UNIVERSITY OF BIRMINGHAM | BIRMINGHAM BUSINESS SCHOOL

Birmingham Business School
Discussion Paper Series

Using Crunchbase to explore innovative ecosystems in the US and UK

Tom Kemeny

Max Nathan

Bader Almeer

2017-01

Using Crunchbase to explore innovative ecosystems in the US and UK

Tom Kemeny, University of Southampton*
Max Nathan, University of Birmingham*
Bader Almeer, University of Southampton

* Corresponding author: m.nathan@bham.ac.uk, t.e.kemeny@soton.ac.uk

Abstract

Innovative, high-technology activities are seen as motors of development, with knock-on effects throughout their local economies. Such activities tend to organise networks that span beyond atomized firms, creating 'ecosystems' of mutual dependence as well as competition. However, such ecosystems remain poorly understood, which in turn constrains the effectiveness of any policy response. This first-steps paper uses the unique, user-generated Crunchbase dataset to fill some of these gaps. With rich information on founders, workers, products and early stage investment activity, Crunchbase has great potential for ecosystem understanding. Like many 'big data' resources, however, Crunchbase requires cleaning and validation to make it suitable for robust analysis. We develop a novel approach to gapfill location data in Crunchbase, exploiting DNS/IP address information, and run a series of tests on a raw sample of 225,000 company-level observations covering the US, UK and Canada. We provide initial descriptive results, and set out steps for further research.

Keywords

cities, clusters, technology, innovation ecosystems, big data, Crunchbase

Acknowledgements

Thanks to Przemek Zientala for outstanding research support, Dani Arribas-Bel and Juan Mateos-Garcia for code, and to Crunchbase and OpenCorporates for data. Thanks also to participants at Uddevalla 2016 for comments. This research is funded through a University of Southampton AAIR grant. This paper represents the views of the authors, not the funders or data providers. Any errors or omissions are our own.

WORK IN PROGRESS – COMMENTS WELCOME

1. Introduction

Innovative, high-technology activities are seen as motors of development, with knock-on effects throughout their local economies (see Storper et al, 2015; Moretti, 2012; Galbraith and Hale, 2006 for reviews). Similarly, governments around the world see high-technology jobs as a foundation of growth strategies, and expend great effort to nurture the entrepreneurs that enable them. Crucially, success is thought to require networks that span beyond atomized firms, creating a system of mutual dependence as well as competition (Saxenian, 1996, 2008).

However, despite considerable supportive theory and case study empirics, we lack rigorous evidence on the generalizability of this idea beyond iconic clusters like Silicon Valley (Chatterji et al, 2013). In fact, evidence suggests that a variety of 'cluster shapes' exist; that constituent firms tend to operate a range of workflows, many of which run some way outside cluster 'boundaries'; and that wider policy and contextual factors may also play important roles (Kerr and Kominers, 2015; Bathelt et al, 2004; Saxenian, 2006). Indeed, it is often argued that inter-firm links are weakly developed in the UK (Nesta, 2013), though this is rarely backed up with systematic evidence. As a result, the broader basis for policy intervention is unclear, and traditional cluster policies have a poor success rate (Nathan and Overman, 2013; Duranton, 2011).

These knowledge gaps affect the prospects for effective policy design and implementation, issues particularly salient now. Since 2008, there has been a reawakening of interest in industrial policy, especially strategies that can encourage high-value / high-tech activity: for example, the EU's Smart Specialisation agenda, the Regional Innovation Clusters programme in the US, or the 'Tech City' initiative in the UK (Foray et al 2012, National Science and Technology Council, 2012; Tech City UK / NESTA, 2016). A better understanding of the determinants of economic vitality in technology ecosystems would usefully build policy and scientific knowledge.

Some key unanswered questions include:

- How locally interconnected are tech clusters outside of Silicon Valley?
- Are local interconnections an independent driver of firm performance?
- Do networks among 'elite' top team members produce different outcomes than those held among more typical employees?
- What non-local connections are most salient – for example, flows of early stage finance, founder / worker movement, B2B collaborations?

