dataset, companies are identified with their BVDID, so it was possible to directly associate them with a technological portfolio. Starting from a set of 3167 companies involved in M&A processes, we were able to link 430 of them to the relative technological portfolios. Crunchbase's data are not labeled by the BVDID, so we had to match the names of Crunchbase's companies to the AMADEUS ones in order to find for each firm the associated BVDID. For a better match, companies' names underwent a "cleaning" process to remove symbols, punctuation, and companies' acronyms. The full process of names' cleaning and matching is described in the S1 File. After the cleaning process, we ended up with 28137 companies with a BVDID associated. From this set, we linked 12017 companies to the relative technological portfolio. Sometimes, due to the cleaning of names, companies resulted associated with multiple BVDIDs and thus multiple portfolios. These are occurrences in which, for example, a multinational company has multiple BVDIDs associated with the various national subsidiaries. In such cases, we merged all the technological portfolios associated with that company. Finally, for each year, in both Zephyr and Crunchbase cases, we kept only the M&A that happened between 2002 and 2012, whose acquirer and target companies deposited at least one patent from 2000 to that year. With this constraint, we managed to build a data set of 1279 M&A (126 from Zephyr and 1153 from Crunchbase), that involves 1974 companies (145 from Zephyr and 1858 from Crunchbase, with 29 present in both).

Sectors classification. Crunchbase companies are organized, concerning their industrial sector, in two levels of aggregation: the lower level counts 744 *categories*, while the upper counts 43 *category groups*. This classification is not directly linked to the official ones (NACE, NAICS, SIC, etc.) but was built independently by Crunchbase. In this classification, each company is assigned several category groups and thus many categories. Starting from this classification we built another level of aggregation consisting of 13 sectors. In this way, we managed to assign one univocal industrial sector to 8069 firms, nearly 70% of the Crunchbase

companies that we had previously linked to their technological portfolio. To have at least a sector linked to each company and a smaller number of sectors is fundamental for the construction of our modified version of CCS and its visualization. Further details on how our classification was built can be found in the S1 File.

The main results presented in this paper were obtained working on a subset of the M&A data set that includes only companies with univocal sectors assigned within our classification. This subset counts 8737 companies and 547 M&A that involve 913 companies of the total subset.

Data processing

Our final data can be used to construct 13 bipartite networks between companies and technology codes, one for each year from 2000 to 2012. We can represent these networks, for each year y, as a matrix with elements M^y_{ft}. Each matrix element represents the weight of the link between the firm f and the technology code t, in the year y; this is equal to the (possibly fractional) number of patents filed by f belonging to the technology t. Under the hypothesis that a patent filed in a certain year is representative of the firm's innovation capabilities also in the following years, for the construction of our relatedness measures we consider a summed version of the matrix M^y_{ft} over the years. We define M^Y_{ft} as the sum of all M^y_{ft} from 2000 to the year Y. From now, we drop the apex Y for simplicity, keeping in mind that all measures can be defined for each year.

Direct measures

Common Tech and Jaffe. The Common Tech and Jaffe metrics consist of a direct projection onto the companies' layer to measure the similarity between each company's pair. We calculate these two quantities by computing the equations:

$$\text{Common Tech}_{ff'} = \sum_t M_{ft} M_{f't},$$

$$\text{Jaffe}_{ff'} = \frac{\sum_t M_{ft} M_{f't}}{\sqrt{\sum_t M^2_{ft}} \sqrt{\sum_t M^2_{f't}}}.$$

where M_{ft} is the adjacency matrix that links firms to technologies. The element of these matrices represents how much a technology t is present in the patenting activity of firm f. The former is a simple scalar product between the technological portfolio of the two firms. This is correlated with the diversification of both f and f'. The latter is a cosine similarity between the two portfolios, introduced in this context by Jaffe [55]. It is bounded between 0 and 1.

These measures represent a projection of the bipartite network onto the firms' layer, so they can be interpreted as the weight of a link that connects the companies f and f' in a monopartite network of firms.

Euclidean distance. To build a relatedness measure between companies based on the euclidean distance we start from the matrix M_{ft}. Each row of this matrix can be seen as the list of coordinates of each company in the space of technology codes. The relatedness measures EU is just the inverse of the euclidean distance between companies in this space:

$$EU_{ff'} = \left(\sum_t (M_{ft} - M_{f't})^2 \right)^{-\frac{1}{2}}$$

Jaffe + Sectors and best α identification. The Jaffe + Sectors scores are computed by considering not only the technological affinity between companies but also the industrial sector. This leaves a degree of freedom (the relative weight) which we optimize as described in the following. The formula is:

$$P_{ff'} = \alpha S_{ff'} + (1 - \alpha) J_{ff'}, \quad (1)$$

where $J_{ff'}$ is the Jaffe measure between firms' technological portfolios and $S_{ff'}$ is 1 if both firms belong to the same sector and 0 otherwise. The weight of these two pieces of information is controlled by the parameter α, bounded between 0 and 1. The higher α, the greater the importance of the sectors' similarity on the measures. To understand the behavior of this measure it is useful to consider that the M&A pairs are distributed in 4 sets according to the respective sectors and the relative distance:

- $S1J1$: M&A with $S_{ff'} = 1$ and $J_{ff'} \neq 0$,
- $S1J0$: M&A with $S_{ff'} = 1$ and $J_{ff'} = 0$,
- $S0J1$: M&A with $S_{ff'} = 0$ and $J_{ff'} \neq 0$,
- $S0J0$: M&A with $S_{ff'} = 0$ and $J_{ff'} = 0$,

