Using Crunchbase for Research in Entrepreneurship: Data Content and Structure

Francesco Ferrati and Moreno Muffatto
School of Entrepreneurship (SCENT), Department of Industrial Engineering, University of Padova, Italy
francesco.ferrati@unipd.it
moreno.muffatto@unipd.it
DOI: 10.34190/ERM.20.120

Abstract: The large amount of business-related data available today allows researchers in entrepreneurship to explore new methodologies for data analysis. This paper aims to present an overview of the database provided by Crunchbase for research purposes. Founded in 2007, Crunchbase collects worldwide data on companies, investors, funding rounds and key people of the entrepreneurial ecosystem. As of May 2019, Crunchbase had collected records on 760,590 organizations (of which 708,558 companies), 121,509 investors of different types, 263,426 funding rounds, 890,429 people, 17,068 initial public offerings (IPO) and 89,959 acquisitions. The main purpose of this work is to give a detailed description of the Crunchbase database in order to highlight its potential and facilitate future researchers who intend to use this source of data. In order to achieve this goal, three main topics are covered. Since the database is provided in seventeen independent datasets, the linking logics have been reconstructed applying a reverse engineering approach. The relationships between the individual files have been identified and then summarized in an original diagram. For each dataset all the available variables are provided. Afterwards, in order to quantify the scope and coverage of the database, some key variables have been analysed, resulting in descriptive statistics for three areas of interest: companies, funding rounds and investors. Specifically, analysis is provided about the geographical distribution of companies, the number of companies per year of foundation and current operating status, the number of companies by amount and number of investments raised and as well as the number of investors by number and amount of investments made. Finally, some indications on the potential uses of Crunchbase for research in entrepreneurship are given. Considering the characteristics of the available variables we focused on the applications of machine learning algorithms for the analysis and modeling of equity investment processes.

Keywords: Crunchbase; startup, venture capital, investments, artificial intelligence, machine learning

1. Introduction

The increasing availability of business-related data as well as the readiness of accessible data mining frameworks allow researchers in entrepreneurship to investigate complex phenomena using innovative approaches. In this context, a very topical issue concerns the assessment of the probability of success of early-stage ventures. Over the last thirty years many studies have tried to codify how equity investors select the best entrepreneurial project to support (Ferrati and Muffatto, 2019) in what is commonly known as the picking winners problem. Given the increasing amount of available information on startups and venture capital funding, advanced data science techniques can now be applied to discover non-trivial, implicit and potentially useful patterns from large volumes of data. This field of research can be found at the intersection between two theoretical frameworks. From a methodological point of view, applying a strongly data-driven approach for the analysis of new ventures profiles, the research is part of the data mining and knowledge discovery strand (Peng et al., 2008). Considering instead the purpose of the research, i.e. the identification of the best companies from an investor's point of view, the lens model from social judgment theory can be taken into account (Brunswik 1956). In fact, this model was in turn considered by Zacharakis and Meyer (1999) to develop a framework for understanding an investor's decision-making process.

Given the complexity of the task, retrieving data that can be used to properly model an early stage company has a huge impact on the final results. In this regard, Crunchbase is an excellent source of information. Crunchbase is an online platform collecting and providing business data about private and public companies on a global scale. Originally built to track startups, the database includes descriptive data about companies, equity investors, funding rounds and people involved in the entrepreneurial ecosystem.

The platform is maintained by Crunchbase Inc. a company founded in July 2007 by Michael Arrington and located in San Francisco, California. The project came to life as a branch of the parent company TechCrunch Inc., a popular online publication started in 2005 by Archimedes Ventures (led by Michael Arrington and Keith Teare) and focused on startups and breaking tech news. From 2007 to 2015, TechCrunch maintained control of the

Crunchbase database using it as a place to track the companies mentioned in their articles. In September 2010, AOL acquired TechCrunch and Crunchbase as one of TechCrunch's portfolio companies for approximately $25 million. In 2015 Verizon acquired AOL for $4.4 billion and in the same year Crunchbase separated from AOL/Verizon to become an independent entity (AOL/Verizon, owner of TechCrunch, still retained a stake in the business). On September 22, 2015, in conjunction with the spin out, Crunchbase announced to be funded $6.5 million by Emergence Capital. On November 22, 2015 a follow-up round of $2 million was made by Salesforce Ventures, SV Angel, Felicis Ventures, Cowboy Ventures and 8VC. On April 6, 2017 a $18 million round was announced involving as investors Mayfield Fund, Felicis Ventures, Emergence, Cowboy Ventures and AOL. On October 31, 2019 a series C funding round was closed, raising $30 million from OMERS Ventures, Mayfield Fund, Emergence, Verizon Ventures, Cowboy Ventures and Felicis Ventures. As regard their product, Crunchbase launched "Crunchbase Pro" in 2016, "Crunchbase Enterprise" and "Crunchbase for Applications" in 2017 and "Crunchbase Marketplace" in 2018. With this last tool, Crunchbase integrated their platform with extra data sets from third-party companies to supplement their own information. Partner companies include, for example, Siftery, Apptopia, BuiltWith, SimilarWeb, IPqwery, Bombora, Owler and Aberdeen. Since its foundation, Crunchbase's popularity has gradually increased over time. Figure 1 represents the Crunchbase's website search frequency worldwide according to Google Trends.

