OECD Science, Technology and Industry Working Papers
2017/08

Using Crunchbase for economic and managerial research

Jean-Michel Dalle,
Matthijs den Besten,
Carlo Menon

https://dx.doi.org/10.1787/6c418d60-en

OECD

OECD SCIENCE, TECHNOLOGY AND INDUSTRY WORKING PAPERS

The release of this working paper has been authorised by Andrew Wyckoff, OECD Director for Science, Technology and Innovation and by Pierre Poret, OECD Director for Financial and Enterprise Affairs Directorate.

Using Crunchbase for Economic and Managerial Research

Jean-Michel Dalle*

Matthijs den Besten†

Carlo Menon‡

October 31, 2017

Abstract

This note describes a new database on innovative start-ups and companies, called Crunchbase, with a focus on its potential for economic and managerial research. Crunchbase is rapidly being discovered by scholars from different fields. It has notably already informed studies on specific sectors as well as studies of networks in the start-up ecosystem. This note first describes the contents of Crunchbase and then reviews academic research that has used it. We further suggest that many more valuable avenues for economic and managerial research can be opened through the combination of Crunchbase with selected supplementary data sources and provide two such examples.

JEL Codes: L26; M13; O31

Keywords: start-ups, venture capital, patents

*Université Pierre et Marie Curie

†Montpellier Research in Management, Montpellier Business School

‡OECD, Science, Technology and Innovation Directorate

Contents

1. INTRODUCTION

Crunchbase is a new commercial database on innovative companies maintained by Crunchbase Inc., an innovative start-up in itself, located in California, US. The database was created in 2007 but its scope and coverage has increased significantly over the past few years. As reported by Kauffman Foundation, the database is increasingly used by the venture capital industry as a "the premier data asset on the tech/startup world".[1]

The database is also becoming increasing popular with scholars and researchers, particularly as a source of information on start-up activity and financing within and across countries. As documented in this paper, more than 90 scientific contributions based on its data have been made available so far.

Given the novelty and the originality of the database, many scholars are interested in understanding the potential of this innovative database for economic and managerial research. This note contributes to fulfill this need, by providing an overview of the scope and coverage of the database; reviewing the contributions that have used it; and illustrating some possible extensions of the database by linking it with supplementary data sources.

Compared to commercial databases covering similar information and frequently used for economic research, Crunchbase is free of access for academic research (conditional on applying for a license and on complying with the terms of use); partially crowd-sourced, i.e., users can add and revise contents, which add to the comprehensiveness and timeliness of the database; is updated on a daily basis; contains cross-linked information on companies, their funders, and their staff; is structured in an accessible way; and does not require a huge amount of data handling before it can be used for econometric analysis.

Similarly to other commercial databases like e.g. ORBIS maintained by Bureau Van Dijk, which are not created with the particular needs of statisticians and economists in mind, the coverage of Crunchbase is not clearly defined and its scope may vary across countries and sectors. This is an issue requiring careful examination by researchers. However, aggregate statistics on VC funding by country and year tend to be reasonably similar to the same figures produced with an alternative and more established source, which is reassuring in term of coverage of funded ventures.

The rest of this note is organized as follows: First, we describe the contents of Crunchbase; next, we review academic research that has used this dataset; and finally, we stress the linkability of Crunchbase with other data sources and provide two examples.

2. SCALE AND SCOPE OF THE CRUNCHBASE DATABASE

2.1 Scope and coverage

Crunchbase [2] provides a depth and breadth of knowledge that ensures its data is recognized as the primary source of business information by over 31 million users globally.".[3] The data are sourced through two main channels: a large investor network and community contributors. As of September 2017, more than 3 000 global investment firms submit monthly portfolio updates to Crunchbase, in exchange for free data access. In addition, around 500 000 executives, entrepreneurs, and investors contributes to update and revise Crunchbase company profile pages. This wealth of data is processed with artificial intelligence (AI) and machine learning algorithms in order to ensure accuracy and scan for anomalies. Additionally, algorithms continuously search the web and thousands of news publications for information to enrich profiles.[4] Given that Crunchbase is quickly becoming a primary data source for investors, it is plausible to assume that dynamic and ambitious entrepreneurs have a strong incentive to register in the website and to keep their information updated. In addition, the crowd-sourcing process, the partnership with investment firms, and the validation with AI and machine-learning algorithms represent important innovations compared to other commercial databases and public data sources commonly used in economic research, which may provide unprecedented opportunities to analyse phenomena that have been under investigated so far because of lack of suitable data.

In the version used for this note, downloaded in January 2017, the database contains information on more than 490 000 distinct companies located in 199 different countries. Of those, around 220 000 reports a founding year later than 2010, and around 363 000 later than 2005 (Figure 1).

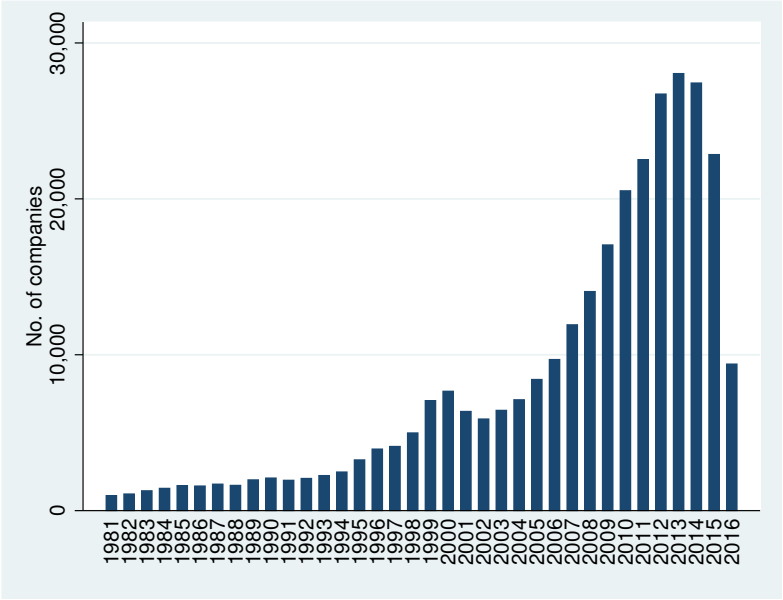
The database started being populated in May 2007, and for every company the date in which the related record was created is reported. The pace of new record creation was however rather limited until the beginning of 2013, when it stabilized at around 200 records per day, on average, with the exceptions of two dates in August 2013 and April 2014, respectively, when several thousand records where added, probably as the consequence of the acquisitions of additional sources (Figure 1). Information on funding deals, however, goes back much earlier in time (with good coverage starting since around 2001), as it is reported later in this Section.

As the predominant source of interest for the database is the coverage of young companies, the graphs of this Section are limited only to the sample of young companies less than 10 year old. The majority of start-ups are located in the United States and operate in the retail, data-analytics, and mobile app sectors (Figure 2).[5] However, the share of companies with at least one registered VC deal is much higher in sectors like biotech and health care. As will be detailed in the following Section, the majority of funding, measured in USD, is concentrated in the retail and biotechnology sector.

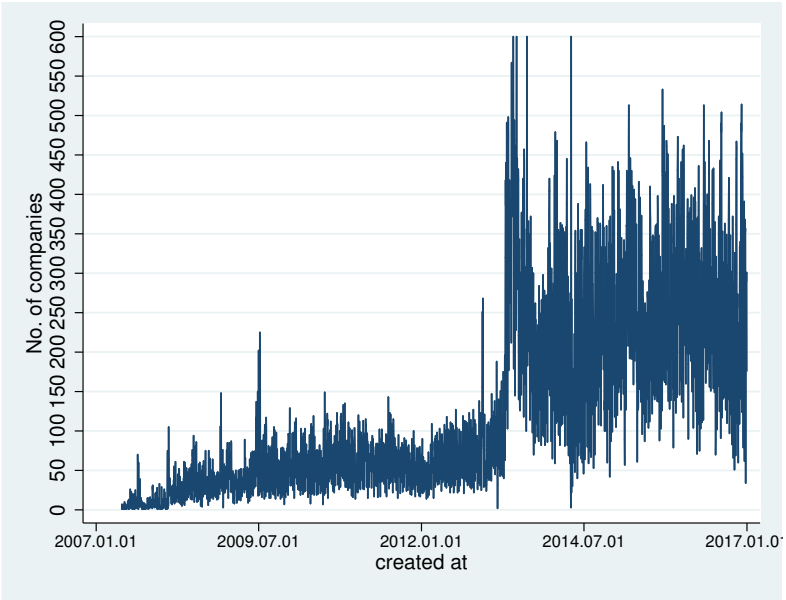
One way to benchmark the coverage of Crunchbase is comparing it with the OECD Entrepreneurship Financing Database. These latter data are typically compiled by national or regional Private Equity and Venture Capital Associations, often with the support of commercial data providers.[6] Figure 3 shows that both the pattern across years (Panel A) and countries (Panel B) are substantially similar across the two data sources. The share of investments accounted for the United States also appear comparable across the two sources. Crounchbase appears to report slightly more investment starting from year 2010, with the difference increasing over time and reaching 22 000 USD million in 2015, corresponding to one third of the total amount reported in the OECD Entrepreneurship Financing Database. Comparisons with other sources at micro-level (e.g., VentureXpert or PwC, available from

