Contents lists available at ScienceDirect

Research Policy

journal homepage: www.elsevier.com/locate/respol

Mapping digital businesses with big data: Some early findings from the UK

Max Nathan[a,*], Anna Rosso[b,**]

[a] National Institute of Economic and Social Research and London School of Economics, UK
[b] National Institute of Economic and Social Research, UK

ARTICLE INFO

Article history:
Received 31 December 2014
Accepted 16 January 2015
Available online xxx

JEL classification:
C55
C81
L63
L86
O38

Keywords:
Big Data
Text mining
ICTs
Digital economy
Industrial policy
Firm-level analysis

ABSTRACT

Governments around the world want to develop their ICT industries. Researchers and policymakers thus need a clear picture of digital businesses, but conventional datasets and typologies tend to lag real-world change. We use innovative 'big data' resources to perform an alternative analysis for all active companies in the UK, focusing on ICT-producing firms. Exploiting a combination of observed and modelled variables, we develop a novel 'sector-product' approach and use text mining to provide further detail on key sector-product cells. We find that the ICT production space is around 42% larger than SIC-based estimates, with around 70,000 more companies. We also find ICT employment shares over double the conventional estimates, although this result is more speculative. Our findings are robust to various scope, selection and sample construction challenges. We use our experiences to reflect on the broader pros and cons of frontier data use.

1. Introduction

This paper uses novel 'big data' sources to expand our understanding of digital businesses in the UK. We produce alternative counts of ICT-producing firms and set out key descriptive characteristics. We then draw on this experience to critically reflect on some of the opportunities and challenges presented by big data tools and analytics for economic research and policymaking.

Information and Communications Technologies (ICTs) – and the 'digital economy' they support – are of enduring interest to researchers and policymakers. Digital sectors and firms are the subject of much analysis both at the organisational level (Bloom et al., 2012; Bresnahan et al., 2002) and in the growth field. Human capital and innovation shape long term economic development (Lucas, 1988; Romer, 1990); high value-added sectors such as ICT make direct contributions to national growth, as well as indirect contributions through spillovers and supply chains (Audretsch and Feldman, 1996; Moretti, 2012).

National and local government are thus keen to exploit the growth potential of digital businesses. Given the recent resurgence of interest in industrial policy across many developed countries (Aghion et al., 2013; Aiginger, 2007; Block and Keller, 2011; Harrison and Rodríguez-Clare, 2009; Mazzucato, 2011; Rodrik, 2004), there is now substantial policy interest in developing stronger, more 'competitive' digital economies. For example, the UK's new industrial strategy agenda (Cable, 2012) combines horizontal interventions with support for seven key sectors, of which the 'information economy' is one (Department for Business Innovation and Skills, 2013). The desire to grow high-tech clusters is often prominent in the policy mix – recent examples include the UK's Tech City initiative, Regional Innovation Clusters in the US and 'smart specialisation' policies in the EU (for a review see Nathan and Overman, 2013).

Real-world features of an industry tend to evolve ahead of any given industrial typology. For researchers, these data challenges

[*] Corresponding author at: National Institute of Economic and Social Research, 2 Dean Trench St, London SW1P 3HE, UK. +44 207 222 7665.
[**] Corresponding author. Present address: Department of Economics, Management and Quantitative Methods (DEMM), University of Milan, Via Conservatorio 7, 20122 Milan, Italy.
E-mail address: m.nathan@niesr.ac.uk (M. Nathan).

http://dx.doi.org/10.1016/j.respol.2015.01.008
0048-7333/© 2015 Elsevier B.V. All rights reserved.

present particular barriers to understanding the extent and nature of ICT production, where the pace of change can be very rapid. Data coverage is often imperfect, industry typologies can lack detail, and product categories do not closely align with sector categories. For policymakers, these information gaps feed through into policy gaps, which can limit the ability to design effective interventions.

To tackle these issues we use an innovative commercial dataset developed by Growth Intelligence (hence Gi). This covers the entire population of active UK companies, and deploys an unusual combination of public administrative data, observed information, and modelled variables built using machine learning techniques. We use this off-the-shelf material to develop a novel 'sector-product' mapping of ICT firms. We also text-mine elements of the underlying raw data to explore key sector-product cells. We run these analyses on a benchmarking sample of companies that allows direct comparisons of conventional and big data-driven estimates. The differences are non-trivial: in our alternative estimates we find that the 'ICT production space' is around 42% larger than SIC-based estimates, with around 70,000 more companies. We also find employment shares over double the conventional estimates, although this result is more speculative.