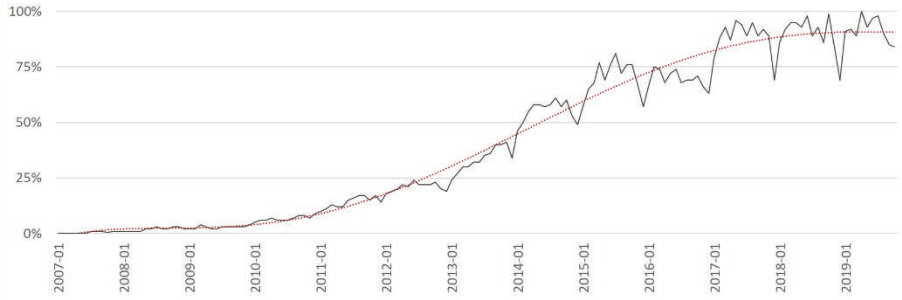

Figure 1: Frequency of Crunchbase website search worldwide. Source: Google Trends

Since Crunchbase is partially a crowd-sourced database, it is important to highlight the methodology with which the company collects and verifies the accuracy of data (Crunchbase, January, 7, 2020). Crunchbase sources, updates and validates their data on a daily basis, using four synergistic activities:

- The Crunchbase Venture Program: investors monthly submit portfolio updates in exchange for discounted access to the Crunchbase API, Excel export, Crunchbase Pro, and Crunchbase Marketplace. More than 3,500 global investment firms update their profile personally. This strategy allows Crunchbase to have access to the most up-to-date data.
- Active Community Contributors: active users can submit information to the database. The community makes the database grow and refine over time. It is important to specify that every submission is subject to registration, social validation, and is often reviewed by a moderator before being accepted and published.
- Artificial intelligence: in order to verify the reliability of data, Crunchbase apply machine learning algorithms to validate data accuracy, scan for anomalies, and alert their data scientists of data discrepancies.
- In-house data science team: Crunchbase data analysts manually validates the collected data. In addition, the team also develops the algorithms internally used and analyses the data to provide business insights, for example in the form of periodic reports.

The combined use of these four strategies of data collection and validation is an element of innovation and competitive advantage over other databases commonly used in the research field of entrepreneurship and economics. In fact, thanks to the quantity and quality of the data collected, Crunchbase is used not only by practitioners (e.g., entrepreneurs, investors or policy makers), but also by academic researchers who intend to apply a quantitative approach to research on entrepreneurship and innovation. However, it should be noted that since the dataset is partially created through a crowd-sourcing approach, data density cannot be guaranteed due to the voluntary nature of information recording. This is especially true for startups that have not collected investments yet and that therefore, not being part of the portfolio of any investor, are not reported by investors participating in the Crunchbase Venture Program. For this reason, to be effectively used by researchers, Crunchbase data requires an accurate pre-processing activity.

This paper aims to provide an overview of Crunchbase in order to highlight its potential and promote its use for research purposes. Specifically, Section 2 describes the methodology applied to connect the different parts of the database and make it possible to analyze it. Section 3 reports the organized content of the database, while Section 4 provides scope and coverage of the available data, giving descriptive statistics on some key information about companies, investors and funding rounds. Section 5 describes how Crunchbase has been used in research on entrepreneurship specifically focusing on the contributions that have applied machine learning techniques for data analysis. Finally, in Section 6 the conclusions are shown and some hints for the future use of Crunchbase in research are given.

2. Methodology

In order to assess the potential of Crunchbase as a source of data to be used for research in Entrepreneurship, a detailed analysis of its content and structure was carried out. The considered version of the database is dated May 21, 2019. Overall, the database is organized into seventeen .csv files. In order to figure out the size of the content, for each file the number of unique records have been detected. Referring to the key entities, the analyzed version of the database provides information on 760,590 organizations (of which 708,558 companies, 38,740 financial organizations and 13,292 schools and/or universities), 121,509 investors of different types (e.g., venture capital firms, angel investors, etc.), 263,426 funding rounds, 890,429 people, 1,346,357 jobs, 17,068 Initial Public Offerings (IPOs) and 89,959 acquisitions. As the use of Crunchbase by entrepreneurs and investors has grown over time and the database is now establishing itself as a primary source of information on startups and funding rounds, it can be expected that the number of companies, investors and people registered voluntarily will grow even more over time.

Since the Crunchbase corpus is organized in seventeen .csv files, in order to use the complete database, it is necessary to understand the way to connect the individual datasets to each other. Figure 2 represents how the different datasets (rectangles) are linked (arrows) with each other and allows a full reconstruction of the complete database starting from the seventeen independent .csv files. The diagram is not provided by other sources and we inferred it by applying a reverse engineering approach. Since the unique elements in each dataset are identified by a unique string (in the column “uuid”, universally unique identifier), the diagram has been designed by checking in different datasets the existence of columns with matching names and containing the same uuid. For example, every company within the “Organizations” dataset is identified by its own uuid. On the other hand, each investment round within the "Funding rounds" dataset is identified by its own uuid and reports the uuid of the company to which the round refers within the "company_uid" column. The correspondence between the two company’s uuid columns (in the “Organizations” and “Funding rounds” datasets) results in a relationship between the two datasets, represented by an arrow in the scheme.