Answering these questions requires robust research designs that can isolate meaningful sources of success and failure. It also depends on high quality data that is able to capture the full range of local firms and institutions inside ecosystems. At present, we lack these data. In the US and UK, public administrative microdata tends to be limited in terms of the kinds of information it yields about firms. Commercial credit-rating agencies like Capital IQ or Dun & Bradstreet are common alternative sources, yet these have blind spots, especially in terms of their coverage of the startups and scaleups that typically engender the strongest scholarly and policy interest. For instance, capturing all the high-technology firms in the San Francisco Bay Area that are available in Capital IQ up to 2009, yields a sample of nearly 5,000 organizations. These organizations have a median start year of 1999, and the cutoff for the 90th percentile of the data is 2005. In short, these data systematically fail to capture nascent entrepreneurial activity.

Qualitative evidence confirms this sample bias. Referring to D&B's DUNS numbers, a serial technology entrepreneur based in Silicon Valley with whom we spoke told us:

> *"…the main reason to get a DUNS number is so that other businesses can run credit checks against you. As a new venture backed company you lack any of the typical markers for creditworthiness, which means a DUNS number isn't very useful. There are specialized actors like Silicon Valley Bank and Square 1 bank that lend to startups in the form of venture debt, bridge loans, etc. But they typically base decisions on other factors (calibre of investors in the company,*

> *cash remaining/burn) and also take small equity stakes as part of their lending."* (personal communication, April 21, 2016)

Our ultimate aim is to build knowledge about the economic value of social networks in these ecosystems. But the initial gap to be filled is to construct a data source that is up to that task – that captures venture-based startups that lack the paper trail of larger, more established concerns. This paper documents initial steps towards this end.

Our strategy begins with Crunchbase (CB), a uniquely comprehensive online crowdsourced platform describing workers and firms involved in high technology activities around the world. CB captures information about companies, company founders, employees and investors in the tech industry, boasting information on more than 650,000 individuals in more than 400,000 firms involved in over 200 countries. Crunchbase provides significantly more coverage and reach of ecosystem activity than conventional datasets, particularly serial entrepreneurship and investment activity; has a rich structure which covers multiple actors, not just workers and firms; and has a flexible design which – arguably – helps represent real-world complexity better than standard employer-employee panels. Crunchbase is thus a potentially hugely valuable resource for economic geographers and those working on local / regional innovation systems.

However, like other 'big data' resources, CB presents researchers with challenges to overcome, including missing information; data quality, and implicit sample issues (Einav and Levin, 2013; Nathan and Rosso, 2015). A key issue for geographers is missing location information, which is blank for 31% of companies in our raw data.

This first-steps paper has three aims. First, we develop a novel strategy for gapfilling locational information for organisations in Crunchbase, using DNS and IP lookup information, and perform extensive cleaning and validation on the raw dataset. Second, we provide some initial descriptive statistics for the US, the most richly populated country in the dataset. Third, we outline the next steps in the project – including at-scale validation of the organisation-level data, (using OpenCorporates and other sources), matching in individual, investor and funding round information. We lay out a high-level research agenda using the completed relational dataset.

The paper makes a number of contributions. To date, a tiny handful of studies have used Crunchbase for academic analysis (Morelix, 2016) these have focused on the investment layer of the dataset, rather than the economic geography / ecosystem issues we look at here. We make substantive contributions to cleaning and improving raw Crunchbase data, developing a dataset that is suitable for serious research. We generate new and highly policy-relevant findings from this data.

More broadly, the project joins a small but growing number of studies that use ‘frontier’ datasets (Feldman et al, 2015) to analyse digital technology activities, or specific sector/product verticals within the digital economy (for example Mateos-Garcia et al, 2014; Williams and Currid, 2014; Nathan and Rosso, 2015; Tech City UK / NESTA, 2016; Bernini et al, 2016).