Figure 1: Number of companies by founding year and by date in which the record was created

Panel A: by founding year



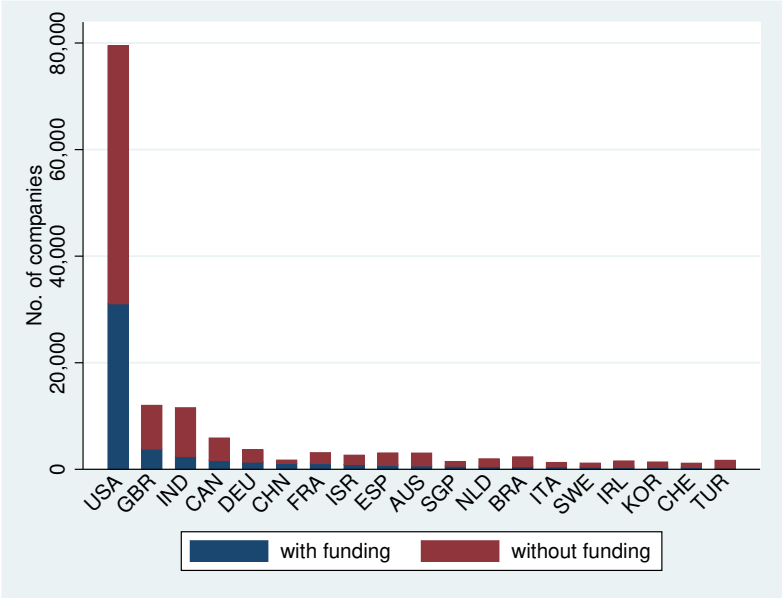
Panel B: by date of entry in the database



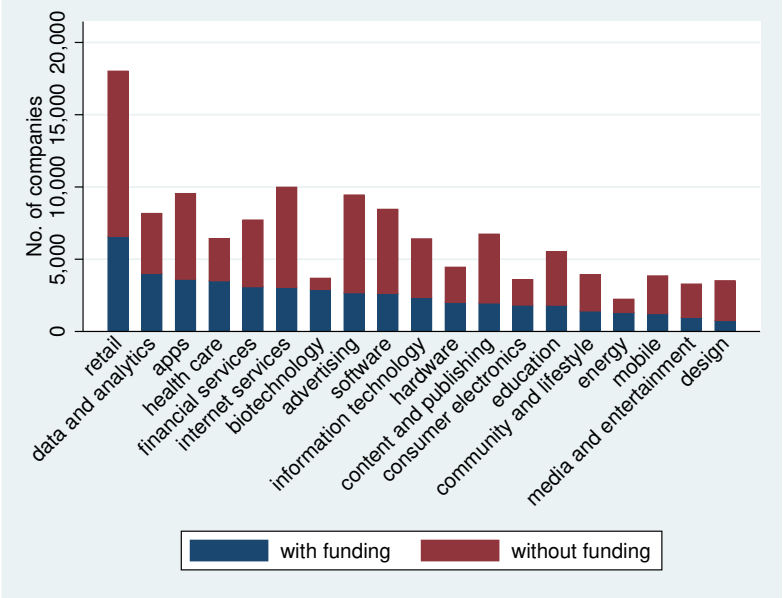
Note: The y-axis of the graph in Panel B is truncated.
Source: http://www.crunchbase.com

Figure 2: Number of companies by country and by category

Panel A: by country



Panel B: by category



Note: the sample is limited to companies less than 10 years old.
Source: http://www.crunchbase.com

the authors upon request) also suggest that the coverage is very comprehensive, especially for start-ups located in the United States.

2.2 Information on companies

The information reported in the database on companies consist of the company size class, its location (city and region), its primary role (firms, group, investor, or school), its status (operating, acquired, IPO, or closed), its founding date, and the dates on which the record was created and updated, respectively. The term "company" here refers to all entities irrespective of her primary role.

As can be seen in Figure 4, the absolute majority of young companies in the database are micro-enterprises. This might be seen as a positive aspect of the database, as generally coverage of micro-enterprises is limited in proprietary databases. Regarding the company roles, more than 90% of companies are private firms (called "companies" in the database); around 5% of the records refer to investors, and the residual share of entities refer either to business groups or to schools (generally universities).

2.3 Information on individuals

Crunchbase contains around 580 000 records on people who are connected to at least one company listed in the database. The following variables are reported: full name, location (city and region), gender, job title, and the dates on which the record was created and updated, respectively. As for graphs in the previous Sections, the figures in this Section are also limited to people linked to a company less than 10 years old.

Figure 5 reports the number of people covered by the dataset aggregated by their primary title and their current company category, respectively. Both graphs also report the gender breakdown (male, female, or unknown). As it is possible to see, the gender gap is marked across all titles and categories, although with some substantial difference in shares. Most people are classified as founder, co-founder, and CEO. The distribution across categories mirror that of companies, but with some differences in the relative ranking.

The people table is linked to both the company table, and to an additional "job" tables reporting almost one million job spells for 521 000 people. Although it is very likely that the only a fraction of the whole professional life is covered for most people, the data can still be an invaluable source of information, e.g. to explore the extent to which the flow of ideas is linked to workers' flow.

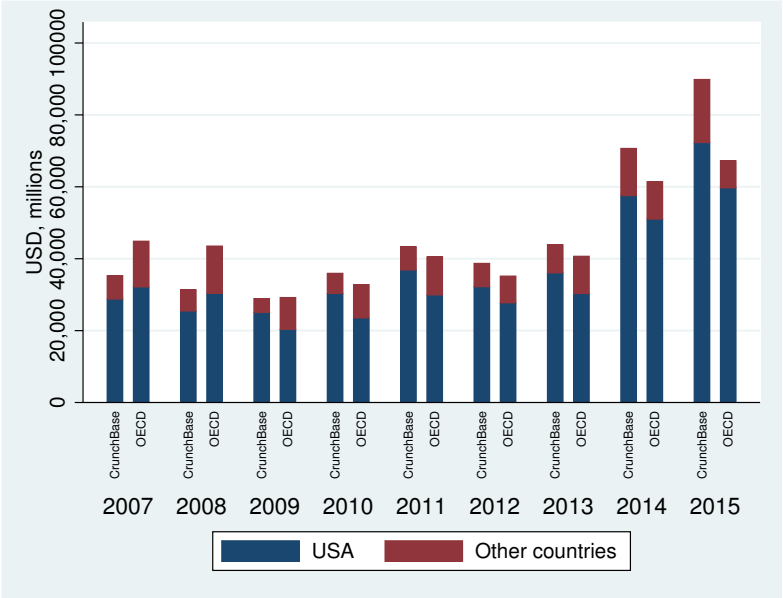
2.4 Information on venture capital and other risk finance providers

Crunchbase contains extensive information on risk financing. A number of different and linked tables list investors and investment rounds, reporting in most cases the amount of capital involved. The number of investors involved and the type (e.g., VC, business angel, private equity, etc.) is always reported. As it is clearly visible in Figure 6, the coverage of VC deals predates the creation of the database by many years, with deals dating as early as the second quarter of 2001 being listed; however, the coverage seems to increase significantly over time.

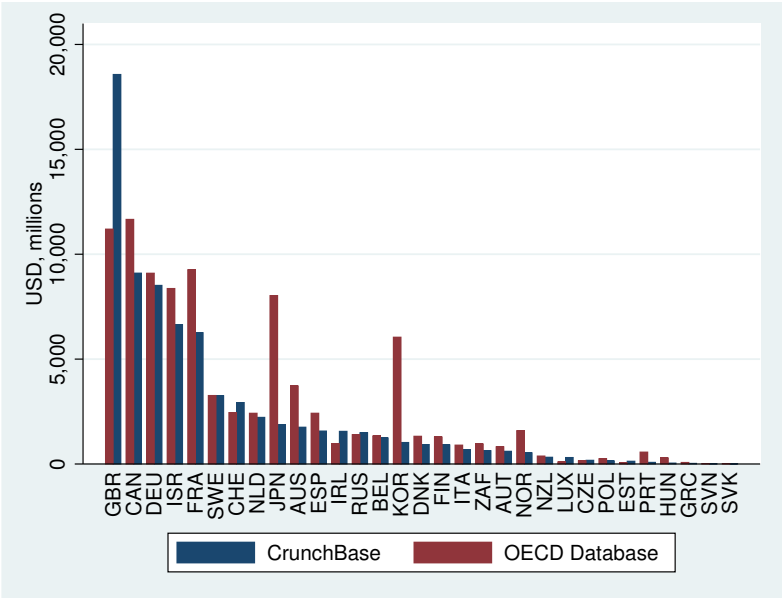
Figure 7 shows the share of funding by type (Panel A) and by recipient category (Panel B). The absolute majority of capital is supplied by venture capitalists, with private equity and debt financing being the other two

Figure 3: Comparison of Crunchbase with the OECD Entrepreneurship Financing Database