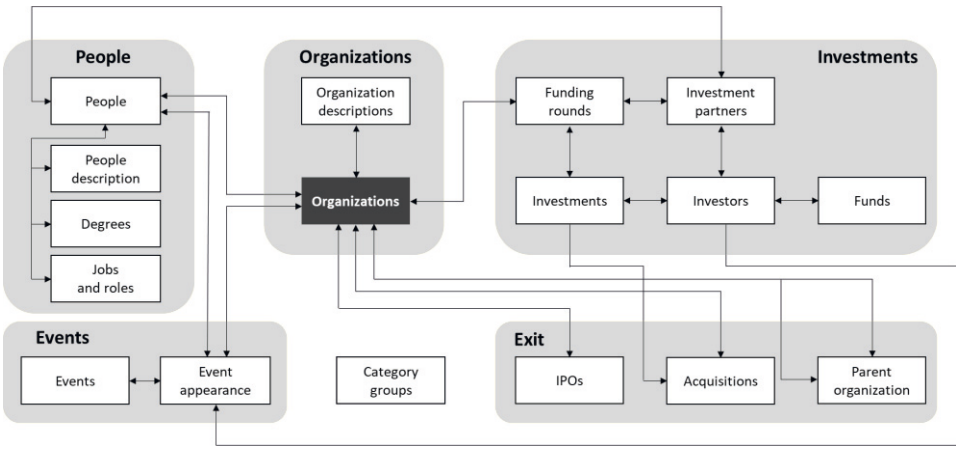

Figure 2: Scheme of the relationships between the different Crunchbase datasets

3. Data content

It can be immediately noticed that the information collected by Crunchbase covers five macro areas concerning respectively to organizations, investment activities, people (e.g., founders, investment partners, etc.), exits and public events. The following points describe the information covered in each area.

- **Organizations**. Crunchbase uses this term to refer to private or public companies, financial organizations as well as schools and universities. For each organization some overview information are given (e.g., legal name, foundation date, headquarter location, operating status, contact references, web and social media urls, etc.). Both a short and extended textual descriptions are available, allowing the use of advanced text analysis techniques to investigate the activities declared by each entity. Particularly interesting is also the information concerning the category which each organization belongs to. In fact, organizations are classified according to 680 unique categories (e.g., "automotive", "marketing", "machine learning") and 46 category groups (e.g., "health care", "commerce and shopping", "information technology"). The "category groups" dataset reports the rationale used to associate each category with one or more category groups.
- **Investments**. Using data in this area, it is possible to reconstruct the entire history of the investments collected by each company (or vice versa the entire history of investments made by each investor). In fact, the "Funding rounds" dataset links companies and investments to each other and represents one of the key elements of the database. It is important to underline that the term investment means the participation of a single investor in a specific round (one or more investors can therefore participate in a funding round by making an investment). Then using the "Investments" dataset the profiles of all the investors involved in a specific round can be identified and the lead investor is specified too. Combining all the investment-related datasets together, for each company it is possible to know the number of rounds collected, the type, date and amount of each round, as well as the number and the identity of the investors involved. On the other hand, for each investor some key information is given such as identification data (name of the firm or individual, location, year of foundation, social media url, etc.), the type of investor (e.g., venture capital, angel, accelerator, corporate venture capital, etc.), and the number as well as the total amount of investments made.
- **People**. Crunchbase provides also information about individuals involved in the entrepreneurial ecosystem, whether they are founders, employees, investors, advisors, etc. For each person, first name, last name, gender and location are provided, as well as the organization to which they belong to and a textual description of their profile. By linking the different datasets together, it is therefore possible to have complete information about the entrepreneurial team or the chief profiles within an organization. Particularly interesting are also the data concerning to the previous work experience of each person (e.g., the companies in which they used to work, the type of job, the title and the period of time when they used to work there). In addition, detailed information on the educational background is also available for each person (e.g., type of degree, subject and the institution where the degree was obtained).
- **Exits**. In analyzing a company, a key piece of information concerns the phase of the business life cycle in which they are in. The variable "status" in the "Organizations" dataset defines whether an organization is currently operating, closed, has already been acquired or went public thorough an IPO. Companies that have been acquired or listed on the stock market are said to have made an exit. For each exit, Crunchbase provides the date of the event and, in the case of an acquisition, the name of the acquirer. Although the dataset includes also other extremely valuable variables such as the price at which the company was acquired or the share value at the time of the IPO, such information is rarely available and cannot therefore be considered for analysis on large samples.
- **Public events**. Finally, Crunchbase also provides information on the appearances of organizations or individuals in public events. For each event, name, date and location are provided along with a detailed description. On the other hand, each appearance is qualified according to the role held by the participant.

Table 1 in Appendix summarizes the full content of the seventeen datasets, providing the name of each .csv file, the number of records as of May 21, 2019 and the list of attributes of each table. Each dataset contains variables of different types (e.g., continuous, ordinal, categorical, etc.) and depending on the considered research hypothesis various combinations of independent, dependent, controlled and moderator variables can be considered.

4. Scope and coverage

Since Crunchbase data is mainly entered on a voluntary basis, the scope of the database is not strictly defined and the coverage may vary across Countries, year of foundation and industry. As the database was created in May 2007, records prior to this date may be incomplete, having been added retrospectively. However, since for each record the date of creation as well as the date of the last update are reported, it is possible to verify a significant increase in the registration rate over time especially in recent years. The record creation rate was

rather low until the beginning of 2013, when the number stabilized on average at about 200 new records per day. The months of August 2013 and April 2014 were an exception, reporting the addition of several thousand records that provided a boost to the coverage of Crunchbase. Looking at the information on investment rounds, the database offers good coverage from around 2001, although in general the registered data also goes far back in time (Breschi, Lassébie and Menon, 2018). Compared to other commonly-used startup databases covering similar information and frequently used for economic research (e.g., ThomsonOne, formerly known as "VentureXpert" and VentureSource), Crunchbase is not only focused on venture-backed companies, but covers both companies that have been funded and that have not been yet. Actually, venture-backed startups represent just a small part of the Crunchbase corpus. All the registered companies therefore represent a more representative sample of the entrepreneurial ecosystem. Moreover, aggregate statistics on funding rounds by Country and year are quite similar to those produced with other established sources, going to validate the use of Crunchbase as a reliable source in term of coverage of funded ventures (Dalle, den Besten and Menon, 2017). For instance, Crunchbase covers about the same number of investment rounds in the analogous sectors as collected by the National Venture Capital Association (Block and Sandner, 2009). In order to understand the coverage of the dataset, some statistics are given below for three key areas of interest: companies, funding rounds and investors. The following evaluations were carried out on a Crunchbase extraction made on May 21, 2019.