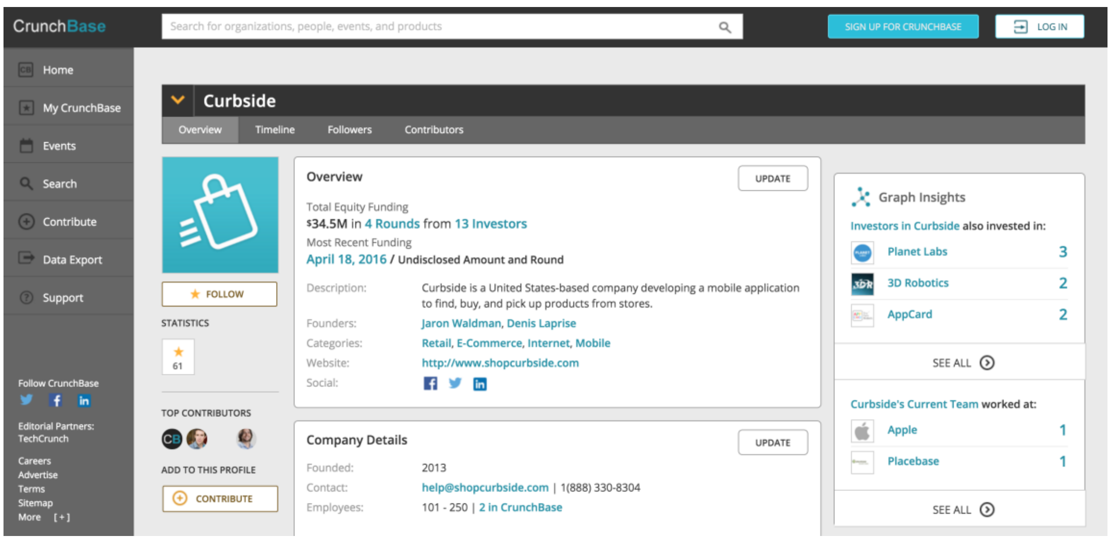
The remainder of this paper is organized as follows. Section 2 describes the Crunchbase data, and its pros and cons for use in innovation ecosystems research. Section 3 lays out our data validation and gap-filling strategy. Section 4 presents initial descriptive statistics. Section 5 concludes by describing future steps in data assembly and some possible research questions.

2. The underlying data: Crunchbase

Crunchbase (www.crunchbase.com, hence CB) was founded in 2007, by Mike Arrington, co-founder of TechCrunch, a web portal dedicated to high-technology activities, and which hosts major technology-centered events, most notably the Disrupt contest. Crunchbase was initially created to track firms that were written about on the TechCrunch industry news site, and has since evolved into a large crowdsourced source of information about technology-oriented firms, workers, investors and investments. It aims to resemble a Wikipedia of high-technology activity, albeit in a more structured format.

As of April 2016, the CB data contains records on over 650,000 individuals in more than 400,000 firms involved in over 200 countries. The information is organized as a relational dataset whose primary elements are organizations and individuals. Organizational variables include history, jobs, investments, products, liquidity events, and management teams. Data on individuals include work history, education, and multiple roles in various organizations over time.

Figure 1. Example of organization-specific webpage on Crunchbase



Source: Crunchbase, accessed 30/05/2016.

Figure 1, above is a screenshot capturing the topmost part of an organization page. It shows a sample of relevant information, regarding funding, industry type, headquarters location, as well as some prior work history for employees in the firm.

Note that CB is offered as a single, rolling cross-section of data, but has many time-specific events (such as company formation or liquidation, employees joining or leaving, financial reporting, investment rounds) which are ‘panellisable’, either within CB (at organisation / individual level) or via matching to other company-level data.

The investment information diverges somewhat from the crowdsourced nature of the data, in that CB's Venture Program tracks and validates investment information from various sources, including partnerships with a large number of venture-oriented investors, ensuring both timeliness and data quality. Crunchbase (2015) suggests that "many of these fundings were never picked up by news sources, and even those that made it into the press lacked the detail CrunchBase provides."

All of this represents a strong starting point, with some evident advantages over databases like D&B. First, Crunchbase provides information on ecosystem activity, in particular investment and funding activity that is not covered in conventional administrative or business datasets. This extra *content* is an important feature of 'big' data sets more generally (Einav and Levin, 2013). Second, Crunchbase contains information on a range of actors in an ecosystem; founders, companies, individual and institutional investors, and employees (as well as some coverage of universities and other educational institutions). This takes it beyond conventional employer-employee datasets, giving significant extra richness in *structure*. Third, and relatedly, Crunchbase allows individuals to be founders, employees, or investors – or any combination of the three at any point in time. This *flexibility* provides a natural representation of the multiple, complex roles individuals can play in real-world ecosystems (Saxenian, 1996, 2008).