Panel A: by year



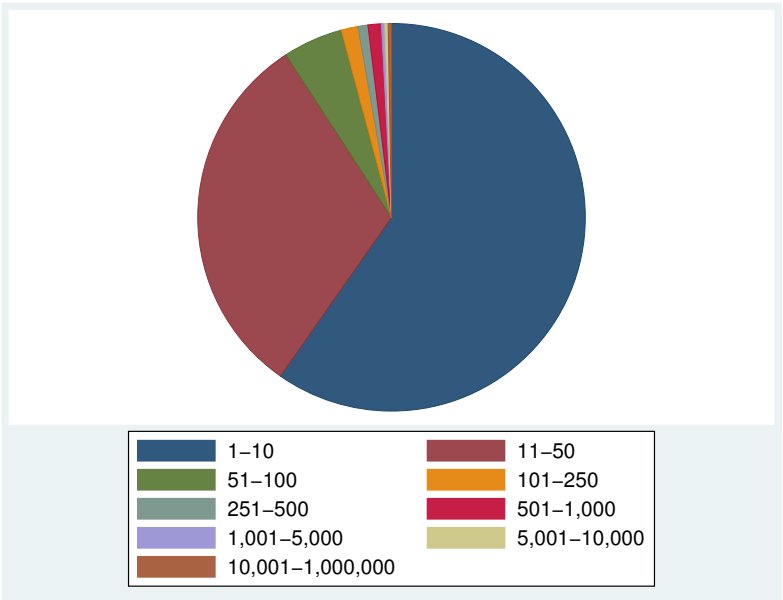
Panel B: by country



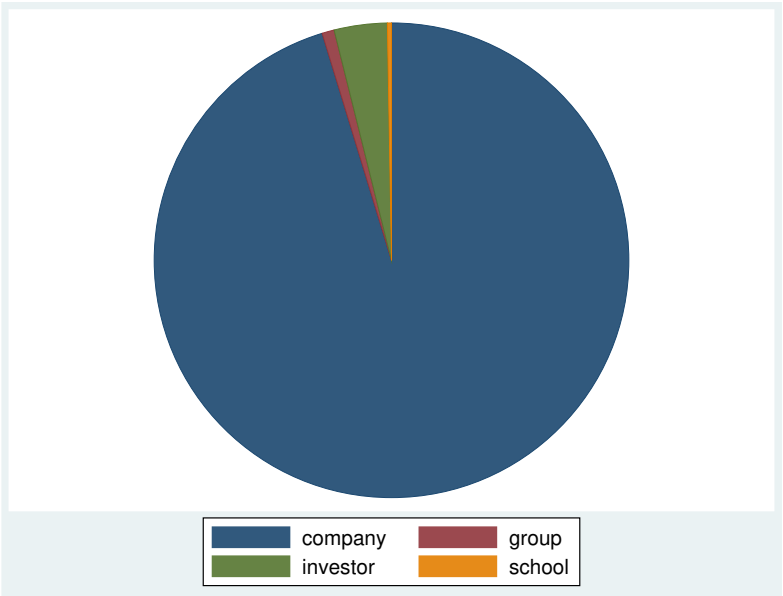
Note: the sample is limited to companies less than 10 years old.
Source: http://www.crunchbase.com

Figure 4: Number of companies by size class and primary role

Panel A: by size



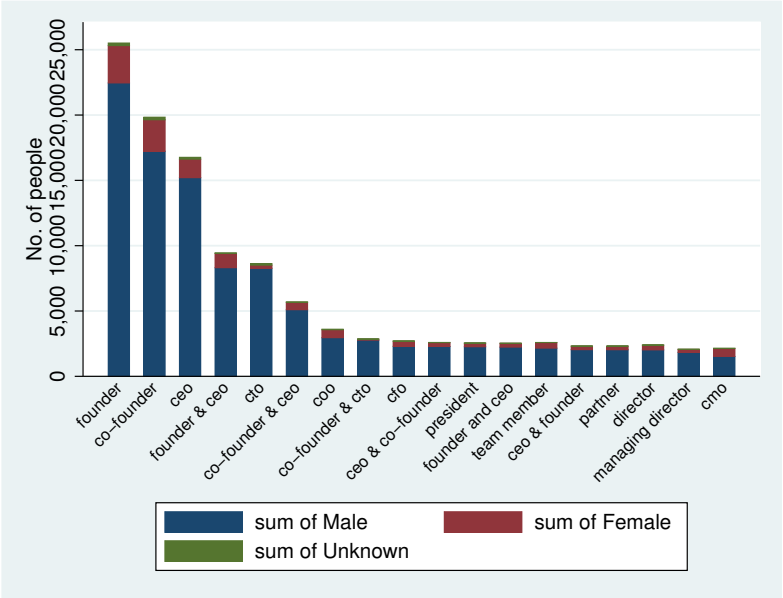
Panel B: by primary role



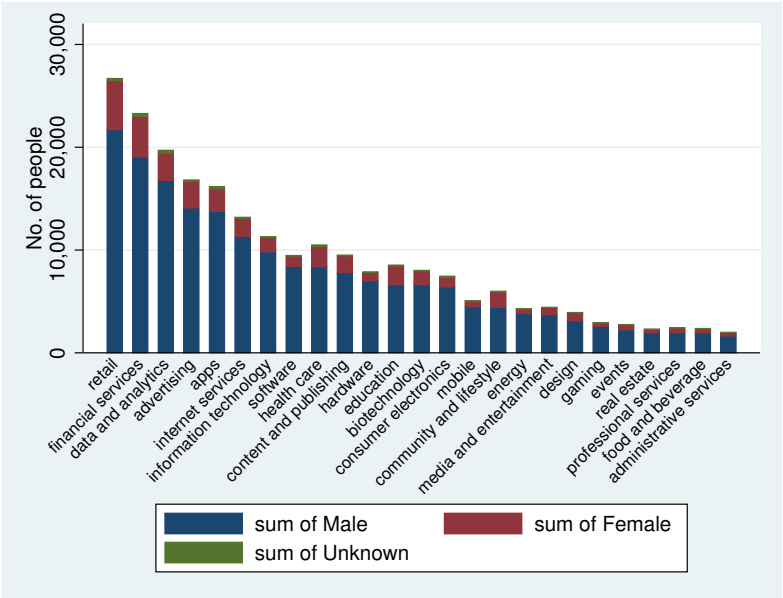
Note: the sample is limited to companies less than 10 years old.
Source: http://www.crunchbase.com

Figure 5: Number of people by primary title and category, with gender breakdown

Panel A: by title

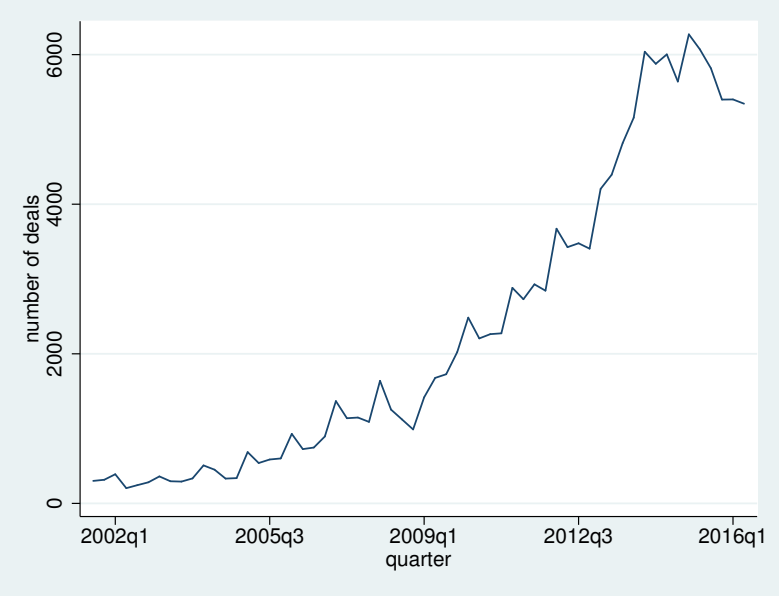


Panel B: by category



Note: the sample is limited to individuals linked to companies less than 10 years old.
Source: http://www.crunchbase.com

Figure 6: VC deals by quarter



Source: http://www.crunchbase.com

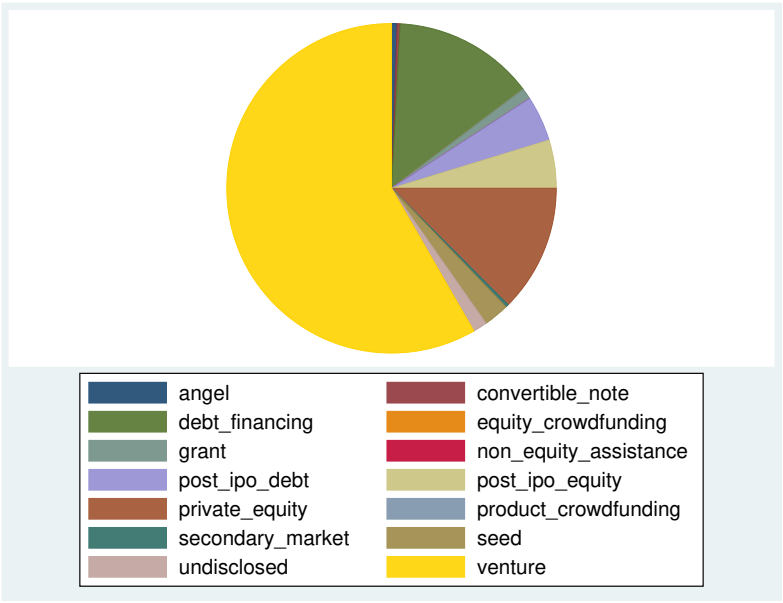
major sources of risk financing. The largest recipient categories are retail and biotechnology, despite the relatively low number of companies listed in the database for the latter category (cf. Figure 2).