Crunchbase provides the profile of 708,558 companies. The variable “country_code” allows to identify the Country in which each company is based. In order to assess the geographical distribution, the number of companies in each Country has been counted and the result is shown in Figure 3. Totally, companies are located in 206 different Countries and Crunchbase does indeed have global coverage. The top 10 Countries by number of companies are: United States (34.75%), United Kingdom (6.16%), India (3.63%), Canada (2.91%), Germany (1.92%), France (1.79%), Austria (1.60%), China (1.48%), Brazil (1.41%) and Spain (1.30%). These ten Countries cover 57% of all registered companies, and the remaining 43% is spread over the 196 remaining locations. This analysis shows that the database has greater coverage in the United States and within these, 53% of companies are located in five states: California (25.63%), New York (11.08%), Texas (6.59%), Florida (5.11%) and Massachusetts (4.70%). The geographical coverage of the database allows researchers to carry out different types of analysis, for example focusing on a single Country, or comparing the dynamics of two or more different areas. However, when using data, it is important to consider the variances between different entrepreneurial ecosystems. Including all the companies in a single population, without taking into account the specific dynamics to each ecosystem, could in fact return misleading results.

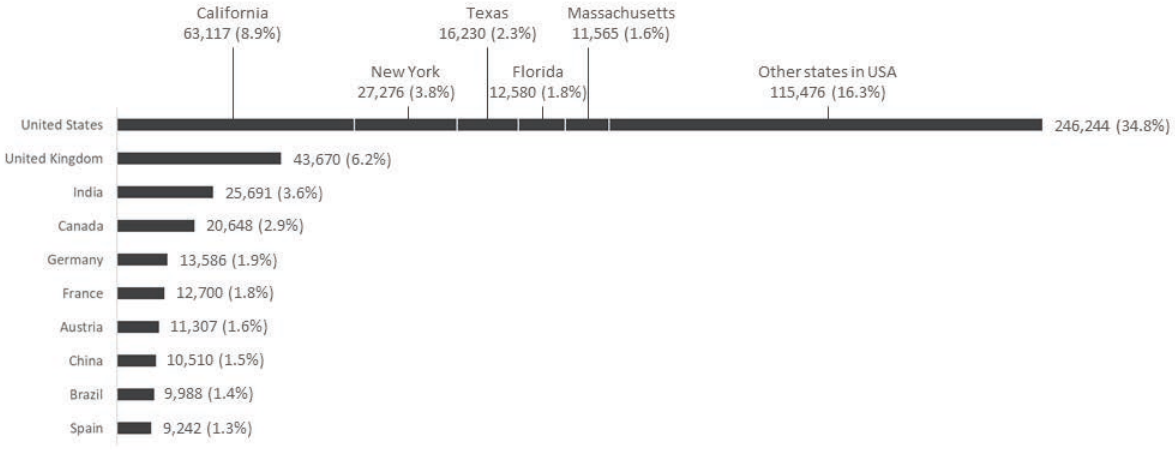

Figure 3: Top 10 Countries by number of companies

Another important element to take into account is the year of foundation of the companies, reported in the variable "founded_on". In fact, Crunchbase not only provides profiles of startups but also of large companies with a long-time history. Figure 4 reports the distribution of companies per year of foundation, starting from 2000. Among the 708,558 companies, 390,821 (55.2%) have been founded since 2000, while 114,689 (16.2%) have been founded before and 203,048 (28.6%) don’t declare their foundation date. The graph shows that the number of companies per year of foundation has an increasing trend until 2015, while thereafter the number is decreasing. This trend must be considered with caution when analyzing the data. In fact, since the Crunchbase data does not constitute an actual census of existing companies but is entered voluntarily by founders or

investors, it is likely that a company will register on the platform when it starts looking for investment, or has become part of the portfolio of an investor, or more generally wishes to gain greater visibility online. There is therefore a certain delay between the foundation of the company and its actual registration on Crunchbase. This phenomenon explains the trend described. Figure 4 also shows the current status declared by the companies, provided by the variable "status". This categorical variable can take four distinct values: *operating*, *acquired*, *ipo* or *closed* and is available for all the 708,558 companies. Specifically, 580,671 (82.0%) companies are still private and operating, 83,345 (11.8%) have been acquired by a bigger organization, 13,171 (1.9%) went public through an Initial Public Offering (IPO) and 31,371 (4.4%) closed. In this respect, it is necessary to treat data with caution. In fact, considering the high failure rate of startups, it is unlikely that only 4.4% of the registered companies have closed. On the contrary, it is more likely that failed companies decide to delete their profile from the database, thus being excluded from the calculation.