But CB suffers from some quality issues common to other frontier datasets that are crowdsourced and/or 'in the wild'. Fundamentally, these datasets are likely to incomplete even if very large; but because they lack an explicit sampling frame, the further work is needed to understand data coverage and structure, in order to have confidence in the results of any analysis (Nathan and Rosso, 2015).

Most seriously, even though CB aims to be comprehensive, it may fail to include some important players in high-technology ecosystems. This problem is likely to be more serious in the early years of CB, where organizations and individuals appear in the data only if an individual has taken the initiative to add them; CB now has an in-house staff which quality-checks user-provided information and adds its own entries to the dataset.

Nevertheless, we cannot rule out the possibility that some important but low-profile actors are not present in our data. Our analysis should therefore be taken as providing a lower bound on the true level of ecosystem presence / activity.

Second, even if firms or individuals are present in the data, some information about these actors may be missing. Most pertinently for our analysis, just under a third of the organizations lack locational information (table 1). The absence of country-level information is very strongly related to the absence of all locational information (city, region etc). In this paper, we develop a novel strategy to gapfill this data by using information from DNS/IP lookups.

Table 1. The distribution of registered organizations by country in Crunchbase.

Country	Organizations	Relative Frequency
AUS	3,443	1%
AUT – CAF	3,330	1%
CANADA	**7,372**	**3%**
CHE – FSM	20,417	7%
GREAT BRITAIN	**14,996**	**5%**
GEO – IND	12,138	4%
IRL	2,588	1%
IRN – URY	23,208	8%
UNITED STATES	**112,075**	**38%**
UZB – ZWE	1,093	0%
Unknown	**91,089**	**31%**
Total	292,897	100%

Note: Authors' calculations, based on data extracted from Crunchbase during September 2015.

Third, and related to this, location and other information may be (un)reported for strategic (or spurious) reasons. For example, even if CB coverage of investments is better than any other dataset, some small-scale funding events may not be reported by firms; conversely, firms may strategically report investments as part of a strategy to attract further funding (Morelix, 2016).

Conversely, there are no barriers to individuals entering information on the platform. As with Wikipedia, anyone can add or edit information on the website. As CB has grown in size and reach, this generates incentives for firms and individuals to be part of the dataset whether or not their activity is relevant. Some actors may use a CB entry as part of a search engine optimisation (SEO) strategy to drive traffic to their site, even if that site's content has no relevance to Crunchbase itself. Such irrelevant or low-value entries may be systematically more likely not to provide location data, or other information that would allow cross-checking. Alternatively, some companies and individuals in the tech sector may be privacy-conscious and be unwilling to provide more than bare-bones information to the public.

Third, some information in CB may be inaccurate. Crunchbase now claims to validate information contributed by users, but it is not clear how extensive this is (or whether some information can be validated straightforwardly.

In this paper we take some initial steps towards disentangling these various channels that affect coverage and sampling frame. At this stage, we perform a series of checks on CB organisation-level data, both via the DNS/IP lookup process and other rule-based cleaning routines, to determine data coverage and quality – as well as gap-filling. In future versions of this paper, we draw on high quality administrative data sources to validate organisation-level information.

3. Dataset construction

Our initial aim is to produce a database of organizations that are active in technology-related activities in technology hubs within the US, Canada, and the UK. We start from a list of Crunchbase organizations extracted from Crunchbase in June 2014, consisting of 225,532 organizations with location information for the United States, Canada, UK or unknown.[1] Our data assembly tasks are 1) to understand the coverage and structure of the organisational level of the Crunchbase dataset – to make the implicit sample explicit, as far as we can – and 2) to fill gaps in the organisation-level data as completely as possible.

[1] De-duplicated data. Since that time, the database has grown considerably, though most of this growth has been organizations with either no locational information, or with locations assigned outside of the US, the UK and Canada.