Due to the fact that $S_{ff'}$ can be only 0 or 1, while $J_{ff'} \in [0, 1]$, for $\alpha > 0.5$ all the Best F1 results are independent of α and equal to the one at $\alpha = 0.5$. In fact, for $\alpha > 0.5$ the elements in the four sets are bounded within their set: in $S1J1$, $P_{ff'} > \alpha$, in $S1J0$, $P_{ff'} = \alpha$, in $S0J1$, $0 < P_{ff'} < (1 - \alpha)$ and in $S0J0$, $P_{ff'} = 0$. Certainly, items in $S1J1$ are classified as positives, while items in $S0J0$ are classified as negatives so the threshold that defines the Best F1 must lie among the elements of $S0J1$ and $S1J0$. Finally, if we sort these elements in descending order, their order does not depend on α, thus neither the threshold nor the Best F1 does. For this reason, we examine the behavior of the measure for $\alpha \leq 0.5$. Another factor that influences the performance of the measure is the class imbalance, which is defined by the parameter N, namely the number of negative examples per positive example. In Fig 3 we present the behavior of the Best F1 as a function of α and the class imbalance N, both for the Pair Prediction and the Target Prediction. To average out the possible fluctuations coming from the random extraction of negative examples, each point in the Figures reports the average Best F1 over 20 realizations of the prediction exercise. Because the Best F1 is highly correlated with the class imbalance [66], for each value of N we rescaled the Best F1 between 0 and 1. In this way, it is possible to better spot the relative maximum of Best F1 as a function of α, for each value of class imbalance.

From both the figures it emerges that if we increase the class imbalance, the maximum of Best F1 moves [66] towards a higher value of α, especially in the Pair Prediction case. This suggests that when choosing a M&A pair among a large pool of options, the similarity between the companies from the point of view of the industrial sector plays a more important role. To deepen the understanding of the model, we also considered the absolute value of Best F1, as it is shown in Fig 4. In this plot, we show the Best F1 vs. α curves relative to the Target Prediction case for some values of class imbalance; the number of negative samples corresponds to lines of different colors. Since the Best F1 is highly correlated with the class imbalance the curves for different values of N never touch each other and can be represented in a single figure. In this case, the figures relative to Pair Prediction and Target Prediction were almost identical, so we decided to show only one of the two.

From Fig 4, it is possible to identify 3 phases, divided by dotted lines, for the Best F1:

- **Low α phase**. Located on the left of the figure; the Best F1 is α-dependent. First, it increases, reaches a maximum, and then decreases.

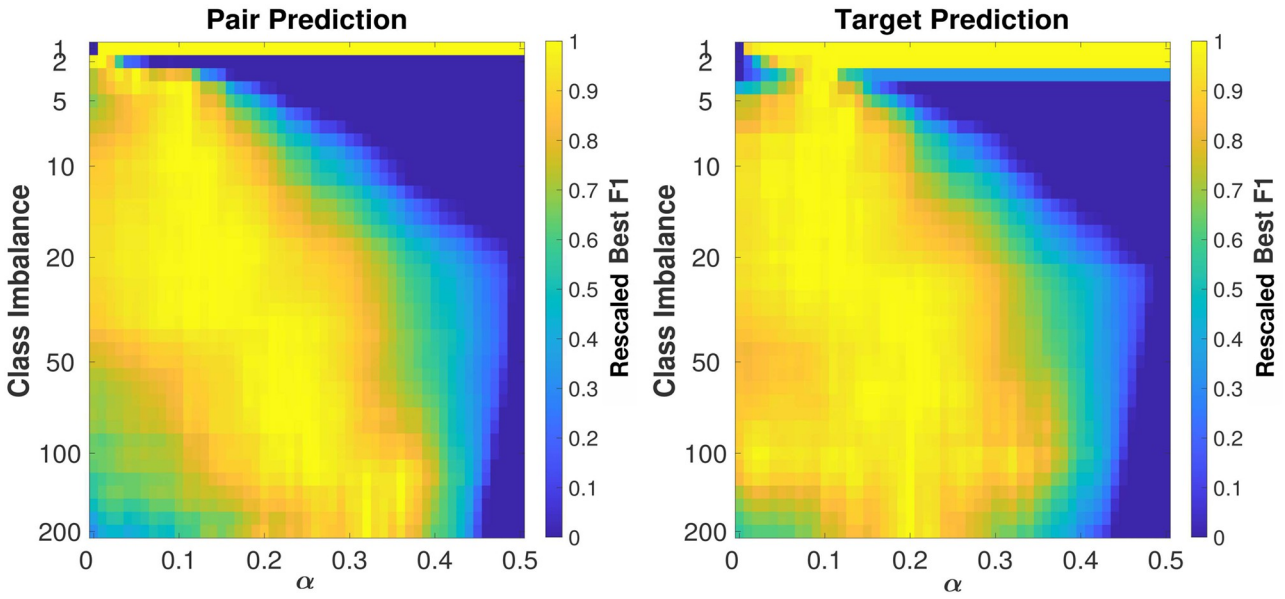

Fig 3. Dependence of maximum Best F1 on α and the class imbalance in the Jaffe + Sectors measure. For each value of class imbalance, we rescale the Best F1 between 0 and 1 to better visualize the maximum as a function of α. The maximum F1 moves towards higher α values when the class imbalance increases. This suggests that when choosing a M&A pair among a large pool of options, the industrial sector plays a more important role.

https://doi.org/10.1371/journal.pone.0283217.g003

- **High α—Low class imbalance phase**. Located at the top-right of the figure; it presents a fixed value of Best F1 independent from alpha.
- **High α—High class imbalance phase**. Located at the bottom-right of the figure; it presents a fixed value of Best F1 independent of alpha.