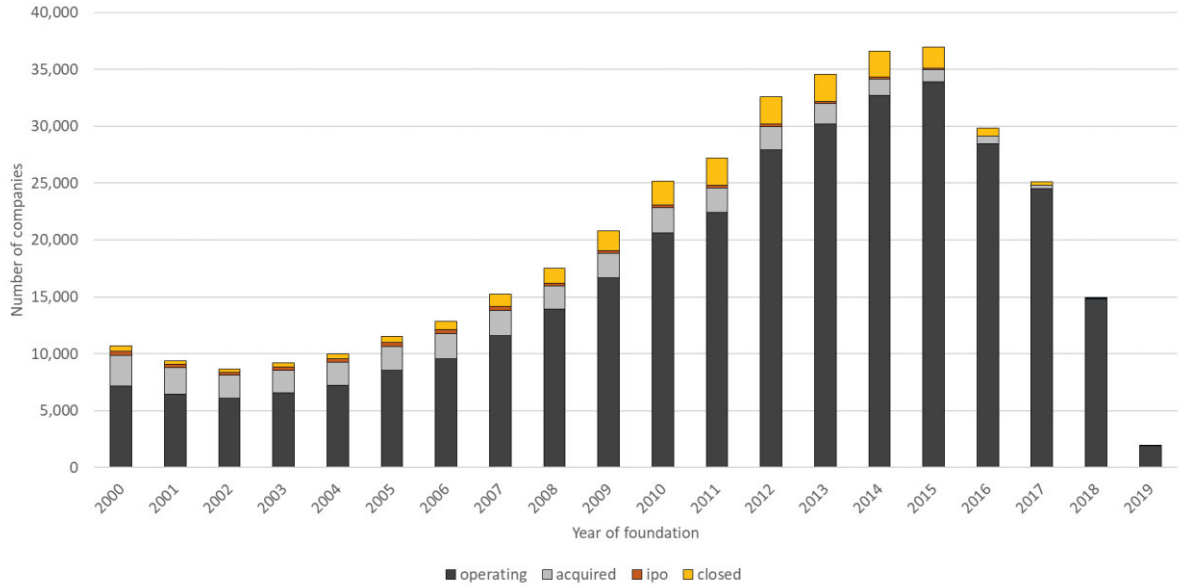

Figure 4: Number of companies per year of foundation VS current status

As Crunchbase provides detailed information on the funding history of each venture, a key analysis to understand the content of the database concerns the amount and total number of rounds collected by companies. First of all, it is important to point out that of the 708,558 companies, 566,079 (79.9%) do not declare any funding round and 38,939 (5.5%) have at least one round but do not specify the total funding amount. On the other hand, Figure 5 show the distribution of the 103,540 (14.6%) companies both having at least one round and declaring their total funding amount. Among these, 37.7% have collected less than $1 million, 26.3% have raised between $1 million and $5 million, 27.7% between $5 million and $50 million and 8.3% have collected more than $50 million. As for the number of rounds collected by the companies in Figure 5, the majority (52.6%) collected only one round of investment. It is therefore important to consider these trends in the light of the analyses to be carried out. The value of the investments must then be considered according to the Country in which the company is based. In fact, while the investments reported by Crunchbase mainly refer to seed and series A rounds, the average values of the rounds of the same stage vary between the different entrepreneurial ecosystems.

Considering investors, Crunchbase provides information on 121,509 professionals of different types (e.g. angel investors, business angel groups, venture capital firms, private equity firms, corporate venture capital branches, accelerators, incubators, entrepreneurship programs, university programs, startup competitions, investment banks, etc.). However, for just 2,388 investors (1.97%) both the information about the total number of investments made and the total amount invested is available. Within this subset, 1,161 (48.6%) have made only one investment, and 859 (36.0%) have invested between one and ten times. **Figure 6** shows the number of investors per total amount invested and total number of investments made. Almost 10% of them have totally invested less than $1 million, 16.3% between $1 million and $5 million, and another 16.0% between $5 million and $15 million. It is also interesting to note that 5.9% of the selected investors have invested more than $1 billion. The database therefore contains information on investors who are very different to each other, both in

terms of frequency in their activity and size of the managed funds. Also in this case, researchers should consider how to manage these differences in order to make consistent analysis or models.