Given our research questions, we are particularly focused on understanding the extent of, and reasons for, missing location data. We thus consider that an initial bifurcation in the CB data is between company-level observations that have locational data and those that do not. Given the discussion in the previous section, we suggest that entries with locational information are more likely to be valid observations. We test this assumption later by matching the organisations against secondary data. 134,443 organisations (59.6% of our starting data) have location information.

We also suggest that some subset of the 91,089 organisations *without* location information are also valid, but we have no prior expectations on how many. We test *this* assumption with a combination of cleaning / validation routines on the raw data, followed by matching against secondary data.

3.1 Organizations with pre-existing location information

For organizations with nonmissing locational information, Table 2 describes the 20 cities containing the most organizations. The distribution roughly conforms to expectations, with the caveat that the city field can contain locations that are properly part of larger regional economies. For instance, the list contains both San Francisco and Palo Alto, which together belong in the broader Bay Area metropolitan region containing Silicon Valley.[2]

[2] Our preferred scale for analysis is the metropolitan area, which is defined as a functionally-integrated economic unit, typically measured using commuting patterns. For a discussion of the appropriateness of this scale for understanding the organization of economic activities in space, see Storper et. al (2015).

Table 2. The top 20 cities by concentration of registered organizations.

City	Country	Organizations
New York	USA	9,592
London	UK	7,692
San Francisco	USA	7,681
Chicago	USA	3,264
Los Angeles	USA	2,953
Seattle	USA	2,193
Toronto	CAN	2,175
Austin	USA	1,867
Dublin	IRL	1,824
Paris	FRA	1,801
Boston	USA	1,755
San Diego	USA	1,608
Washington	USA	1,604
Bangalore	IND	1,585
Madrid	ESP	1,535
Singapore	SGP	1,478
Palo Alto	USA	1,434
Atlanta	USA	1,401
Berlin	GBR	1,381
Mumbai	IND	1,311

Note: Authors' calculations, based on data extracted from Crunchbase during September 2015.

This aggregation issue notwithstanding, one might assume that this list of cities ought to remain relatively unchanged after accounting for organizations with missing locational information, on the basis that such missing information is likely to be missing at random in terms of geographical origin. Hence, while the number of organizations will increase, the technology hubs identified in Table 2 are likely to remain among the largest.

3.2 Organizations with missing location information

Almost a third of organisations in our Crunchbase sample have missing location information. Missing locational information can mean absent country, city or region – though missing information in one such field is very strongly correlated with missing values in the others. Our workflow for these organisations is as follows:

A. Validating the list of organizations with missing locational information using URL information;
B. From those that pass stage A, conduct DNS and IP lookups to determine location information;
C. Explore any remaining non-disclosive DNS/IP addresses.

Initial validation

Form the set of organisations with missing location data, we first identify a subset of ‘valid’ organizations. Each organisation in Crunchbase provides a website address or Uniform Resource Locator (URL). We use this to perform an initial validation. We discard invalid / dead / missing URLs; non US, Canada or UK-based URLs, and non-English language sites. We also set aside (for now) websites that are non US / Canada / UK- hosted; use a third party platform such as Blogger or Wordpress; or, based on content, are personal websites rather than those for businesses. Taken together, these steps remove XXXX observations from the data.

DNS / IP Lookups

For organisations that pass this first set of filters, we use the Domain Name System (DNS) to extract location information. The DNS system is a hierarchical, decentralised naming system for computers, websites, and other services connected to the internet.[3] Crucially, running a DNS lookup on websites will give us the underlying IP (Internet Protocol) address, which should in turn provide identifying name / location information for the IP address-holder.[4]

For example, https://en.wikipedia.org has the IP address 208.80.154.224, which in turn provides zipcode, city, state and country information for the Wikimedia Foundation, owners of Wikipedia.[5]

[3] https://en.wikipedia.org/wiki/Domain_Name_System#DNS_resource_records, accessed 16 May 2016.
[4] https://en.wikipedia.org/wiki/IP_address, accessed 16 May 2016.
[5] DNS and IP lookups done through https://www.whatismyip.com/, 16 May 2016.