To understand this behavior we need to define some other sets in which the M&A examples can be divided with respect to the measure $P_{ff'}$. Specifically these are three subsets of $S0J1$, so they have $S_{ff'} = 0$ and differentiate according to the $J_{ff'}$ value:

- $\boldsymbol{J_+}$: M&A with $(1-\alpha)J_{ff'} > \alpha$,
- $\boldsymbol{J_-}$: M&A with $(1-\alpha)J_{ff'} < \alpha$,
- $\boldsymbol{J_\alpha}$: M&A with $(1-\alpha)J_{ff'} = \alpha$. In this subset $P_{tt'} = \alpha$ as in the $S1J0$ set, so in the subsequent analysis, it will be considered jointly with this last one.

Now, if we consider the $P_{ff'}$ values for each set in ascending order we have: $S0J0$ (i.e. M&A with $S_{ff'} = 0$ and $J_{ff'} = 0$ in Eq 1) with $P_{ff'} = 0$, J_- with $0 < P_{ff'} < \alpha$; $S1J0$ (i.e. M&A with $S_{ff'} = 1$ and $J_{ff'} = 0$ in Eq 1) with $P_{ff'} = \alpha$; J_+ and $S1J1$ (i.e. M&A with $S_{ff'} = 1$ and $J_{ff'} \neq 0$ in Eq 1) with $P_{ff'} > \alpha$. The observations on the behavior of the Best F1 are similar in style to the ones made before for $\alpha > 0.5$. In the upper-right side of Fig 4, in the High α—Low-class imbalance phase, the Best F1 is independent of α because the relative threshold is always less than α. In this case, the threshold lies between the elements of J_-, whose order (in terms of the magnitude of $P_{ff'}$) is independent of α. On the contrary, in the Low α phase, the threshold is always above α. This means that it lies among the elements of J_+ and $S1J1$. If we order these elements according to their $P_{ff'}$ value, the two sets are, in general, mixed and the order of elements depends on α. So the Best F1 is α dependent. This remains true until both α and the class imbalance become too large. In this phase, the High α—High-class imbalance one, the J_+ set becomes negligible to the

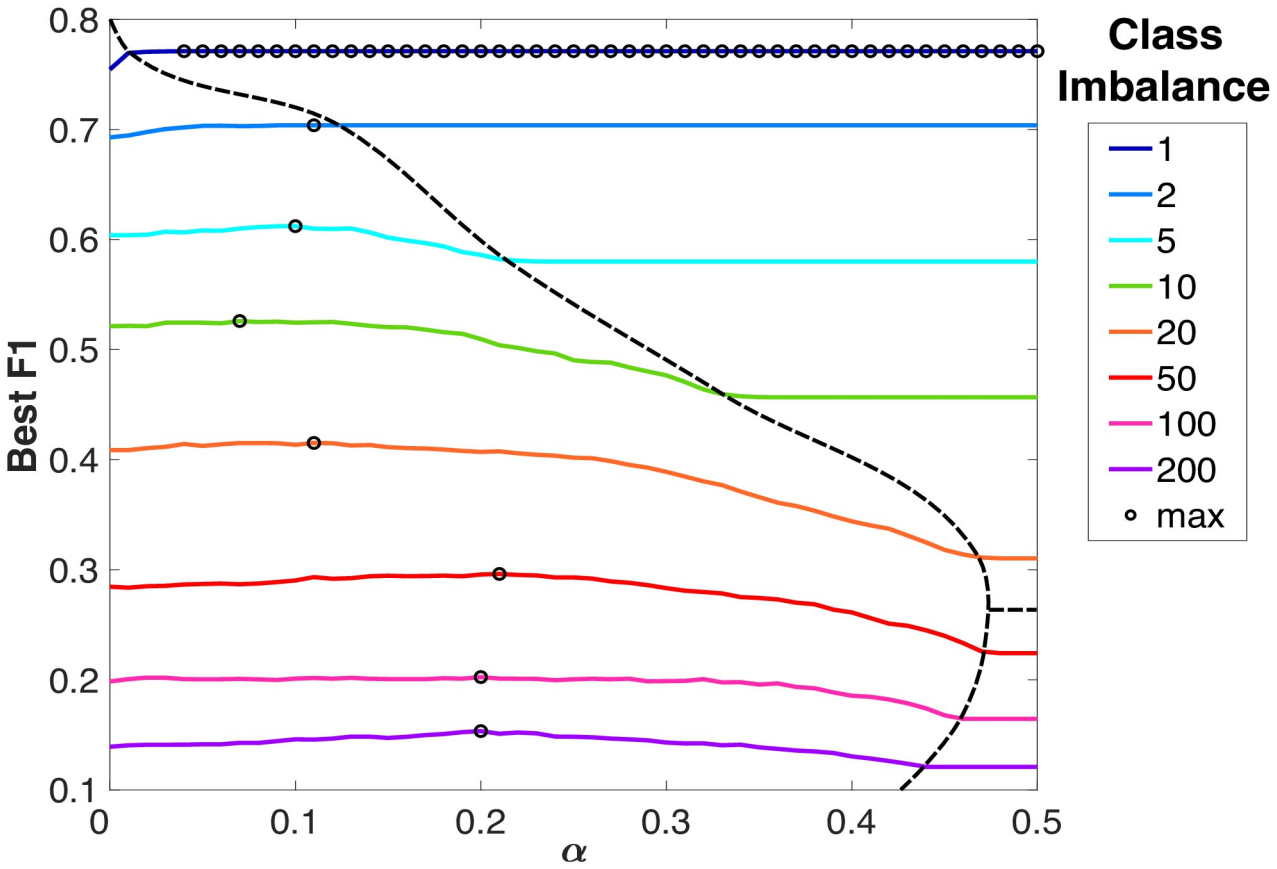

Fig 4. Behavior of Best F1 versus α for various class imbalance values in the Jaffe + sectors measure. Due to the correlation between Best F1 and class imbalance the various curves never intersect. In this plot, three phases divided by dotted lines can be spotted: a Low α phase, at the bottom-left, where there exists an α-dependent maximum for the Best F1, a High α—Low-class imbalance phase, at the top-right, and the High α—High-class imbalance phase at the bottom-right, where the Best F1 is independent of α.

https://doi.org/10.1371/journal.pone.0283217.g004

$S1J1$, and the threshold always falls between the elements of $S1J1$. Therefore, also in this phase, the Best F1 is independent of α.