2.5 Information on IPOs and acquisitions

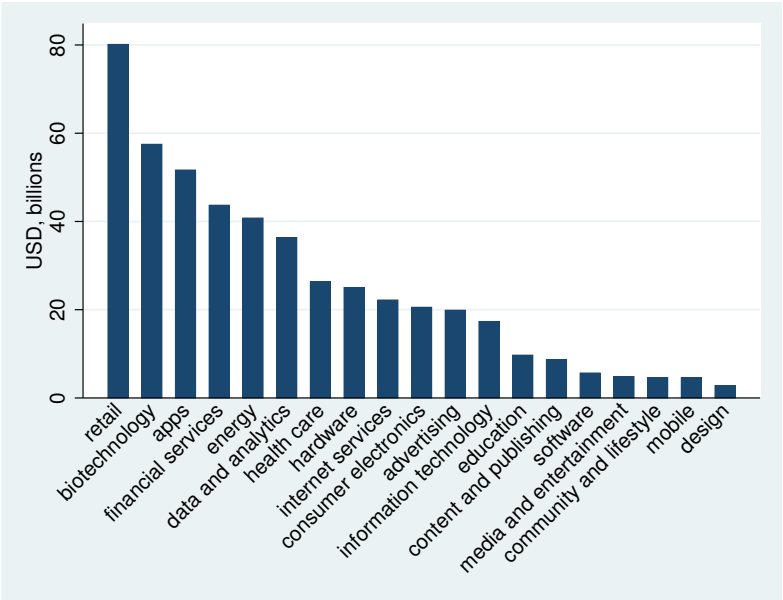
The tables on IPOs and acquisitions contain information on around 11 thousand and 26 thousand individual events, respectively. The exact date of the event is always reported, while the amount paid by the acquirer, or raised through the IPO, is non-missing only for a minority of observations. Panel A of Figure 8 shows a steeply increasing trend in the quarterly flow of acquisitions in the database starting at the beginning of 2013. However, this trend may also depend on idiosyncratic characteristics of the database coverage and it would need to be benchmarked with alternative sources. The quarterly flow of IPOs, conversely, appears to be more regular. The distribution of events by country (the location of the acquired company is used in the case of acquisitions) shows a clear predominance of United States, especially for acquisitions.

Figure 7: Total funding by type and recipient category

Panel A: by type



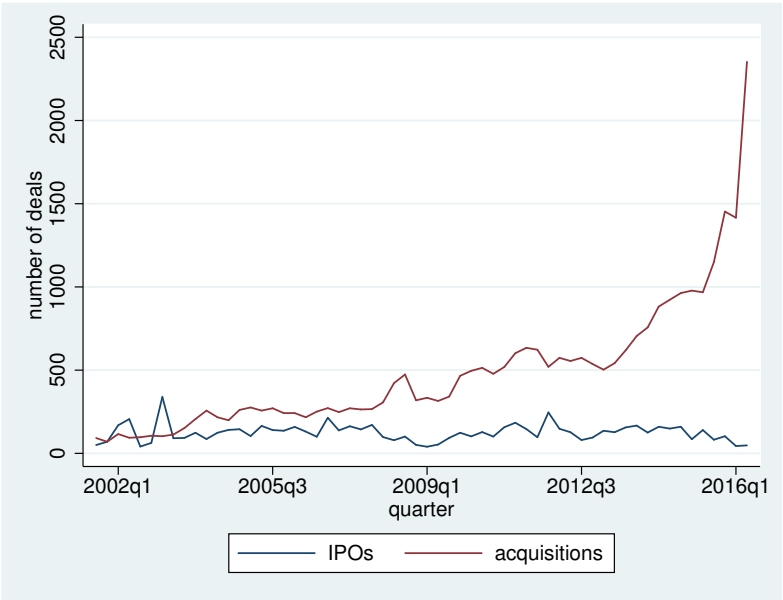
Panel B: by recipient category



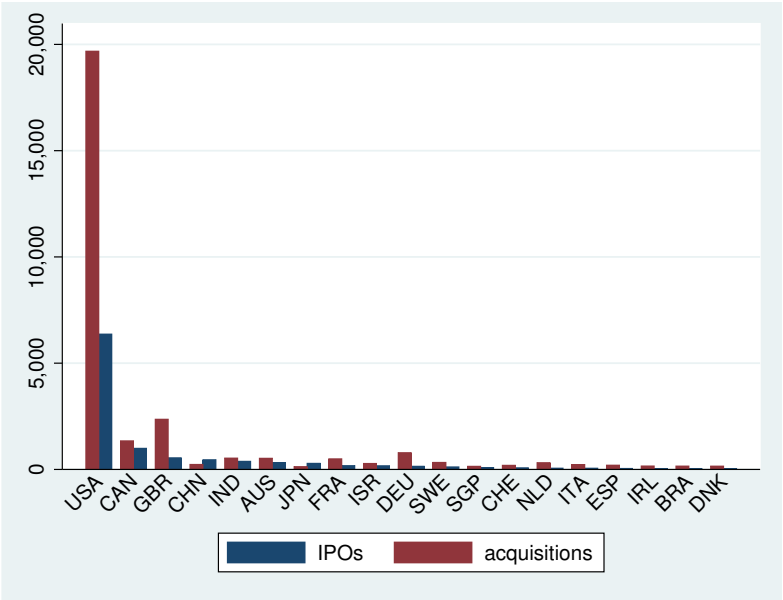
Note: the sample is limited to companies less than 10 years old.
Source: http://www.crunchbase.com

Figure 8: Number of IPOs and acquisitions by quarter and country

Panel A: over time



Panel B: by country



Note: the sample is limited to companies less than 10 years old.
Source: http://www.crunchbase.com

3. SCHOLARLY USE OF CRUNCHBASE

In this section, we provide an overview of scholarly research that has used Crunchbase. With help of various sources including Google Scholar, we identified more than 90 articles and papers that relied on Crunchbase. Most of these works are recent, having been published in 2015 or 2016, while the first was published in 2009. In order to classify these works, we consider where they were published, the research questions they address, and the methods used. Our focus is primarily on empirical studies that use Crunchbase as their main source of data. Thus, we do not focus on studies like Hasenpusch and Baumann (2016), which employs merely Crunchbase to complement ORBIS, nor do we cover studies that rely on databases like VICO (Colombo et al., 2017; Cumming and Vismara, 2017), which combines information from Crunchbase with other secondary sources, nor studies which complement VICO with handpicked data from Crunchbase (Bertoni and Tykvová, 2015).

3.1 Journal Articles

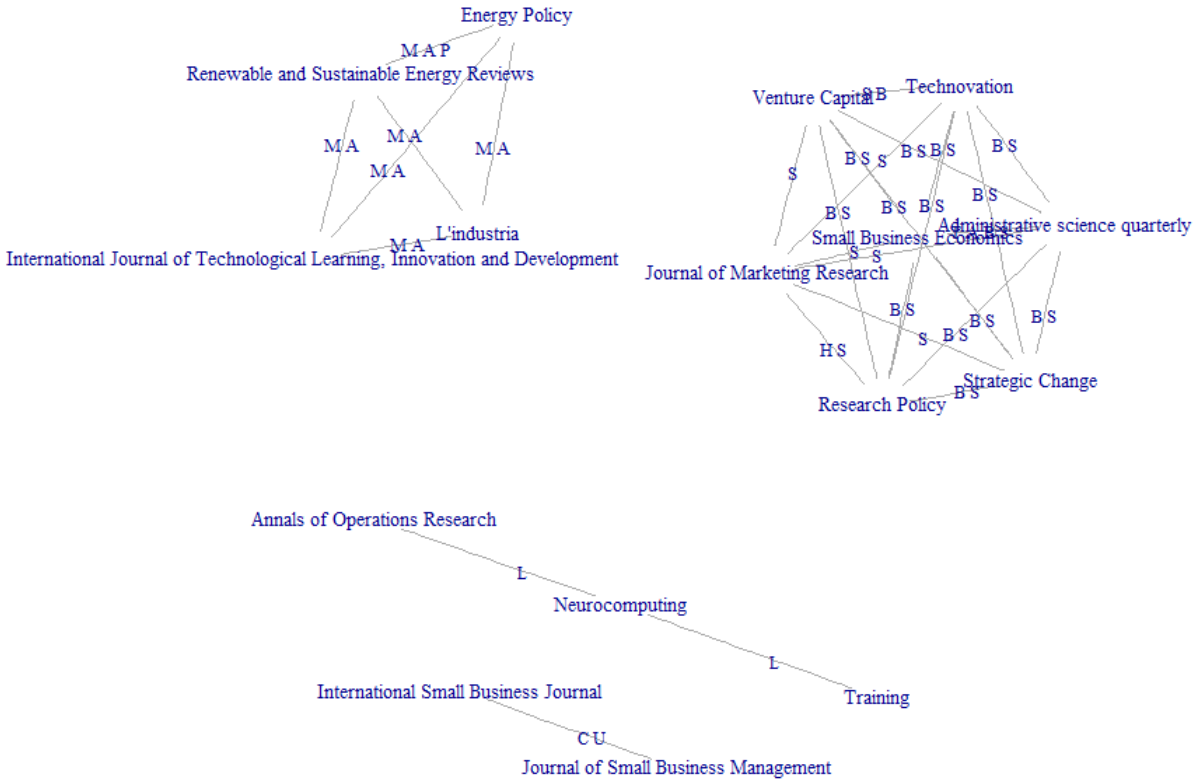
As one would expect from their recent date of publication, a minority of the works we identified has yet been published in peer-reviewed journals, although we found some instances of working papers that reappeared in a different form as journal articles (e.g. O. Alexy et al., 2010; O. T. Alexy et al., 2012; Santana et al., 2014; Santana et al., 2017; Tata et al., 2016; Tata et al., 2017). However, the articles we identified were published in a wide variety of journals. While most journals are related to economics and/or management, the list also includes articles in journals dedicated to information professionals (Feldmann, 2016), IT professionals (Dumas, 2014), and journals focusing on science policy (Barnes, 2016). Apart from English, articles have also been published in Italian (Marra et al., 2014), and in Japanese (Uenoyama et al., 2014).