Starting with URLs in Crunchbase, we use DNS and IP lookups, at scale, to ascertain location information for CB companies. The working assumption is that IP address holders share locations with the companies whose websites they own. This may not always be the case, however. In future versions of this paper we will run a sensitivity check using trading information from company websites to determine the extent and nature of any error.

Table 3. Distribution of locations after DNS lookup.

Country2	Freq.	Percent	Cumulative
CA	**3,906**	4.87	4.87
UK	**5,032**	6.27	11.14
US	**34,993**	43.60	54.74
Other	23,154	28.85	83.59
n/a	13,181	16.42	100.00
Total	80,266	100.00	

Authors' calculations.

Table 3 shows the distribution of country location information after the DNS lookup exercise. In future versions of the paper we will cross-check a sample of these firms using URLs and website content: for now, in Section 5 we present descriptive results for these firms separately from the rest of the CB data.

As an initial sense check, we can see that just under 55% of formerly missing observations are from Canada, the US and UK, with the largest single group from the US (in line with the overall distribution of country obs in Table 1). Overall, we achieve a gapfill rate of over 83% on the missing data, which is more than satisfactory. We add 43,931 entries to our data, bringing the US-Canada-UK sample size up to 178,374. The remaining 13,181 missings represent 7.4% of our data, giving us location coverage of over 90%.

<u>Non-disclosive IP addresses</u>

Table 4 provides a breakdown of the 14,000-odd company observations for which DNS/IP lookups provide no location data. We can see that the vast majority of these use third party services to protect this information: typically the location given is that of the third party rather than the client company. As discussed in section two, there are various reasons why firms might do this: they could simply be publicity-shy; operating in a sector which demands secrecy; engaged in irregular / illegal activity; or URLs could represent shell companies where there is no economic activity to speak of. Again, in future versions of the paper we will cross-check a sample of these firms using URLs and website content.

Table 4. Distribution of organisations with missing DNS information.

Registration Private	6086
WhoisGuard Protected	1621
Domain Admin	1566
PERFECT PRIVACY, LLC	1519
Whois Agent	1014
Domain Administrator	851
PRIVATE REGISTRANT	503
DOMAIN PRIVACY SERVICE FBO REGISTRANT	435
Oneandone Private Registration	368
Total	13,693

Authors' calculations.

4. Initial descriptive statistics

This section sets out some (very) preliminary descriptives for the USA, the best-populated country in the Crunchbase dataset.

Table 5 shows the 20 US metropolitan statistical areas (MSAs) and consolidated MSAs (CSAs) with the largest numbers of CB companies. Not surprisingly, we can see a spiky distribution of activity, with twin peaks in the San Francisco Bay Area and New York region.

Table 5. Top 20 US Metros and CSAs in terms of numbers of organizations present in Crunchbase

Name of Metro or CSA	Organizations
San Jose-San Francisco-Oakland, CA	18159
New York-Newark, NY-NJ-CT-PA	13989
Los Angeles-Long Beach, CA	9371
Boston-Worcester-Providence, MA-RI-NH-CT	4952
Chicago-Naperville, IL-IN-WI	4752
Washington-Baltimore-Arlington, DC-MD-VA-WV-PA	4481
Seattle-Tacoma, WA	3460
Denver-Aurora, CO	2437
Miami-Fort Lauderdale-Port St. Lucie, FL	2278
San Diego-Carlsbad, CA	2272
Dallas-Fort Worth, TX-OK	2272
Atlanta--Athens-Clarke County--Sandy Springs, GA	2163
Austin-Round Rock, TX	1959
Philadelphia-Reading-Camden, PA-NJ-DE-MD	1847
Houston-The Woodlands, TX	1436
Phoenix-Mesa-Scottsdale, AZ	1370
Portland-Vancouver-Salem, OR-WA	1161
Minneapolis-St. Paul, MN-WI	1113
Detroit-Warren-Ann Arbor, MI	1091
Salt Lake City-Provo-Orem, UT	1089

Note: Area definitions built from 2013 CBSA and CSA definitions, using crosswalks from the Missouri Data Census Center’s Geographic Correspondence Engine

For selected regions, Table 6 compares the number of organizations present in Crunchbase to those available in Capital IQ, a conventional commercial dataset which is claimed to have among the most comprehensive cross-sectional coverage on entrepreneurial firms available in the United States. The Capital IQ data capture firms that have received bank, private equity or venture capital financing. Data from this source has been filtered to include firms described as operating in ‘information technology’ and ‘life science’ fields – broad umbrellas that ought to roughly correspond to those organizations that fit with Crunchbase’s target base of organizations. We can see that in raw form, Crunchbase has hugely higher coverage than Capital IQ’s data. Of course, not all businesses in Crunchbase will have received early stage funding (something we will check in future versions of this paper).