From a practical point of view, this means that, when looking for the best α to optimise the prediction performances at fixed class imbalance, one should always search in the Low α phase. In fact, only in that phase performance metrics like Best F1 are α dependent and can reach a maximum.

Indirect measures

The idea of these measures is to compute firstly a measure of similarity between technology or of relatedness between companies and technologies, and secondly a measure of similarity between companies. Even if indirect, these approaches are discussed here because they use well-known tools of both mainstream and Economic Complexity literature.

Network-based approaches. The network-based approach to the construction of a relatedness measure between companies starts from the observation that our data constitute a bipartite network between companies and technology codes described by **M** matrices. The described bipartite network can be naturally projected onto both the company layer and the technology layer using the co-occurrences method. This gives us the possibility to construct different measures of relatedness.

The construction of a relatedness measure based on the projection of the bipartite network onto the technologies' layer follows the work of Pugliese et al. [26]. Firstly, we build a measure of relatedness between technologies, which is the weight of the link between technology codes in their monopartite network. We employ two kinds of normalizations for this measure.

The first is the one introduced in [23] and normalizes the co-occurrences of two technology codes in the same portfolio by the maximum of their ubiquity. We refer to this measure as *Technology Space*.

$$B_{tt'}^{TS} = \frac{1}{max(u_t, u_{t'})} \sum_f M_{ft} M_{ft'}.$$

where $u_t = \sum_f M_{ft}$ is called ubiquity. Normalizing by the maximum of the ubiquities allows us to weigh less the co-occurrences between highly ubiquitous technologies. Moreover, this normalization avoids undesirable effects caused by technologies patented by a single firm. In fact, if the technology t is patented only by the firm f, for every other technology code t', $\sum_{f'} M_{f't} M_{f't'} = 1$.

The other type of normalization we make use of in this work was first introduced by Teece et al. [17] in the context of firms' diversification; it has been later employed in several other studies on technological diversification [18–20]. We refer to this measure as *Micro-Partial* because the null model hardly constrains one layer and randomizes everything else [52]. To calculate the Micro-Partial measure we start from a matrix whose elements count the co-occurrences between technology codes within companies' portfolios $C_{tt'}$.

$$C_{tt'} = \sum_f M_{ft} M_{ft'}.$$

The $C_{tt'}$ matrix is then normalized with respect to a null model in which technologies are randomly assigned to companies' technological portfolios, keeping their ubiquity fixed. If we call u_t the ubiquity of the technology t and N the total number of companies, the random variable $x_{tt'}$, the number of companies innovating in technologies t and t' in the random case, follows a hypergeometric distribution with average and variance:

$$\mu_{tt'} = \frac{u_t u_{t'}}{N}, \qquad \sigma_{tt'}^2 = \mu_{tt'} \frac{(N - u_t)(N - u_{t'})}{N(N-1)}.$$

We therefore define:

$$B_{tt'}^{MP} = \frac{C_{tt'} - \mu_{tt'}}{\sigma_{tt'}},$$

which, in a similar fashion to a t-Student variable, measures the number of $\sigma_{tt'}$ the observed value of $C_{tt'}$ deviates from $\mu_{tt'}$. In other words, we are comparing the weight of the links in $C_{tt'}$ with the average values generated by a partial Microcanonical null model, in which only one of the two layers, the technologies' one, is fixed. Given these two measures of similarity between technology codes, we define the *coherence* [26] between a firm f and a technology t as:

$$\gamma_{ft} = \sum_{t'} M_{ft'} B_{tt'}.$$

For our scope, we need a similarity measure between companies. Therefore, to study the relatedness between the two firms involved in a M&A, we look at the mean coherence (Mean

γ) between the acquirer firm f and technologies of the target company f'

$$\bar{\gamma}_{f,f'} = \frac{1}{d_{f'}} \sum_{t \in f'} \gamma_{ft},$$

where $d_{f'}$ is the diversification of the target firm, i.e. the number of technologies in its portfolio.

Finally, as we mentioned in the Results section, the coherence γ_{ft} is correlated with the number of technologies f is linked to, i.e. the diversification d_f of the acquirer. To test if this correlation has some effect on the results of our forecast exercise, we compute a rescaled version of the coherence measure. We call $\hat{\gamma}_f$ the vector of all the values of coherence between the firm f and all the technologies (i.e. the f-th row of the matrix with elements γ_{ft}) and we rescale each of these vectors between 0 and 1:

$$\hat{\gamma}'_f = \frac{\hat{\gamma}_f - \min_t(\hat{\gamma}_f)}{\max_t(\hat{\gamma}_f) - \min_t(\hat{\gamma}_f)}.$$

In this way, the coherence metric γ'_f is less dependent on the diversification of the acquirer firm.