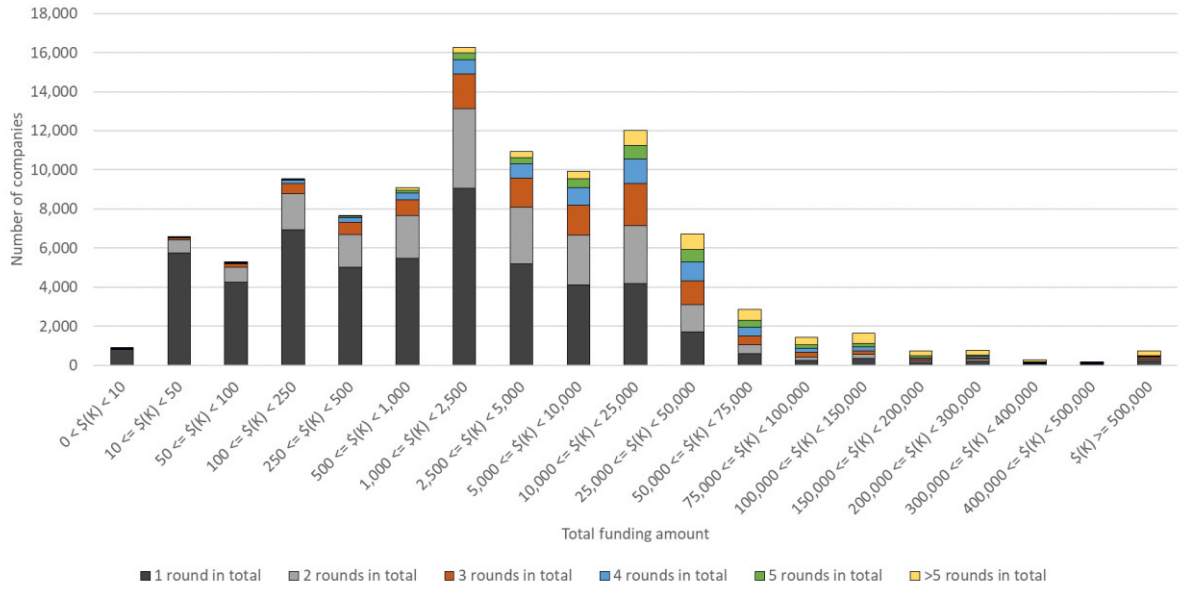

Figure 5: Number of companies per total funding amount VS total number of rounds

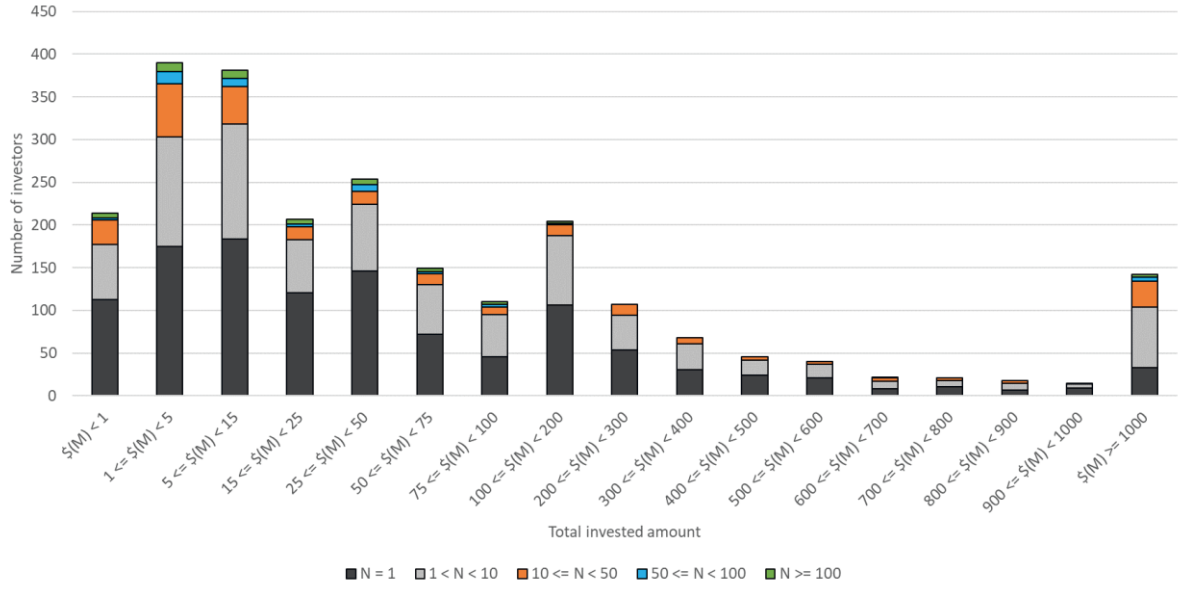

Figure 6: Number of investors per total amount invested VS total number of investments made

5. Using Crunchbase for academic research

From an academic point of view, many researchers have begun to explore the potential of Crunchbase, generally to do research in the areas of entrepreneurship, innovation and finance. By January 20, 2020 forty-seven works related to Crunchbase have been published on Scopus, forty-one of them since 2016. In addition to indexed publications, other contribution can be identified and analyzed in order to fully understand the state of the art in the use of Crunchbase (Dalle, Den Besten and Menon, 2017). Considering the large amount of structured data available, some researchers have recently started to analyze Crunchbase data using machine learning algorithms. In this regard, we identified five different research areas.

The most considered objective concerns the prediction of a company's exit event (Xiang et al., 2012, Ünal, 2019). In fact, it is commonly accepted that the critical milestone that classify a venture-backed company as financially successful is the so-called exit event. A venture-backed company can make an exit through two main strategies as it can either make an Initial Public Offering (IPO) or can be acquired by a larger company through merger and

acquisition (M&A). Crunchbase provides information about the status of an organization through a dedicated categorical variable in the "organizations" dataset. The variable can assume four different values: operating, closed, acquired or IPO. Because of the direct availability of this information, the value of the status variable can be used as a target variable in a machine learning classification problem.

Another problem addressed using a machine learning approach is the prediction of the next funding event within a certain period of time (Sharchilev et al., 2018). This kind of problem is generally formulated as follows: for a given company that has already secured at least one funding round (e.g., seed, angel, etc.), predict whether it will raise a further round of investment (i.e., Series A, B, etc.) during a given period of time (e.g., one year). Series A rounds are generally the first round collected from VCs and they usually let angel investors to exit the company and eventually make a positive return on their previous investment. For this reason, series A rounds are considered as an important milestone in the lifecycle of a startup company and having insights about their occurrence in the future can have an impact on the funding decision of early-stage investors.

Another research theme concerns the possibility of predicting whether a specific investor will invest in a specific company (Liang et al., 2016). In the context of this problem, Crunchbase still represents a valuable data source. In fact, for each funding round the information of all the investors involved is given. Companies and investors can therefore be represented as nodes in a graph data structure, where investment relationships are modeled as edges in a bipartite network (i.e., a network whose nodes are divided into two sets, and where the only connections allowed are those between nodes in different sets).