This proof of concept exercise highlights both affordances and limitations of big data-driven analysis. This is critically important for the research community, as the use of non-traditional/unstructured sources, and scraping/mining/learning tools, is growing rapidly in the social sciences (Einav and Levin, 2013; King, 2013; Varian, 2014). Enthusiasts point to huge potential in closing knowledge gaps, and taking research closer to the policy cycle. Sceptics highlight potentially limited access and relevance of these 'frontier' datasets. We talk through issues of access and relevance, as well as coverage, reliability, quality and working practices that researchers are likely to encounter.

The paper is structured as follows. Section 2 sets out a basic analytical framework. Section 3 introduces the Growth Intelligence dataset and other data resources, and outlines potential pros and cons of 'big data' approaches. Sections 4 and 5 detail our sample construction and mapping strategies. Sections 6 and 7 give descriptive results. Section 8 concludes.

2. Framework

2.1. Definitions

The 'digital economy' is an economic system based on digital technologies (Negroponte, 1996; Tapscott, 1997). This is an interlocking set of sectors (industries and firms), outputs (products and services, and the content these are used to generate), and a set of production inputs used at varying intensities by firms and workers across all sectors (OECD, 2011, 2013). We focus on the production side, and map both industries and outputs. We ignore inputs, as it is now hard to think of any economic activity where digital inputs do not feature (Lehr, 2012; OECD, 2013).

The standard OECD/UN definitions of digital producer activity are detailed product/service groups identified by an expert panel: which are then aggregated to less detailed 4-digit standard industry codes (SICs) (OECD, 2011).[1] That is, the definition moves from fine-grained to rougher grained, and is typically one-dimensional. By contrast, we are able to use industry and product information for our alternative mapping and analytics, as we explain in Section 5 below.

The OECD's three main ICT producer groups are a) information and communication technologies (ICT), covering computer manufacture, IT and telecoms networks and services and software publishing; b) digital content, covering digital/online activities in music, TV, film, advertising, architecture, design, and e-commerce; and c) wholesale, leasing, installation and repair activities in both ICT and content 'space'. In this paper we focus on the production of ICT goods and services, rather than content developed using these tools and platforms. Specifically, we are interested in the producer sectors delineated in the UK Department of Business' 'information economy strategy' (Department for Business Innovation and Skills, 2012, 2013). We refer to firms in these industries as 'information economy businesses'.

The boundaries of the UK information economy are still a matter of debate. Some analysts prefer a very narrow definition including only ICT manufacturing; conversely, some industry voices want a much broader approach that includes manufacturing, services and supply chain activity (such as wholesale, retail, installation and repair). We need to take these different opinions into account: we therefore take ICT services and manufacturing as our base case (see Table 1), and show that our results are robust to narrower and broader starting sets.[2]

In an earlier paper (Nathan and Rosso, 2013) we conduct exploratory analysis on both ICT and digital content activities. The latter is substantially harder to delineate in sector terms, not least because most content sectors are rapidly shifting from physical to multi-platform, online and offline outputs (Bakhshi and Mateos-Garcia, 2012; Foord, 2013) and because many product categories bleed across sector boundaries (see below).

2.2. Data challenges

Counting information economy businesses is challenging, particularly when conventional administrative datasets are used. In the UK there are three principal issues.

The first issue is data coverage. The main UK administrative source for firm-level data is the Business structure database (BSD) (Office of National Statistics, 2010, 2012). However, the BSD only includes firms paying UK sales tax and/or those with at least one employee on the payroll. The BSD covers 99% of all UK enterprises, but for sectors with large numbers of start-ups and small young firms – such as the digital and information economies, or nanotech – coverage will be substantially poorer.

The second issue is industry code precision. SICs are designed to represent a firm's principal business activity, but also aggregate information about inputs and clients (Office of National Statistics, 2009). As the OECD (2013) has noted, SICs can be too broad to describe new industries. For this reason, firm counts for 'other' or 'not elsewhere classified' based SIC cells are often very large, even at the most detailed five-digit level. In the 2011 BSD, for example, the second largest ICT cell is 'Other information technology service

[1] We use the most recent agreed definitions available at the time of writing, as developed by the OECD Working Party on Indicators for the Information Society (WPIIS). WPIIS agrees product lists using UN Central Product Classification (CPC) codes, then crosswalks these onto SIC 2007 4-digit cells. See OECD (2011) for detail.