In order to analyze these publications, we looked for common names among co-authors. Figure 9 displays the links between journals with the first letter of the common names next to each link. There are four separate groups of interconnected journals in the figure. The largest group concerns management journals; the second group of journals has a narrower scope with green innovation as common denominator; the third group seems to have a more technological focus, while journals in group four focus on studying small businesses. It should be noted that some of the journals in these groups are among the top journals in their fields, which shows that Crunchbase is already accepted as legitimate source for research by many experts.

Given the variety of publication outlets, it is no surprise that the research questions that have been tackled with help of Crunchbase cover a wide range of topics. In some cases, Crunchbase is used as a source of examples of the phenomenon that is described, e.g. in Dumas (2014). In other cases, the data are used to corroborate theories, e.g. in Ter Wal et al. (2016), or else, Crunchbase is used to illustrate the potential of prediction methods, e.g. in Zhong et al. (2016b).

Focusing now on the four distinct groups of related journals in more detail, and while the labels of the links may suggest that there are only one or two authors who lead a group, upon closer inspection the range of perspectives adopted even within a group can still be large. In the largest group, the first publication (J. Block and Sandner, 2009) uses aggregate Crunchbase data on funding rounds to investigate the effect of the 2007 financial crisis. This theme is further explored by J. H. Block and Sandner (2011). Soon after, however, the authors of the first publication seem to have started working on other issues with different co-authors, and a particularly fruitful line of research appears with the exploration of the relationships among investors and startups that are recorded in Crunchbase (O. T. Alexy et al., 2012; Ter Wal et al., 2016). Besides, the authors in this group tried to exploit

Figure 9: Grouping of journals where articles using Crunchbase have been published based on common author names



Note: The letters on the edges correspond to the initials of the family names of the authors the connected journals have in common.

information on the startup team (Homburg et al., 2014), on trademarks (J. H. Block et al., 2015) and on patents (Zhou et al., 2016).

The approach of the authors of articles published in the second group of journals is generally less theory driven and more descriptive. After pondering methodology (Marra et al., 2014), they investigated phenomena such as social innovation (Maiolini et al., 2015), and industrial clusters (Marra et al., 2017a; Marra et al., 2017b). On their part, the authors of group three mainly analyze investment behavior (Zeng et al., 2016; Zhong et al., 2016b) while those of group four analyze the behavior of business angels (Croce et al., 2016a; Croce et al., 2016b).

In addition, a few more articles, thematically close to one more more of these groups, have been written by other authors: The *Journal of Business Venturing Insights* published a study that uses Crunchbase to identify entrepreneurs among Twitter users and study their communication (Tata et al., 2017); the *Journal of Cleaner Production* features a social network analysis of clean technology entrepreneurs (Lange, 2016); the journal *Social Networks* accepted a multi-layer network analysis of investment in serial entrepreneurs (Santana et al., 2017) and the journal *Expert Systems with Applications* accepted and exercise in data-mining (Martens et al., 2011); finally, the *International Journal of Entrepreneurial Venturing* contains an article analyzing co-investment networks of business angels (Werth and Boeert, 2013).

Finally, a few articles are more difficult to relate to one of these four groups: *Internet Research* published research that proposes a machine learning approach to predicting investment behavior (Liang and Yuan, 2016), which could be related to group 3; *Stanford Technology Law Review* published research on patents in the Smartphone industry (Reidenberg et al., 2014), which relates to some of the themes of group 1; while recent research on accelerators could fit either with group 1 or group 4 (D. J. Smith et al., 2010; Ko and McKelvie, 2015; Barnes, 2016) or give birth to a new group. Given its interest in a particular sector, social media startups, the study published in the *International Journal of Information Management* (Ghezzi et al., 2016) could fit with group 2.

3.2 Non-journal literature

Apart from peer reviewed articles in academic journals, numerous works based on Crunchbase have been presented at workshops and conferences. To a great extent these research endeavors also explore the themes identified above, noting that authors who work on topics closer to computer science tend to put more weight on publications in conference proceedings than other scholars. Hence topics like data mining or machine learning are particularly well-represented in the non-journal literature.

Related to the first group, which focuses on management, we find papers that report the preliminary results of research subsequently published in journals (J. H. Block et al., 2010; O. Alexy et al., 2010). There are also other works by the same authors, such as a study that seeks to exploit the information in Crunchbase on the location of company headquarters and their sector of activity (Berchicci et al., 2011). Several very recent papers have further been written by authors who are likely to target the same group of journals (Hallen et al., 2017; Huang and M. Z. Shi, 2015; Huang and Z. Shi, 2016; Howell, 2017; Kaminski et al., 2016; Dams et al., 2016, for instance), while some other works seem to be thematically very close (Bradic, 2012; Wang, 2016).

Related to the second group, which focuses on sector specific studies with a particular interest in green innovation, we find preliminary results on social innovation by Marra et al. (2015). Papers on green innovation include Townsend (2014) and Townsend (2015). Other sector-specific studies include research on the Fintech market (Haddad and Hornuf, 2016; Dorfleitner et al., 2016), a survey of recent business models in the music industry (Waldner et al., 2012), one of business models in the mobility sector (Remane et al., 2016), and a study

of Web entrepreneurship (Spiegel et al., 2013). There is also a couple of papers that focus on regional factors associated with startup growth (Motoyama and Bell-Masterson, 2014; Desai and Motoyama, 2015).

Related to the third group, which encompasses contributions that experiment with new methods to select investment opportunities, we find papers with preliminary results (Zhong et al., 2016a; Liang and Yuan, 2012; Liang and Yuan, 2013; Santana et al., 2014). Several other papers are thematically close (Gupta et al., 2015; Raghuvanshi et al., 2015). Besides, there is a large stream of research on data-mining that rather fits in this group as well (Cheng et al., 2016; Spiegel, 2012; Parisot et al., 2016; Xiang et al., 2012; Batista and Carvalho, 2015a; Batista and Carvalho, 2015b; Krishna et al., 2016; Yan et al., 2016).

Related to the fourth group, with its focus on business angels, we could add, thus extending its scope, research on venture capital more in general (Cumming et al., 2014; Nuscheler, 2016), research on crowdfunding (McGuire, 2017; Cheng et al., 2016; Kaminski et al., 2016), and research on accelerators (den Besten and Dalle, 2015; S. W. Smith and Hannigan, 2015; Perotti and Yu, 2015; Dams et al., 2016).

In addition, it should be noted that a few doctoral dissertations have relied on Crunchbase: Two at the Computer Science department of University College London (Stone, 2014; Zhao, 2016), two in US management schools (Wu, 2016; Strickling, 2016), and one at the Management and Economics department at the University of Liège in Belgium (Gillain, 2016). Apart from (Zhao et al., 2015), no publication from these works has yet appeared on Google Scholar. Finally, Crunchbase appears in a Honor's Thesis at the Economics Department of Stanford University, (Wei, 2015) as well as in a series of student reports for a computer science course (Adcock et al., 2013; Zhang et al., 2015; Fougner et al., 2013). The use of Crunchbase in a class assignment is also reported by (den Besten, 2014) and a study using Crunchbase has even been published in the University of Chicago *Undergraduate Business Journal*.

4. LINKING CRUNCHBASE WITH OTHER SOURCES

Several works have benefitted from the "linkability" of Crunchbase, by integrating it with other datasets in order to augment the possibilities offered to research through to the data thus merged. Methodologically speaking, a small group of scholars have even proposed to use more generic mechanisms to link Crunchbase data with other data[7], while the question of how Web-based data like Crunchbase can be integrated with the Semantic Web is treated by O'Riain et al. (2012), Goto et al. (2013), and Färber et al. (2016).

Among the sources of data that have been used to complement Crunchbase we find KickStarter (Kaminski et al., 2016), Twitter (Tata et al., 2016; Tata et al., 2017), and LinkedIn (Nuscheler, 2016), but two databases, PATSTAT (Zhou et al., 2016; Ter Wal et al., 2016; Lerman, 2015; Reidenberg et al., 2014; Wang, 2016) and Seeddb (D. J. Smith et al., 2010; den Besten and Dalle, 2015; Porat, 2014; Perotti and Yu, 2015; Dams et al., 2016), for which we now provide examples, seem to have garnered more specific attention.