In order to support equity investors in their investment decision making process, some studies have also developed recommender systems to learn the specific investment preferences of a venture capital firm in order to identify the most suitable company to suggest in the screening activity (Zhong, 2019).

Finally, another area of research concerns the implementation of machine learning models for a company's industry classification (Batista and Carvalho, 2015). An accurate company's classification by sector has a key role in many applications, such as identifying similar companies, matching investors with companies in their specific sector of interest, or providing valuable features to support machine learning algorithms in the pattern recognition task.

6. Conclusion and opportunities for future researchers

From an academic point of view, the Crunchbase database represents an important asset for research in entrepreneurship. In fact, the large amount collected information allows researchers to analyze phenomena not investigated so far due to lack of sufficient data. In order to facilitate future researchers, a detailed description of the database was presented, providing information about its content, structure, scope and coverage. In order to use Crunchbase effectively, researchers should first analyze the content of the database in detail. In fact, in accordance with the typical data science workflow, the use of Crunchbase requires careful pre-processing and data cleaning activities. For example, attention should be paid to the preliminary analysis of missing data in the database. In fact, being built on a voluntary basis by companies and investors, some information is not always provided. Another important step concerns the identification of outliers especially regarding both the amount and the number of investments collected by different companies. The location of the considered companies should also be taken into account at a preliminary stage, in view of the specific dynamics of different entrepreneurial ecosystems. Having the confidence to fully understand the data provided, Crunchbase offers researchers the opportunity to explore new methodological approaches in entrepreneurship. As described in the last section the database can be successfully used in machine learning for the analysis of startups, equity investments different types of investors.

For future research, it would be interesting to further integrate the information provided by Crunchbase with that collected from other sources (e.g. social networks, intellectual property databases, socio-economic characteristics of different geographical locations, etc.). Although this activity has already been partially explored by some authors, there is still plenty of room for future research. The integration of different types of data, together with the use of advanced data mining techniques, could in fact provide new elements to better understand the key elements of successful companies.

Appendix 1

Table 1: Crunchbase corpus (as of May 21, 2019)

Dataset name (.csv file)	# records	List of dataset variables
Organizations	760,590	uuid, company_name, type, primary_role, roles, short_description, category_list, category_group_list, founded_on, country_code, state_code, region, city, address, status, closed_on, employee_count, funding_rounds, funding_total_usd, last_funding_on, permalink, cb_url, aliases, domain, homepage_url, email, phone, linkedin_url, twitter_url, facebook_url, logo_url, created_at, updated_at
Organization descriptions	535,355	uuid, description
Category groups	680	uuid, category_name, category_group_list
Funding rounds	263,426	funding_round_uuid, company_uuid, company_name, investor_uuids, investor_names, investor_count, investment_type, announced_on, raised_amount_usd, raised_amount, raised_amount_currency_code, post_money_valuation_usd, post_money_valuation, post_money_currency_code, country_code, state_code, region, city, cb_url, created_at, updated_at
Investors	121,509	uuid, investor_name, roles, investor_type, founded_on, closed_on, investment_count, total_funding_usd, country_code, state_code, region, city, cb_url, domain, twitter_url, facebook_url, logo_url, updated_at
Investments	400,432	funding_round_uuid, investor_uuid, is_lead_investor
Investment partners	73,050	funding_round_uuid, investor_uuid, partner_uuid
Funds	11,658	fund_uuid, fund_name, entity_uuid, announced_on, raised_amount, raised_amount_currency_code, created_at, updated_at
People	890,429	uuid, first_name, last_name, gender, country_code, state_code, city, primary_organization_uuid, primary_affiliation_organization, primary_affiliation_title, cb_url, linkedin_url, twitter_url, facebook_url, logo_url, created_at, updated_at
People descriptions	475,278	uuid, description
Jobs	1,346,357	job_uuid, person_uuid, org_uuid, job_type, title, started_on, ended_on, is_current
Degrees	335,414	degree_uuid, person_uuid, institution_uuid, degree_type, subject, started_on, completed_on, is_completed, created_at, updated_at
Acquisitions	89,959	acquisition_uuid, acquisition_type, acquiree_uuid, acquiree_name, acquiree_country_code, state_code, acquiree_region, acquiree_city, acquiree_cb_url, acquirer_uuid, acquirer_name, acquirer_country_code, acquirer_state_code, acquirer_region, acquirer_city, acquirer_cb_url, acquired_on, price_usd, price, price_currency_code, created_at, updated_at
Ipos	17,068	ipo_uuid, company_uuid, name, country_code, company_state_code, region, city, went_public_on, money_raised_usd, stock_exchange_symbol, stock_symbol, price_usd, price, price_currency_code, cb_url, created_at, updated_at
Organization parents	13,593	uuid, parent_uuid, created_at, updated_at
Event appearances	320,320	event_uuid, participant_uuid, appearance_type, participant_type, created_at, updated_at
Events	17,744	uuid, name, event_roles, short_description, description, started_on, ended_on, location_uuid, venue_name, continent, country_code, region, city, permalink, cb_url, registration_url, logo_url, created_at, updated_at

Acknowledgements

This research was made possible thanks to the support of Crunchbase Inc. http://www.crunchbase.com

References

Batista, F., & Carvalho, J. P. (2015). Text based classification of companies in CrunchBase. In 2015 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE) (pp. 1-7).