Table 6. Crunchbase Coverage in Selected Metros versus Capital IQ

Name of Metro or CSA	Orgs in Cruncbase	Orgs in Capital IQ
San Jose-San Francisco-Oakland, CA	18159	4837
Boston-Worcester-Providence, MA-RI-NH-CT	4952	2478
Seattle-Tacoma, WA	3460	917
San Diego-Carlsbad, CA	2272	1025
Phoenix-Mesa-Scottsdale, AZ	1370	420

Note: Firms in Capital IQ are identified as those in either 'information technology' or 'life sciences' industries.

Table 7. Top 20 US Metros and CSAs in terms of numbers of organizations validated through DNS lookups

Name of Metro or CSA	Organizations
Phoenix-Mesa-Scottsdale, AZ	6539
San Jose-San Francisco-Oakland, CA	2143
Los Angeles-Long Beach, CA	1994
New York-Newark, NY-NJ-CT-PA	1832
Jacksonville-St. Marys-Palatka, FL-GA	1618
Seattle-Tacoma, WA	1532
Boston-Worcester-Providence, MA-RI-NH-CT	837
Philadelphia-Reading-Camden, PA-NJ-DE-MD	691
Denver-Aurora, CO	679
Miami-Fort Lauderdale-Port St. Lucie, FL	659
Salt Lake City-Provo-Orem, UT	642
Chicago-Naperville, IL-IN-WI	635
Washington-Baltimore-Arlington, DC-MD-VA-WV-PA	583
Dallas-Fort Worth, TX-OK	417
San Diego-Carlsbad, CA	361
Atlanta--Athens-Clarke County--Sandy Springs, GA	334
Houston-The Woodlands, TX	266
Austin-Round Rock, TX	230
Minneapolis-St. Paul, MN-WI	220
Portland-Vancouver-Salem, OR-WA	216

Note: Area definitions built from 2013 CBSA and CSA definitions, using crosswalks from the Missouri Data Census Center's Geographic Correspondence Engine

Table 7 presents additional organizations that can be assigned to locations using DNS lookups. We can see that many of the same locations feature in this table as in Table 5, but the distribution of activity across space has some differences.

As discussed in section 3, if the organisations with missing location information were a random sample of all companies in CB, we would expect the same distribution of activity in Tables 5 and 7. Further versions of this paper will investigate the differences, including tests of the statistical significance of the differences in distributions, as well as cross-checks on the quality of the location information and validity of the companies. For this reason we currently present these results separately from those shown in Tables 5 and 6.

5. Next Steps

This paper introduces the Crunchbase dataset as a research tool to understand technology clusters and ecosystems, as well as the actors (founders, firms, workers, investors) operating in and around them over time. Crunchbase represents a potentially very rich information resource; like many 'frontier' datasets, however, it requires cleaning and validation to make it suitable for robust academic work. This paper develops an extensive set of cleaning/testing routines for the organisation level of CB, most notably the use of DNS/IP address information for company location. Preliminary descriptive work for the US suggests, first, a recognisable geography of economic activity in tech sectors, and second, the extensive coverage advantage that CB enjoys over conventional datasets deployed in research to date.

The paper presents the very first steps in our analysis, and there are a number of next steps. As discussed throughout the text, we have a range of further validation tests still to run on the organisation-level data. Beyond this, we will also be matching companies in CB against high-quality administrative datasets, such as Companies House in the UK (accessed through the OpenCorporates API). This matching process will also provide us with a useful vector of additional variables, including information on company directors and financial performance. Further ahead, we will clean individual level data in CB, combining this with the organisation level information to create a full founder-worker-firm-investor relational dataset.