4.1 Matching with PATSTAT

Tarasconi and Menon (2017) describe in detail a procedure to match Crunchbase with information on intellectual property (IP) contained in PATSTAT, the worldwide database on IP maintained by the European Patent Office (EPO). While other scholars have matched Crunchbase with IP data for specific subsample of the two databases, as mentioned in the previous Section (e.g. Zhou et al., 2016; Ter Wal et al., 2016; Lerman, 2015), this is the only matching exercise that covers the entirety of the Crunchbase database, to the best of the authors' knowledge. The match covers both companies and inventors. Given that neither administrative nor other unique identifiers are available in either of the two databases, the matching is based on a "fuzzy" procedure that exploits the available overlapping information across the two databases: the company names, their location, and the names of the people linked to them. The matching procedure needs to be carefully designed in order to maximise the number of correct matches, while at the same time minimising both "false positive" and "false negative" errors. This is not straightforward as the spelling of companies' and people's name is not always consistent across the two databases. Furthermore, in the PATSTAT database there is neither a unique internal identifier for either applicants and inventors, which have therefore to be adequately disambiguated before engaging in the matching exercise. This is particularly critical for inventors, where homonymy is very frequent (the so-called "John Smith" problem).

Almost 50 thousand companies, out of the 447 thousand listed in Crunchbase in January 2017 (excluding venture capital companies), are found to own one or more patents, for a total of around 12 million patents. Around 220 thousand of those have been applied for by companies created after 2005. The share of patentees for US companies is 15%, but the share doubles for companies reporting at least one funding round. Regarding individuals, out of the 578 thousand professionals listed in Crunchbase who could be potential patent inventors, around 25 thousand are found to have a correspondent in PATSTAT. These inventors account for 2.2 million patent applications.

A first analysis of the resulting database shows some interesting facts. For instance, it clearly appears that patenting start-ups are generally more likely to be VC recipients than non-patentees across a variety of countries and technological categories. When looking at individual inventors, the figures suggest that more than one out of five start-uppers working in biotechnology are patent inventors. Furthermore, those start-uppers who report their job title to be founder, president, co-founder, chief executive officer (CEO), and chief technology officer (CTO), appears to be significantly more likely to be inventors, with a share higher than 10%. However, the share of inventors is significantly higher for male start-uppers than for female ones.

Given the international dimension of both the Crunchbase and PATSTAT database, on the one hand, and the richness and granularity of the available micro-data for both databases, on the other hand, it easy to recognize the potential of the matched database for meaningful comparisons across countries, technologies, and types of start-ups. Directions for further research include e.g. the analysis of the role of IP assets in securing venture capital; the characterization of the IP portfolio of high-growth patenting start-ups, of start-ups developing radical or breakthrough innovations, and of inclusive start-ups; the analysis of the linkages between patenting innovative start-ups and public research, exploiting patent citations to other patents, as well as to non-patent literature (NPL).

4.2 Matching with Seed-db to study Accelerators

Recent years have seen the surge of a new model for business incubation known as "accelerators". Although there are many variations, accelerators are generally private entities that take small equity positions in the startups they select. They provide strong coaching programs to sequential cohorts of founders during short periods of time (three to six months on average). Unlike incubators, they rarely provide office space to accelerated startups. Accelerators have been associated in the media, with the success of startups Dropbox and Airbnb, which were "accelerated" by Y-Combinator, a prominent and early accelerator in the Silicon Valley.

The accelerator model has first been debated with respect to the uncertainty associated with accelerators' business models, as they need to rely on an extremely quick growth and exit path for startups, and as some of the most prominent accelerators have later raised large venture-capital funds. Putting aside the question about whether accelerators can exist as stand-alone entities, another issue, closer to the preoccupations of both entrepreneurs and the research community interested in the economic and management of innovation, has to do with the impact of accelerators on the startups they "accelerate".

In this respect, by linking Crunchbase with Seed-DB, a dataset curated independently by Jed Christiansen that lists several thousand companies that have benefited from the support of accelerators while identifying the accelerator they have benefited from and the cohort to which they belonged, (den Besten and Dalle, 2015; Morfin et al., forthcoming) have focused on the early stage funding (seed capital and Series A) of startup companies that have taken part in programs provided by US-based accelerators. Comparing with similar companies that did not, by using standard propensity matching techniques, they observe that US accelerators have a very different impact on early stage funding, even amongst the most prominent ones. These observations are corroborated when observing the centrality of the investors that accelerated startups attract, as measured by Katz centrality within the network of investors, comparing again to the investors in a control of non-accelerated startups. Investor networks are straightforwardly defined here on behalf of co-investment in startups. The same accelerators that have a statistically significant and positive effect on early stage funding of startups also have a positive effect on the Katz of startup investors, and conversely. Taken together, these results are coherent with the fact that more prominent investors are able to pay a higher price, which would be reflected by an increased funding for the same share of the startup company.

They also point to the fact that accelerators play a role of platform between startups and investors: some accelerators attract higher-potential startups and thus more prominent investors, and thus even more high-quality startups, etc. This platform role is also in line with the heterogeneity of results observed, some accelerators being even associated with a significantly negative impact of early stage startup funding. As a consequence, and even if claims according to which accelerators would have developed a new model could be challenged, at least on behalf of startup fundraising, there might actually exist an accelerator "model" that would also apply to other actors dedicated to coaching early-stage startups: accelerators or incubators "as a platform".

The platform model does not seem to be accessible to all actors, notably in relation to their belonging, or not, to a thriving ecosystem: the 3 “best” accelerators according to their statistical impact on startup fundraising are located either in the Silicon Valley (two of them) and in New York. In a thriving ecosystem, attract “good” startups and “good” investors, whichever side of the platform is developed at first, appears more feasible. In a less thriving ecosystem, the construction of a successful accelerator or incubator platform is probably a bigger challenge, that could sometimes be accessible to an actor that would manage to build a sufficiently strong and qualitative track-record allowing it to create positive externalities on the other side of the platform, and thus to attract high-level investors and thus even higher-quality startups.

5. CONCLUSION

This paper provides an overview of the current and prospective use of the Crunchbase database as an original data sources on innovative companies and start-ups around the world. Considering the potentialities offered by Crunchbase through its contents and its linkability with other relevant data sources, it is not surprising that numerous works have started using it as a source of relevant data for their research, some of which is already published in top-tier journals. For the same reasons, we definitely expect this trend to continue and probably to grow rapidly in the coming years. In this context, several scholars may be interested in assessing the potential of the database for economic and managerial research, and therefore they may find this paper to be an useful resource. Increased coordination among different teams and even research communities - e.g. by a dedicated conference - might prove particularly fruitful, not less with respect to methodologies and tools.

ACKNOWLEDGMENTS

The authors wish to thank Nick Johnstone for useful comments and suggestions. They also thank Jérémie Morfin for his contributions to Section 4.2. Usual disclaimers apply. The views expressed here do not necessarily represent those of the OECD or its Member governments. MdB is supported by the Agence Nationale de Recherche (ANR-10-LabX-11-01).

References

Adcock, A. B., M. Lakkam, and J. Meyer (2013), "CS 224W Final Report Group 37", Stanford University.

Alexy, O. T., J. H. Block, P. Sandner, and A. L. Ter Wal (2012), "Social capital of venture capitalists and start-up funding", *Small Business Economics*, Vol. 39, No. 4, pp. 835–851.

Alexy, O., J. H. Block, P. G. Sandner, and A. L. Ter Wal (2010), "The social capital of venture capitalists and its impact on the funding of start-up firms".

Barnes, A. (2016), "Mass Producing Innovation: A Case Investigation on Why Accelerators Might Not Be a Paradox", *STI Policy and Management Journal*, Vol. 1, No. 2.

Batista, F. and J. P. Carvalho (2015a), "Text based classification of companies in CrunchBase", in: *Fuzzy Systems (FUZZ-IEEE), 2015 IEEE International Conference on*, IEEE, pp. 1–7.

Batista, F. and J. P. Carvalho (2015b), "Text based classification of companies in CrunchBase", in: *FUZZ-IEEE*.

Berchicci, L., J. H. Block, and P. G. Sandner (2011), "The influence of geographical proximity and industry similarity in a business angel's investment choice".

Bertoni, F. and T. Tykvová (2015), "Does governmental venture capital spur invention and innovation? Evidence from young European biotech companies", *Research Policy*, Vol. 44, No. 4, pp. 925–935.

Block, J. H., G. De Vries, and P. G. Sandner (2010), "Venture capital and the financial crisis: An empirical study across industries and countries".

Block, J. H., C. O. Fisch, A. Hahn, and P. G. Sandner (2015), "Why do SMEs file trademarks? Insights from firms in innovative industries", *Research Policy*, Vol. 44, No. 10, pp. 1915–1930.

Block, J. H. and P. Sandner (2011), "Venture capital funding in the middle of the year 2011: are we back to pre-crisis boom levels?", *Strategic Change*, Vol. 20, No. 5-6, pp. 161–169.

Block, J. and P. Sandner (2009), "What is the effect of the financial crisis on venture capital financing? Empirical evidence from US Internet start-ups", *Venture Capital*, Vol. 11, No. 4, pp. 295–309.