Random Forest. To compute the relatedness between companies and technologies, following [27], we use the Random Forest algorithm [56]. The Random Forest (RF) is a supervised machine learning algorithm based on decision trees. These are based on rules to divide the input space and learn to classify the input data according to subsequent splittings of this space. Although decision trees have several strengths, including high interpretability, little data preparation, and low computational cost, they suffer from high variance and overfitting. RF can correct these problems. As for the high variance, this indicates that if we take a tree and train it using a training set, the result of the prediction on a subsequent test will be very different from the same decision tree that has different training and a test (but still belonging to the same database). RF solves this problem by averaging the results of multiple decision trees. The overfitting problem is due to the adaptability of single trees to the training data. The RF solves this problem by performing each division of each branch of the tree using only a part of the complete input set.

In order to obtain a measure of the relatedness between a firm and a technology, we train one model (one RF) for each target technology t.

In general, for supervised machine learning algorithms we have to quantify three quantities:

- The matrix of samples **X**. The rows represent the different samples (the technological portfolios of companies); we have to identify correctly those that will become active in t. Each row is a company in a given year and each column represents a feature (a technology). In our case, **X** is the matrix obtained by concatenating vertically each matrix **M** with $y \in [2000, 2010]$ considering only companies with high diversification. In [27], the authors show that training the RF with firms with higher diversification increases the forecast results.
- The vector of classes **y**: in a generic classification problem it is a vector to which the class of the sample is associated. In our work, we associate the past technological portfolios of firms with the possible future activity in technology t, that is **y** = [0, 1]. The size of this array is equal to the number of companies multiplied by the number of years used in the training. So for each RF **y** is a column (the one corresponding to the technology t) of the matrix $\mathbf{M}^y$ shifted by two years with respect to the one used in the training, so $y \in [2002, 2012]$. We have binarized this matrix by setting the elements equal to 1 if that technology will be made by the corresponding firm after two years, and 0 otherwise i.e. we put 1 if the element of the matrix is different from 0, and if it is equal to 0, we leave it as it is.

We use **X** and **y** to train our Random Forest, i.e. to learn how the features of the samples (that is, the technologies of the companies) are associated with patenting/not patenting in the target technology after two years.

- The matrix of samples $\mathbf{X}_{\text{test}}$: after the training, we provide samples that are never seen by the RF in the training process, organized in a matrix $\mathbf{X}_{\text{test}}$. In our case, we use as $\mathbf{X}_{\text{test}}$ the matrix obtained by concatenating vertically each matrix with $y \in [2000, 2010]$ considering the companies about which we have either sector or M&A information. The companies present in $\mathbf{X}_{\text{test}}$ and which would have been among the ones used in the training have been removed from **X** to avoid overfitting problems.

As for the parameters, the number of trees, the min samples leaf (MSL), and the max depth (MD) have been investigated and optimized. The first represents the number of trees in the forest. This is a parameter that cannot be tuned in the classical sense but should be set high enough [67]; we set it equal to 50 to avoid higher computational time. MSL represents the minimum number of samples required to be at a leaf node; we set it equal to 4. MD is the maximum depth of the tree; we set it equal to 40. We choose these values after a grid search parameters test.

Finally, the output of the Random Forest is a matrix that contains the relatedness RF_{ft} between companies and technology codes, that is how much the RF finds the technology t close to the technological portfolio of firm f, where f corresponds to a row of $\mathbf{X}_{\text{test}}$. We can interpret this measure as another form of *coherence* between firms and technologies and use it in the same way. Therefore, to study the similarity between two firms involved in a M&A, we look at the mean values of RF_{ft}, i.e. $\bar{RF}_{f,f'}$ where f is the acquirer firm, and the average computed on all the technologies of the target firm f'

$$\bar{RF}_{f,f'} = \frac{1}{d_{f'}} \sum_{t \in f'} RF_{ft}. \tag{13}$$

As for the coherence, also in this case the RF_{ft} is correlated with the number of technologies f is linked to, and also in this case we rescale it by

$$\hat{RF}'_f = \frac{\hat{RF}_f - \min_t(\hat{RF}_f)}{\max_t(\hat{RF}_f) - \min_t(\hat{RF}_f)},$$

where $\hat{RF}_f$ is the f-th row of the matrix with elements RF_{ft}.

Continuous Company Space (CCS)

To obtain a two-dimensional representation of the proximity between companies, and therefore to obtain a greater interpretability of the results, we introduce the Continuous Company Space (CCS). Tacchella et al. have proposed in [24] the Continuous Projection Space by applying it to the exported products, and Straccamore et. al [27] have reformulated this concept by applying it to technologies.

The construction of our CCS starts from a matrix of distances $D_{ff'}$ between companies. In particular, we consider the subset of 8079 companies to which it was possible to assign a sector in our classification. We use two distance matrices. The first is derived from the Jaffe measure between companies. Since Jaffe's is a similarity measure, to obtain a distance we consider:

$$D^J_{ff'} = 1 - J_{ff'}.$$

The other distance measure is derived from the Jaffe + Sectors approach. In this case, we consider:

$$D^{JS}_{ff'} = \alpha(1 - S_{ff'}) + (1 - \alpha)D^{J}_{ff'}.$$

where $J_{ff'}$ is the Jaffe measure between firms' technological portfolios and $S_{ff'}$ is 1 if both firms belong to the same sector and 0 otherwise.

The columns of the distance matrix can be seen as the coordinates in a high-dimensional space for each company, with a dimension equal to 8079. Because it is impossible to visualize these coordinates in this such a high dimensional space, we project it in a 2*D* space we call CCS. Following [24], this operation consists of two steps. First, we reduce the number of dimensions from 8737 to 150 using a Variational—Autoencoder Neural Network [68]. Then, we reduce from 150 to 2 dimension using the t-SNE algorithm [57]. Within the 2*D* representation, the similarity between companies is simply given by the relative euclidean distance.