This resource will help us answer a number of important questions about the nature and shape of technology ecosystems, including:

- How locally interconnected are tech clusters outside of Silicon Valley?
- What non-local connections are most salient – for example, flows of early stage finance, founder / worker movement, B2B collaborations?

We will also be able to look at issues of cluster 'performance', separating out ecosystem-level features and individual / group-level factors:

- Are local interconnections an independent driver of firm performance?
- Do networks among 'elite' top team members produce different outcomes than those held among more typical employees?

Finally, we will look for relevant policy changes – such as the introduction of Seed EIS tax breaks in the UK – that can be seen as shifters on ecosystems. Our data should provide us with a strong basis to assess the impact of these policies.

References

Bathelt, H., A. Malmberg and P. Maskell (2004). "Clusters and knowledge: local buzz, global pipelines and the process of knowledge creation." Progress in Human Geography 28(1): 31-56.

Bernini, M., M. Barbera, S. Addison, R. Mulhall, M. Nathan, P. Ramirez and N. Sambin (2016). Industrial Clusters in England. Report for BIS. London, NIESR.

Crunchbase (2015) CrunchBase's Venture Program Members Are Making Startup Data Better Than Ever. Downloaded April 19, 2016, from: https://info.crunchbase.com/category/venture-program/

Chatterji, A., Glaeser, E. L., & Kerr, W. R. (2013). *Clusters of entrepreneurship and innovation*. National Bureau of Economic Research, Working Paper 19013.

Duranton, G. (2011). California dreamin': The feeble case for cluster policies.*Review of Economic Analysis*, *3*(1), 3-45.

Einav, L. and J. D. Levin (2013). The Data Revolution and Economic Analysis. National Bureau of Economic Research Working Paper Series No. 19035. Cambridge, MA, NBER.

Feldman, M., M. Kenney and F. Lissoni (2015). "The New Data Frontier." Research Policy 44(9): 1629-1632.

Foray, D., J. Goddard, X. G. Beldarrain, M. Landabaso, P. McCann, K. Morgan, C. Nauwelaers and R. Ortega-Arguilés (2012). Guide to Research and Innovation Strategies for Smart Specialisation (RIS3). Brussels, DG Regio.

Galbraith, J. K., & Hale, J. T. (2006). American inequality: from IT bust to big government boom. *The Economists' Voice*, *3*(8).

Kerr, W. and S. Kominers (2015). "Agglomerative Forces and Cluster Shapes." Review of Economics and Statistics 97(4): 877-899.

Mateos-Garcia, J., H. Bakhshi and M. Lenel (2014). A Map of the UK Games Industry. London, NESTA.

Morelix, A. (2016). "The future of entrepreneurship data – getting to know CrunchBase." Retrieved 16 May, 2016, from http://www.kauffman.org/blogs/growthology/2016/02/the-future-of-entrepreneurship-data-getting-to-know-crunchbase.

Moretti, E. (2012). *The new geography of jobs*. Houghton Mifflin Harcourt.

Nathan, M., & Overman, H. (2013). Agglomeration, clusters, and industrial policy. *Oxford Review of Economic Policy*, *29*(2), 383-404.

Nathan, M. and A. Rosso (2015). "Mapping digital businesses with Big Data: some early findings from the UK " Research Policy 44(9): 1714-1733.

National Science and Technology Council (2012) A national strategic plan for advanced manufacturing. Executive Office of the President, Office of Science and Technology Policy.

Saxenian, A-L. (1996). *Regional advantage*. Harvard University Press.

Saxenian, A-L. (2006). The New Argonauts: Regional Advantage in a Global Economy. Cambridge, MA, Harvard University Press.

Storper, M, Kemeny, T., Makarem, N, and Osman, T. (2015) *The rise and fall of urban economies: Lessons from San Francisco and Los Angeles*. Stanford University Press.

Tech City UK and NESTA (2016). Tech Nation 2016: Transforming UK industries. London, TCUK.

Williams, S. and E. Currid-Halkett (2014). "Industry in Motion: Using Smart Phones to Explore the Spatial Network of the Garment Industry in New York City." PLoS ONE 9(2): e86165.