Bradic, A. (2012), "The Role of Social Feedback in Financing of Technology Ventures", *Papers*, arXiv.org.

Cheng, M., A. Sriramulu, S. Muralidhar, B. T. Loo, L. Huang, and P.-L. Loh (2016), "Collection, exploration and analysis of crowdfunding social networks", in: *Proceedings of the Third International Workshop on Exploratory Search in Databases and the Web*, ACM, pp. 25–30.

Colombo, M. G., D. D'Adda, P. Malighetti, and A. Quas (2017), "The Impact of Venture Capital Monitoring in Europe", *techreport*, Available at SSRN: https://ssrn.com/abstract=2906236orhttp://dx.doi.org/10.2139/ssrn.2906236, Politecnico di Milano, Milan, Italy.

Croce, A., M. Guerini, and E. Ughetto (2016a), "Angel Financing and the Performance of High-Tech Start-Ups", *Journal of Small Business Management*.

Croce, A., F. Tenca, and E. Ughetto (2016b), "How business angel groups work: Rejection criteria in investment evaluation", *International Small Business Journal*, p. 0266242615622675.

Cumming, D. J. and S. Vismara (2017), "De-segmenting research in entrepreneurial finance", *Venture Capital*, Vol. 19, No. 1-2, pp. 17–27.

Cumming, D. J., U. Walz, and J. C. Werth (2014), "Entrepreneurial Spawning: Experience, Education, and the Role of Venture Capital", *techreport*, Available at SSRN: https://ssrn.com/abstract=2503532 or http://dx.doi.org/10.2139/ssrn.2503532, York University.

Dams, M. C., V. Sarria-Allende, R. Pasquini, and G. Robiolo (2016), "Accelerators, Networks and Venture Capital Financing", in: *Academy of Management Proceedings*, vol. 2016, 1, Academy of Management, p. 11047.

den Besten, M. (2014), "Crunchbase Case Study", in: *2nd International Conference on the Knowledge Commons*, IASC, New York, NY.

den Besten, M. and J.-M. Dalle (2015), "A comparative analysis of funds raised by startups that benefitted from US accelerators", in: *Interdisciplinary European Conference on Entrepreneurship Research*, Montpellier, France.

Desai, S. and Y. Motoyama (2015), "The Regional Environment: Indianapolis - Insights from High-Growth Companies", *techreport*, Available at SSRN: https://ssrn.com/abstract=2660938 or http://dx.doi.org/10.2139/ssrn.2660938, SSRN.

Dorfleitner, G., L. Hornuf, M. Schmitt, and M. Weber (2016), "The Fintech Market in Germany".

Dumas, M. (2014), "The Rise of the Estonian Start-Up Sphere", *IT Professional*, Vol. 16, No. 4, pp. 8–11.

Färber, M., C. Menne, and A. Harth (2016), "A Linked Data wrapper for CrunchBase", *Semantic Web*, No. Preprint, pp. 1–11.

Feldmann, L. M. (2016), "TechCrunch and CrunchBase", *The Charleston Advisor*, Vol. 17, No. 3, pp. 34–37.

Fougner, C., J. Lloyd, and S. Wong (2013), "Platonic Graph Inference", CS224w group report 55, Stanford University.

Ghezzi, A., L. Gastaldi, E. Lettieri, A. Martini, and M. Corso (2016), "A role for startups in unleashing the disruptive power of social media", *International Journal of Information Management*, Vol. 36, No. 6, pp. 1152–1159.

Gillain, A. (2016), "Determinants of venture capitalists' exit strategies: An empirical study through survival analysis", PhD thesis, Université de Liège, Liège, Belgique.

Goto, M., B. Hu, A. Naseer, and P.-Y. Vandenbussche (2013), "Linked data for financial reporting", in: *Proceedings of the Fourth International Conference on Consuming Linked Data-Volume 1034*, CEUR-WS. org, pp. 123–135.

Gupta, S., R. Pienta, A. Tamersoy, D. H. Chau, and R. C. Basole (2015), "Identifying Successful Investors in the Startup Ecosystem", in: *Proceedings of the 24th International Conference on World Wide Web*, ACM, pp. 39–40.

Haddad, C. and L. Hornuf (2016), "The Emergence of the Global Fintech Market: Economic and Technological Determinants", *Workin Paper* No. 6131, Available at SSRN: https://ssrn.com/abstract=2830124, CESifo.

Hallen, B. L., J. P. Davis, and P.-L. Yin (2017), "Network Isolates: Entrepreneurial Bootstrapping and the Social Disconnection of New Organizations in the Mobile App Ecosystem", *Working Paper* No. 17-22, Available at SSRN: https://ssrn.com/abstract=2907526, Marshall School of Business.

Hasenpusch, T. C. and S. Baumann (2016), "Strategic media venturing: Private equity investments as a strategic tool for media firms", in: *Mergers and Acquisitions, Entrepreneurship and Innovation*, Emerald Group Publishing Limited, pp. 77–110.

Homburg, C., A. Hahn, T. Bornemann, and P. Sandner (2014), "The role of chief marketing officers for venture capital funding: endowing new ventures with marketing legitimacy", *Journal of Marketing Research*, Vol. 51, No. 5, pp. 625–644.

Howell, S. T. (2017), "Learning from Feedback", in: *The 14th Annual Conference in Financial Economics Research*, Available at https://www.idc.ac.il/en/schools/business/annual-conference/Documents/learning-feedback.pdf.

Huang, J. and M. Z. Shi (2015), "With a Little Help of My (Former) Employer: Past Employment and Entrepreneurs' External Financing", in: *Academy of Management Proceedings*, vol. 2015, 1, Academy of Management, p. 12050.

Huang, J. and Z. Shi (2016), "Entrepreneurial Spawning Across Product Markets: Skills and Performance", *Research paper* No. 15-3, Available at SSRN: https://ssrn.com/abstract=2544521 or http://dx.doi.org/10.2139/ssrn.2544521, Columbia Business School.

Kaminski, J. C., C. Hopp, and T. Tykvova (2016), "New Technology Assessment in Entrepreneurial Financing-Can Crowdfunding Predict Venture Capital Investments?", *SSRN Working paper*, Available at SSRN: https://ssrn.com/abstract=2829777, RWTH Aachen.

Ko, E.-J. and A. McKelvie (2015), "The Influence of Different Types of Legitimacy across Resource Acquisition Stages in Entrepreneurial Firms", *Frontiers of Entrepreneurship Research*, Vol. 35, No. 6, p. 1.

Krishna, A., A. Agrawal, and A. N. Choudhary (2016), "Predicting the Outcome of Startups: Less Failure, More Success", in: *2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW)*, pp. 798–805.

Lange, D. E. de (2016), "A social capital paradox: Entrepreneurial dynamism in a small world clean technology cluster", *Journal of Cleaner Production*, Vol. 139, pp. 576–585.

Lerman, C. (2015), "Patent Strategies of Technology Startups: An Empirical Study", *techreport*, Available at SSRN: https://ssrn.com/abstract=2610433 or http://dx.doi.org/10.2139/ssrn.2610433.

Liang, Y. E. and S.-T. D. Yuan (2016), "Predicting investor funding behavior using CrunchBase social network features", *Internet Research*, Vol. 26, No. 1, pp. 74–100.

Liang, Y. E. and S.-T. D. Yuan (2013), "Investors Are Social Animals: Predicting Investor Behavior using Social Network Features via Supervised Learning Approach", in: *Proceedings of the Workshop on Mining and Learning with Graphs (MLG-2013)*, Chicago, IL.

Liang, Y. E. and S.-T. D. Yuan (2012), "Where's the Money? The Social Behavior of Investors in Facebook's Small World", in: *Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012)*, IEEE Computer Society, pp. 158–162.

Maiolini, R., A. Marra, and M. Luciani (2015), "Social Innovation: An Observational Study in High-Tech Industries", *L'industria*, Vol. 36, No. 4, pp. 535–552.

Marra, A., P. Antonelli, and C. Pozzi (2017a), "Emerging green-tech specializations and clusters–A network analysis on technological innovation at the metropolitan level", *Renewable and Sustainable Energy Reviews*, Vol. 67, pp. 1037–1046.

Marra, A., E. Cassetta, and P. Antonelli (2017b), "Emerging specialisations and software metropolitan clusters-a comparative network analysis on San Francisco, New York and London", *International Journal of Technological Learning, Innovation and Development*, Vol. 9, No. 1, pp. 17–41.

Marra, A., M. Luciani, and P. Antonelli (2014), "Data analytics nel comparto high-tech statunitense: impatti, evoluzione e tutela della privacy", *L'industria*, Vol. 35, No. 2, pp. 319–338.

Marra, A., R. Maiolini, and C. Baldassarri (2015), "Social innovation in the US high-tech industries: its core business and main drivers of innovation", in: *DRUID*.

Martens, D., C. Vanhoutte, S. De Winne, B. Baesens, L. Sels, and C. Mues (2011), "Identifying financially successful start-up profiles with data mining", *Expert Systems with Applications*, Vol. 38, No. 5, pp. 5794–5800.

McGuire, E. (2017), "Can Equity Crowdfunding Mitigate the Gender Gap in Entrepreneurship?", in: *2017 ASSA Annual Meeting*, American Economic Association, Chicago, IL.