Prediction performance metric: Best-F1

To compare the algorithms and techniques used in this work we use as a performance metric the Best-F1 [24, 25, 28, 69]. The F1 score is defined as:

$$F1 = 2\left(\frac{1}{\text{precision}(\tau)} + \frac{1}{\text{recall}(\tau)}\right)^{-1} \tag{2}$$

i.e. is the harmonic mean between precision $= \frac{TP(\tau)}{TP(\tau)+FP(\tau)}$ and recall $= \frac{TP(\tau)}{TP(\tau)+FN(\tau)}$, where *TP* represents the number of True Positives (i.e. the elements equal to 1 that are correctly predicted) and, analogously, *FP* are the False Positives and *FN* the False Negatives. These quantities depend on the scores' binarization threshold τ, that is, the number above which the prediction score is associated with a predicted 1 (if the score is lower than τ, the measure predicts a zero). The Best-F1 is the metric associated with the value of τ that maximizes the F1 a posteriori. The highest possible value of Best-F1 is 1, which indicates that both precision and recall are equal to 1, and the lowest possible value is 0 if one of the precision or recall is zero. Since the Best F1 depends on the class imbalance of the problem, we show the forecast results with respect to a random prediction.

Supporting information

S1 File.
(PDF)

Acknowledgments

The authors thank Louis Barlascini for his preliminary investigations, Arianna Martinelli for providing the Zephyr data, Lorenzo Napolitano for providing the bipartite company-technology data, and Luciano Pietronero and Giambattista Albora for scientific discussions. The authors also acknowledge the CREF project "Complessità in Economia".

Author Contributions

Conceptualization: Lorenzo Arsini, Matteo Straccamore, Andrea Zaccaria.

Data curation: Lorenzo Arsini, Matteo Straccamore, Andrea Zaccaria.

Formal analysis: Lorenzo Arsini, Matteo Straccamore, Andrea Zaccaria.

Investigation: Lorenzo Arsini, Matteo Straccamore, Andrea Zaccaria.

Methodology: Lorenzo Arsini, Matteo Straccamore, Andrea Zaccaria.

Resources: Andrea Zaccaria.

Supervision: Matteo Straccamore, Andrea Zaccaria.

Writing – original draft: Lorenzo Arsini, Matteo Straccamore, Andrea Zaccaria.

Writing – review & editing: Lorenzo Arsini, Matteo Straccamore, Andrea Zaccaria.

References

1. Bruner RF, Perella JR. Applied mergers and acquisitions. vol. 173. John Wiley & Sons; 2004.
2. King DR, Dalton DR, Daily CM, Covin JG. Meta-analyses of post-acquisition performance: indications of unidentified moderators. Strategic Management Journal. 2004; 25(2):187–200. https://doi.org/10.1002/smj.371
3. Gomes E, Angwin DN, Weber Y, Yedidia Tarba S. Critical Success Factors through the Mergers and Acquisitions Process: Revealing Pre- and Post-M&A Connections for Improved Performance. Thunderbird International Business Review. 2013; 55(1):13–35. https://doi.org/10.1002/tie.21521
4. Ismail TH, Abdou AA, Annis RM. Review of literature linking corporate performance to mergers and acquisitions. The Review of Financial and Accounting Studies. 2011; 1(1):89–104.
5. Rossi M, Tarba S, Raviv A. Mergers and acquisitions in the hightech industry: A literature review. International Journal of Organizational Analysis. 2013; 21:66–82. https://doi.org/10.1108/19348831311322542
6. Hidalgo CA. Economic complexity theory and applications. Nature Reviews Physics. 2021; 3(2):92–113. https://doi.org/10.1038/s42254-020-00275-1
7. Balland PA, Broekel T, Diodato D, Giuliani E, Hausmann R, O'Clery N, et al. The new paradigm of economic complexity. Research Policy. 2022; 51(3):104450. https://doi.org/10.1016/j.respol.2021.104450 PMID: 35370320
8. Hidalgo CA, Hausmann R. The building blocks of economic complexity. Proceedings of the National Academy of Sciences. 2009; 106(26):10570–10575. https://doi.org/10.1073/pnas.0900943106 PMID: 19549871
9. Pietronero L, Cristelli M, Gabrielli A, Mazzilli D, Pugliese E, Tacchella A, et al. Economic Complexity: "Buttarla in caciara" vs a constructive approach. arXiv preprint arXiv:170905272. 2017;.
10. Hidalgo CA, Balland PA, Boschma R, Delgado M, Feldman M, Frenken K, et al. The principle of relatedness. In: International conference on complex systems. Springer; 2018. p. 451–457.
11. Hidalgo CA, Klinger B, Barabási AL, Hausmann R. The Product Space Conditions the Development of Nations. Science. 2007; 317(5837):482–487. https://doi.org/10.1126/science.1144581 PMID: 17656717
12. Zaccaria A, Cristelli M, Tacchella A, Pietronero L. How the taxonomy of products drives the economic development of countries. PloS one. 2014; 9(12):e113770. https://doi.org/10.1371/journal.pone.0113770 PMID: 25486526
13. Cefis E. The impact of M&A on technology sourcing strategies. Economics of Innovation and New Technology. 2010; 19(1):27–51. https://doi.org/10.1080/10438590903016385
14. Penrose ET. The Theory of the Growth of the Firm. Wiley; 1959.
15. Teece DJ. Towards an economic theory of the multiproduct firm. Journal of Economic Behavior & Organization. 1982; 3(1):39–63. https://doi.org/10.1016/0167-2681(82)90003-8
16. Knecht M. Diversification, Industry Dynamism, and Economic Performance: The Impact of Dynamic-related Diversification on the Multi-business Firm. Springer Fachmedien Wiesbaden; 2013.
17. Teece DJ, Rumelt R, Dosi G, Winter S. Understanding corporate coherence: Theory and evidence. Journal of Economic Behavior & Organization. 1994; 23(1):1–30. https://doi.org/10.1016/0167-2681(94)90094-9
18. Breschi S, Lissoni F, Malerba F. Knowledge-relatedness in firm technological diversification. Research Policy. 2003; 32(1):69–87. https://doi.org/10.1016/S0048-7333(02)00004-5
19. Piscitello L. Relatedness and coherence in technological and product diversification of the world's largest firms. Structural Change and Economic Dynamics. 2000; 11:295–315. https://doi.org/10.1016/S0954-349X(00)00019-9

20. Bottazzi G, Pirino D. Measuring Industry Relatedness and Corporate Coherence. Laboratory of Economics and Management (LEM), Sant'Anna School of Advanced Studies, Pisa, Italy; 2010.