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

Breschi, S., J. Lassébie and C. Menon (2018). A portrait of innovative start-ups across countries. OECD Science, Technology and Industry Working Papers, 2018/02, OECD Publishing, Paris.

Brunswik E (1956) Perception and the representative design of psychological experiments, University of California Press.

Crunchbase (January, 7, 2020). Where does Crunchbase get their data? Accessed January, 20, 2020 https://support.Crunchbase.com/hc/en-us/articles/360009616013

Dalle, J. M., Den Besten, M., & Menon, C. (2017). Using Crunchbase for economic and managerial research. OECD Science, Technology and Industry Working Papers, 2017/08, OECD Publishing, Paris.

Ferrati, F., & Muffatto, M. (2019). A Systematic Literature Review of the Assessment Criteria Applied by Equity Investors. 14th European Conference on Innovation and Entrepreneurship, (p. 304-312). Kalamata, Greece.

Liang, Y. E., & Yuan, S. T. D. (2016). Predicting investor funding behavior using Crunchbase social network features. Internet Research, 26(1), 74-100.

Peng, Y., Kou, G., Shi, Y., & Chen, Z. (2008). A descriptive framework for the field of data mining and knowledge discovery. International Journal of Information Technology & Decision Making, 7(04), 639-682.

Sharchilev, B., Roizner, M., Rumyantsev, A., Ozornin, D., Serdyukov, P., & de Rijke, M. (2018). Web-based startup success prediction. In Proceedings of the 27th ACM International Conference on Information and Knowledge Management (pp. 2283-2291).

Ünal, C. (2019). Searching for a Unicorn: A Machine Learning Approach Towards Startup Success Prediction (Master's thesis, Humboldt-Universität zu Berlin).

Xiang, G., Zheng, Z., Wen, M., Hong, J., Rose, C., & Liu, C. (2012). A supervised approach to predict company acquisition with factual and topic features using profiles and news articles on TechCrunch. In Sixth International AAAI Conference on Weblogs and Social Media.

Zacharakis AL, Meyer GD (1999). A lack of insight: Do venture capitalists really understand their own decision process?. The Journal of Private Equity, 56-71.

Zhong, H. (2019). Venture capital investment: from rule of thumb to data science (Doctoral thesis, Rutgers University-Graduate School-Newark).

Francesco Ferrati. School of Entrepreneurship (SCENT), Department of Industrial Engineering, University of Padova, Padova, Italy. PhD student in Management Engineering at the University of Padova, Italy. His research activity regards the identification of specific relationships between the attributes of technology-driven startups and the investments raised along the business life-cycle.

António José Fernandes is an Associate Professor and Head of Social and Exacta Sciences Department from Agriculture School, Institute Polytechnic of Bragança, Portugal. Since 2006, he is a PhD in Management from the University of Trás-os-Montes and Alto Douro

Olaf Flak Associate Professor at the University of Silesia (since 2010), Assistant Professor at University of Economics in Katowice (2002-2012). Scientist and a specialist in business management, Managing Director in a consulting company konsultanci24.pl. He investigates how automatic pattern recognition techniques can be applied in the management science in order to replace managers with robots – TransistorsHead.com.

Esteban Galán is a lecturer and researcher of the group "Communication, Art and Digital Culture" (ARTICOM) in Universitat Politècnica of Valencia (Spain). He has worked as an audiovisual storyteller and he has more than 15 years of experience as producer in different broadcasters. He has articles and international conferences about transmedia communication.

Tadeusz A. Grzeszczyk is an associate professor in Faculty of Management at Warsaw University of Technology and conducts scientific and didactic activity regarding project management and evaluation (over 100 publications in management and social sciences). His interests and research work also include methodology of management sciences and the use of AI methods in decision support.

Fabian Hecklau, M. Sc., studied industrial engineering at the Otto-von-Guericke University Magdeburg and started working in applied research at Fraunhofer IFF in Magdeburg. Since 2015, he works for Fraunhofer IPK in Berlin and is involved in international research and consulting projects in the field of strategic management of organizations and innovation institutions. He is the head of the Competence Center Innovation Systems & Structures at Fraunhofer IPK since 2020.

Shital Jayantilal - Head of School of Management and Economics, and assistant professor, of Universidade Portucalense, Porto, Portugal; Coordinator of the research group Strategy & Competitiveness in REMIT (Research on Economics, Management, and Information Technologies) investigation centre.

Andrew Jenkins is Principal Lecturer at Huddersfield Business School. His areas of research interest include HRM, work & employment, older workers, hospitality, tourism and logistics. Andrew is the Subject Leader for Marketing, Events, Hospitality and Tourism, the Module Leader for Research Methods and the former Chair of Huddersfield Business School Research Ethics Committee.

Carina Jordão is a researcher at the Department of Social, Political and Territorial Sciences of the University of Aveiro, Portugal. Her current interests focus on gender in/equality in research and higher education. She has also been studying the phenomenon of in/equality between women and men in the labour market, especially in European Union countries.

Anna Kimberley A senior lecturer and a Ph.D. candidate. Lives and works in Finland, and pursues her doctorate studies at the University of Westminster, London, UK. Interested in cultural diversity, diversity management, communication management. Her research areas are: Interpretative Phenomenological Analysis, Narrative Analysis, identity and sensemaking.

Olga Koropets is an Associate Professor of the Personnel Management and Psychology Department at the Ural Federal University. Her main scientific interests cover psychological aspects of personnel management, labour and organizational psychology. She has published more than 60 papers and 4 monographs.