Morfin, J., J.-M. Dalle, and M. L. Den Besten (Forthcoming), "Startup accelerators as platforms: an analysis of early-stage fundraising of accelerated companies", *Working paper in preparation*, SSRN.

Motoyama, Y. and J. Bell-Masterson (2014), "Beyond Metropolitan Startup Rates: Regional Factors Associated with Startup Growth", *techreport*, Available at SSRN: https://ssrn.com/abstract=2393516 or http://dx.doi.org/10.2139/ssrn.2393516, Ewing Marion Kauffman Foundation.

Nuscheler, D. (2016), "Regularly change a running system! An analysis of stage-specific criteria for attracting venture capital and changing the likelihood for getting funded", in: *International Finance and Banking Society*.

Oreif, K. (2015), "Structuring data around a topical matter and a.i./n.l.p./ machine learning knowledge system that enhances source content by identifying content topics and keywords and integrating associated/related contents", US Patent App. 14/251,447, URL: https://www.google.com/patents/US20150294220.

O'Riain, S., E. Curry, and A. Harth (2012), "XBRL and open data for global financial ecosystems: A linked data approach", *International Journal of Accounting Information Systems*, Vol. 13, No. 2, pp. 141–162.

Parisot, O., P. Hitzelberger, Y. Didry, G. Vierke, and H. Rieder (2016), "Text analytics on start-up descriptions", in: *Research Challenges in Information Science (RCIS), 2016 IEEE Tenth International Conference on*, IEEE, pp. 1–2.

Perotti, V. and Y. Yu (2015), "Startup Tribes: Social Network Ties that Support Success in New Firms", in: *AMCIS2015*, http://aisel.aisnet.org/amcis2015/e-Biz/GeneralPresentations/17/.

Porat, J. (2014), "Exploring the Policy Relevance of Startup Accelerators", *Issue Brief* No. 4, Available at https://www.sba.gov/sites/default/files/advocacy, SBA.

Raghuvanshi, A., T. Balakrishnan, and M. Balakrishnan (2015), "Predicting Investments in Startups using Network Features and Supervised Random Walks", in:

Reidenberg, J. R., N. C. Russell, M. Price, and A. Mohan (2014), "Patents and Small Participants in the Smartphone Industry", *Stanford Technology Law Review*, Vol. 18, No. 3.

Remane, G., B. Hildebrandt, A. Hanelt, and L. M. Kolbe (2016), "Discovering new digital business model types–a study of technology startups from the mobility sector", in: *Proceedings of 20th Pacific Asia Conference on Information Systems (PACIS 2016), Chiayi, Chinese Tapei*.

Santana, J., R. Hoover, and M. Vengadasubbu (2017), "Investor commitment to serial entrepreneurs: A multilayer network analysis", *Social Networks*, Vol. 48, pp. 256–269.

Santana, J., R. Hoover, and M. Vengadasubbu (2014), "Multilayer Network Analysis of Investment Patterns", in:

Smith, D. J., R. T. Harrison, and C. M. Mason (2010), "Experience, heuristics and learning: The angel investment process", *Frontiers of Entrepreneurship Research*, Vol. 30, No. 2, p. 3.

Smith, S. W. and T. J. Hannigan (2015), "Swinging for the fences: How do top accelerators impact the trajectories of new ventures?", in: *Paper presented at DRUID*, Rome, Italy.

Spiegel, O. (2012), "How employee turnover impacts social capital and performance of companies", in: *Proceedings of the 50th annual conference on Computers and People Research*, ACM, pp. 109–114.

Spiegel, O., P. Abbassi, D. Schlagwein, and K. Fischbach (2013), "Going it all alone in web entrepreneurship?: a comparison of single founders vs. co-founders", in: *Proceedings of the 2013 annual conference on Computers and people research*, ACM, pp. 21–32.

Stone, T. R. (2014), "Computational analytics for venture finance", PhD thesis, UCL (University College London).

Strickling, J. A. (2016), "Developing Entrepreneurial Ecosystems: Integrating social evolutionary theory and signaling theory to explain the role of media in entrepreneurial ecosystems", PhD thesis, University of Tennessee.

Tarasconi, G. and C. Menon (2017), "Matching CrunchBase with patent data", *working paper*, OECD.

Tata, A., D. Laureiro-Martínez, and S. Brusoni (2016), "Don't Look Back? The Effect of Attention to Time and Self on Startup Funding", in: *Academy of Management Proceedings*, vol. 2016, 1, Academy of Management, p. 13926.

Tata, A., D. L. Martinez, D. Garcia, A. Oesch, and S. Brusoni (2017), "The psycholinguistics of entrepreneurship", *Journal of Business Venturing Insights*, Vol. 7, pp. 38–44.

Ter Wal, A. L., O. Alexy, J. Block, and P. G. Sandner (2016), "The Best of Both Worlds: The Benefits of Open-specialized and Closed-diverse Syndication Networks for New Ventures' Success", *Administrative Science Quarterly*, Vol. 61, No. 3, pp. 393–432.

Townsend, J. H. (2015), "Digital Taxonomy for Sustainability", in: *Proceedings of ICT for Sustainability 2015*, ICT for Sustainability 2015, pp. 1–11.

Townsend, J. H. (2014), "Web for sustainability: tackling environmental complexity with scale", in: *ICT for Sustainability (ICT4S 2014)*, Stockholm, Sweden.

Uenoyama, K., S. Osawa, and Y. Matsuo (2014), "Exit prediction of a venture company with feature change information on talented people", *Journal of Information Processing Society of Japan*, Vol. 55, No. 10, pp. 2309–2317.

Waldner, F., M. Zsifkovits, and K. Heidenberger (2012), "Emerging service-based business models in the music industry: an exploratory survey", in: *International Conference on Exploring Services Science*, Springer, pp. 321–329.

Wang, X. (2016), "Catering Innovation: Entrepreneurship and the Acquisition Market", in: *Finance Workshop*, http://faculty.chicagobooth.edu/workshops/finance/pdf/xinxinwangcateringinnov.pdf, Chicago Booth, Chicago, IL.

Wei, L. H. (2015), "BUZZ: The Impact of Positive User Sentiment on Startup Company Valuations in the USA", Under the direction of Timothy Bresnahan, Honors Thesis, Stanford University.

Werth, J. C. and P. Boeert (2013), "Co-investment networks of business angels and the performance of their start-up investments", *International Journal of Entrepreneurial Venturing*, Vol. 5, No. 3, pp. 240–256.

Wu, A. (2016), "Organizational resource assembly in technology ventures", PhD thesis, Wharton School, University of Pennsylvania.

Xiang, G., Z. Zheng, M. Wen, J. I. Hong, C. P. Rosé, and C. Liu (2012), "A Supervised Approach to Predict Company Acquisition with Factual and Topic Features Using Profiles and News Articles on TechCrunch.", in: *ICWSM*.

Yan, J., S. Xiao, C. Li, B. Jin, X. Wang, B. Ke, X. Yang, and H. Zha (2016), "Modeling Contagious Merger and Acquisition via Point Processes with a Profile Regression Prior", in: *IJCAI*.

Zeng, X., Y. Li, S. C. Leung, Z. Lin, and X. Liu (2016), "Investment behavior prediction in heterogeneous information network", *Neurocomputing*, Vol. 217, pp. 125–132.

Zhang, C., E. Chan, and A. Abdulhamid (2015), "Link Prediction in Bipartite Venture Capital Investment Networks", CS224-w report, Stanford.

Zhao, X. (2016), "Cold-Start Collaborative Filtering", PhD thesis, University College London.

Zhao, X., W. Zhang, and J. Wang (2015), "Risk-hedged venture capital investment recommendation", in: *Proceedings of the 9th ACM Conference on Recommender Systems*, ACM, pp. 75–82.

Zhong, H., C. Liu, X. Lu, and H. Xiong (2016a), "To be or Not to be Friends: Exploiting Social Ties for Venture Investments", in: *Data Mining (ICDM), 2016 IEEE 16th International Conference on*, IEEE, pp. 699–708.

Zhong, H., C. Liu, J. Zhong, and H. Xiong (2016b), "Which startup to invest in: a personalized portfolio strategy", *Annals of Operations Research*, pp. 1–22.

Zhou, H., P. G. Sandner, S. L. Martinelli, and J. H. Block (2016), "Patents, trademarks, and their complementarity in venture capital funding", *Technovation*, Vol. 47, pp. 14–22.

Notes

[1] http://www.kauffman.org/microsites/state-of-the-field/topics/finance/equity/venture-capital visited on 25th January, 2017.

[2] Crunchbase is presented in its website as “Crunchbase is the destination for discovering industry trends, investments, and news about businesses, from startups to the Fortune 1000.

[3] http://about.crunchbase.com/ visited on September 11th, 2017.

[4] https://about.crunchbase.com/products/the-crunchbase-difference/ visited on September 11th, 2017.

[5] Crunchbase each company is tagged with several (up to 14) category group tags; to keep things simple, the graphs in this Section are based on the first tag only. The group "commerce and shopping" has been renamed "retail" for visualization purposes.

[6] OECD (2015) stresses that, whereas the quality and availability of aggregate data on venture capital have improved considerably in recent years, international comparisons remain complicated, because of the lack of a standard international definition of venture capital, and the diverse methodologies employed by data compilers.

[7] A patent application covers this activity (Oreif, 2015)