21. Ernst H. Patent information for strategic technology management. World Patent Information. 2003; 25 (3):233–242. https://doi.org/10.1016/S0172-2190(03)00077-2

22. Strumsky D, Lobo J, van der Leeuw S. Using patent technology codes to study technological change. Economics of Innovation and New Technology. 2012; 21(3):267–286. https://doi.org/10.1080/10438599.2011.578709

23. Hausmann R, Klinger B. The structure of the product space and the evolution of comparative advantage. CID Working Paper Series. 2007;.

24. Tacchella A, Zaccaria A, Miccheli M, Pietronero L. Relatedness in the era of machine learning. arXiv preprint arXiv:210306017. 2021;.

25. Albora G, Pietronero L, Tacchella A, Zaccaria A. Product Progression: a machine learning approach to forecasting industrial upgrading. Scientific Reports. 2023; 13(1):1481. https://doi.org/10.1038/s41598-023-28179-x PMID: 36707529

26. Pugliese E, Napolitano L, Zaccaria A, Pietronero L. Coherent diversification in corporate technological portfolios. PLOS ONE. 2019; 14(10):1–22. https://doi.org/10.1371/journal.pone.0223403 PMID: 31600259

27. Straccamore M, Pietronero L, Zaccaria A. Which will be your firm's next technology? comparison between machine learning and network-based algorithms. Journal of Physics: Complexity. 2022; 3 (3):035002.

28. Albora G, Zaccaria A. Machine learning to assess relatedness: the advantage of using firm-level data. Complexity. 2022; 2022. https://doi.org/10.1155/2022/2095048

29. Liu H. Influencing Factors and Risk Control in Cross-Border Mergers and Acquisitions. In: 2022 International Conference on Economics, Smart Finance and Contemporary Trade (ESFCT 2022). Atlantis Press; 2022. p. 787–794.

30. Satapathy DP, Patjoshi PK. EFFECT OF MERGERS AND ACQUISITIONS ON EFFICIENCY OF INDIAN ACQUIRING BANKS: EVIDENCE FROM INDIA. Journal of Pharmaceutical Negative Results. 2022; p. 3434–3438.

31. Ding H, Fan H, Li C, Qiu LD. The effects of discriminatory protections on cross-border mergers and acquisitions. Journal of Comparative Economics. 2022;. https://doi.org/10.1016/j.jce.2022.11.003

32. Kooli C, Lock Son M. Impact of COVID-19 on mergers, acquisitions & corporate restructurings. Businesses. 2021; 1(2):102–114. https://doi.org/10.3390/businesses1020008

33. Ogendo JL, Ariemba J. Mergers and Acquisitions for Business Sustainability in Emerging Markets During a Vague Era: A Literature Analysis. AD-minister. 2022;(41):35–56. https://doi.org/10.17230/Ad-minister.41.2

34. Novita F, Rasyid R. Comparative Analysis of Financial Performance Before and After Mergers and Acquisitions. Financial Management Studies. 2022; 2(3):59–68.

35. Chiaramonte L, Dreassi A, Piserà S, Khan A. Mergers and Acquisitions in the Financial Industry: A bibliometric review and future research directions. Research in International Business and Finance. 2022; p. 101837.

36. Kaneko K, Kajikawa Y. Novelty Score and Technological Relatedness Measurement Using Patent Information in Mergers and Acquisitions: Case Study in the Japanese Electric Motor Industry. Global Journal of Flexible Systems Management. 2022; p. 1–15.

37. Cohen W, Levinthal D. Absorptive Capacity: A New Perspective on Learning and Innovation. Administrative Science Quarterly. 1990; 35:128–152. https://doi.org/10.2307/2393553

38. Lane P, Lubatkin M. Relative absorptive capacity and interorganizational learning. Strategic Management Journal. 1998; 19:461–477. https://doi.org/10.1002/(SICI)1097-0266(199805)19:5%3C461::AID-SMJ953%3E3.0.CO;2-L

39. Kaul A, Wu B. A Capabilities-Based Perspective on Target Selection in Acquisitions. Corporate Governance & Economics eJournal. 2015;.

40. Chakrabarti A, Mitchell W. The role of geographic distance in completing related acquisitions: Evidence from U.S. chemical manufacturers. Strategic Management Journal. 2016; 37(4):673–694. https://doi.org/10.1002/smj.2366

41. Bettinazzi ELM, Miller D, Amore M, Corbetta G. Ownership similarity in mergers and acquisitions target selection. Strategic Organization. 2020; 18:330–361. https://doi.org/10.1177/1476127018801294

42. Kennedy KH, Payne GT, Whitehead CJ. Matching Industries between Target and Acquirer in High-Tech Mergers and Acquisitions. Technology Analysis & Strategic Management. 2002; 14(2):149–162. https://doi.org/10.1080/09537320220133839

43. Cefis E, Rigamonti D. The importance of Industry Relatedness in M&A. University of Bergamo. 2013;.

44. Ahuja G, Katila R. Technological acquisitions and the innovation performance of acquiring firms: a longitudinal study. Strategic Management Journal. 2001; 22(3):197–220. https://doi.org/10.1002/smj.157

45. Cloodt M, Hagedoorn J, Kranenburg H. Mergers and Acquisitions: Their Effect on the Innovative Performance of Companies in High-Tech Industries. Research Policy. 2006; 35:642–654. https://doi.org/10.1016/j.respol.2006.02.007

46. Cassiman B, Colombo MG, Garrone P, Veugelers R. The impact of M&A on the R&D process: An empirical analysis of the role of technological- and market-relatedness. Research Policy. 2005; 34 (2):195–220. https://doi.org/10.1016/j.respol.2005.01.002

47. Hagedoorn J. Inter-Firm R&D Partnerships: An Overview of Major Trends and Patterns Since 1960. Research Policy. 2002; 31:477–492. https://doi.org/10.1016/S0048-7333(01)00120-2

48. Valentini G, Dawson A. Beyond knowledge bases: Towards a better understanding of the effects of M&A on technological performance. In: Advances in Mergers and Acquisitions. vol. 9. Emerald Group Publishing Limited; 2010. p. 177–197.

49. Jo GS, Park G, Kang J. Unravelling the link between technological M&A and innovation performance using the concept of relative absorptive capacity. Asian Journal of Technology Innovation. 2016; 24 (1):55–76.

50. Makri M, Hitt MA, Lane PJ. Complementary technologies, knowledge relatedness, and invention outcomes in high technology mergers and acquisitions. Strategic Management Journal. 2010; 31(6):602–628.

51. Orsi L, Ganzaroli A, Noni ID, Marelli F. Knowledge utilisation drivers in technological M&As. Technology Analysis & Strategic Management. 2015; 27(8):877–894.

52. Cimini G, Carra A, Didomenicantonio L, Zaccaria A. Meta-validation of bipartite network projections. Communications Physics. 2022; 5(1):1–12. https://doi.org/10.1038/s42005-022-00856-9

53. Wei CP, Jiang YS, Yang CS. Patent analysis for supporting merger and acquisition (m&a) prediction: A data mining approach. In: Workshop on E-Business. Springer; 2008. p. 187–200.

54. Futagami K, Fukazawa Y, Kapoor N, Kito T. Pairwise acquisition prediction with SHAP value interpretation. The Journal of Finance and Data Science. 2021; 7:22–44. https://doi.org/10.1016/j.jfds.2021.02.001

55. Jaffe AB. Technological Opportunity and Spillovers of R & D: Evidence from Firms' Patents, Profits, and Market Value. The American Economic Review. 1986; 76(5):984–1001.

56. Breiman L. Random forests. Machine learning. 2001; 45(1):5–32. https://doi.org/10.1023/A:1010933404324

57. Van der Maaten L, Hinton G. Visualizing data using t-SNE. Journal of machine learning research. 2008; 9(11).

58. Dice LR. Measures of the amount of ecologic association between species. Ecology. 1945; 26(3):297–302. https://doi.org/10.2307/1932409

59. Van Rijsbergen CJ. Foundation of evaluation. Journal of documentation. 1974; 30(4):365–373. https://doi.org/10.1108/eb026584

60. Han J, Kamber M, Pei J. 2—Getting to Know Your Data. In: Han J, Kamber M, Pei J, editors. Data Mining (Third Edition). third edition ed. The Morgan Kaufmann Series in Data Management Systems. Boston: Morgan Kaufmann; 2012. p. 39–82. Available from: https://www.sciencedirect.com/science/article/pii/B9780123814791000022.

61. Aggarwal C, Hinneburg A, Keim D. On the Surprising Behavior of Distance Metric in High-Dimensional Space. First publ in: Database theory, ICDT 200, 8th International Conference, London, UK, January 4—6, 2001 / Jan Van den Bussche (eds) Berlin: Springer, 2001, pp 420-434 (= Lecture notes in computer science; 1973). 2002;.

62. Brockmann D, Helbing D. The hidden geometry of complex, network-driven contagion phenomena. science. 2013; 342(6164):1337–1342. https://doi.org/10.1126/science.1245200 PMID: 24337289

63. Melo HP, Henriques J, Carvalho R, Verma T, Da Cruz JP, Araújo N. Heterogeneous impact of a lockdown on inter-municipality mobility. Physical Review Research. 2021; 3(1):013032. https://doi.org/10.1103/PhysRevResearch.3.013032

64. Ribeiro SP, Menghinello S, De Backer K. The OECD ORBIS database: Responding to the need for firm-level micro-data in the OECD. OECD. 2010;.

65. Pugliese E, Napolitano L, Zaccaria A, Pietronero L. Coherent diversification in corporate technological portfolios–Supplementary information. PLOS ONE. 2019;. https://doi.org/10.1371/journal.pone.0223403

66. Jeni LA, Cohn JF, De La Torre F. Facing imbalanced data–recommendations for the use of performance metrics. In: 2013 Humaine association conference on affective computing and intelligent interaction. IEEE; 2013. p. 245–251.

67. Probst P, Wright MN, Boulesteix AL. Hyperparameters and tuning strategies for random forest. Wiley Interdisciplinary Reviews: data mining and knowledge discovery. 2019; 9(3):e1301.

68. Kingma DP, Welling M. Auto-encoding variational bayes. arXiv preprint arXiv:13126114. 2013;.

69. Cruz R, Fernandes K, Cardoso JS, Costa JFP. Tackling class imbalance with ranking. In: 2016 International joint conference on neural networks (IJCNN). IEEE; 2016. p. 2182–2187.