[2] We use the whole UN/OECD set of digital economy SIC4 codes as a starting point for our analysis, then crosswalk these to 5-digit level and make some adjustments for the information economy in a UK context. BIS have not formally defined a set of SIC codes for the information economy, but the department's internal working definition is all of SIC3 cells 58.2, 61, 62 and 63 (personal communication, 28 November 2013). Following consultation with BIS we exclude the SIC5 cells 71121 ('engineering design activities for industrial processes and production') and 71122 ('engineering-related scientific and technical consulting activities') specified by the OECD (personal communication, 2 December 2013). Conversely, we exclude the BIS-specified cells 63910 ('news agency activities') and 63990 ('other information service activities not elsewhere classified') because they are included in the UN/OECD list of content sectors, rather than ICT production. Our robustness checks cover ICT services only (excluding all the sectors in the ICT manufacturing, code 26) and a broader set of SICs comprising manufacturing, services and supply chain activity. See Section 6.

Please cite this article in press as: Nathan, M., Rosso, A., Mapping digital businesses with big data: Some early findings from the UK. Res. Policy (2015), http://dx.doi.org/10.1016/j.respol.2015.01.008

Table 1
ICT products and services. List of SIC 2007 codes.

	ICT manufacturing
26	*Manufacture of computer, electronic and optical products, of which we use*
26110	Manufacture of electronic components
26120	Manufacture of loaded electronic boards
26200	Manufacture of computers and peripheral equipment
26301	Manufacture of telegraph and telephone apparatus and equipment
26309	Manufacture of other communication equipment
26400	Manufacture of consumer electronics
26511	Manufacture of electronic measuring, testing equipment not for industrial process control
26512	Manufacture of electronic process control equipment
26513	Manufacture of non-electronic measuring, testing equipment
26514	Manufacture of non-electronic process control equipment
26701	Manufacture of optical precision instruments
26702	Manufacture of photographic and cinematographic equipment
26800	Manufacture of magnetic and optical media
	ICT services
58	*Publishing activities, of which we use*
58210	Publishing of computer games
58290	Other software publishing
61	*Telecommunications, of which we use*
61100	Wired telecommunications activities
61200	Wireless telecommunications activities
61300	Satellite telecommunications activities
61900	Other telecommunications activities
62	*Computer programming, consultancy and related activities, of which we use*
62011	Ready-made interactive leisure and entertainment software
62012	Business and domestic software development
62020	IT consultancy activities
62030	Computer facilities management activities
62090	Other information technology service activities
63	*Information service activities, of which we use*
63110	Data processing, hosting and related
63120	Web portals

Source: OECD (2011), BIS (2013), authors' adjustments.

Notes: We follow the core definitions in OECD (2011) but use 5-digit not 4-digit SIC codes. In consultation with BIS we make minor adjustments for the UK context at 5-digit level: we remove 71121 and 71122 but include 62030. Following BIS (2013) we also separate out ICT services and manufacturing groups.

activities' (62090) which contains 22,444 enterprises (compared to 66,090 in 'Information technology consultancy activities', cell 62020).

A third, related issue is that products and services often cross sector boundaries. In the OECD analysis 'software publishing', SIC 5820, contains 10 product/service groups; conversely, the products 'data transmissions services' and 'broadband internet services' are present in multiple SIC cells (6110 through 6190). Cross-sector product types are even more prevalent in digital content activities (OECD, 2011).

2.3. *Can big data help?*

These data challenges highlight a more fundamental issue. Real-world industries, products and services are constantly evolving, while administrative typologies designed to describe them are essentially static with periodical revisions. This means that for any given iteration of an administrative typology, there is always a gradual divergence between the real features of a given economy and the means of representing those features in code form. In industries such as ICT, where entry barriers are low and the pace of innovation rapid, this divergence will be particularly marked.

It is for these reasons that we might turn to big data sources and techniques. 'big data' is a complex concept that needs careful specification. We follow Einav and Levin (2013), who define 'big' datasets as those that a) are available at massive scale, often millions or billions of observations; b) can be accessed in (close to) real time; c) have high 'dimensionality', that is, cover many variables including phenomena previously hard to observe quantitatively, and d) are much less structured than 'conventional' sources, such as administrative data.

The use of such datasets and associated analytical techniques – web scraping, text mining and statistical learning – is growing in the social sciences (King, 2013; Varian, 2014). Well-known examples include analysis of internet search data (Askitas and Zimmermann, 2009; Choi and Varian, 2012; Ginsberg et al., 2009) proprietary datasets, such as those derived from mobile phone networks (Lorenzo et al., 2012); and material derived from texts, both historic (Dittmar, 2011) and contemporary textual information taken from the web political speeches, social media or patent abstracts (Couture, 2013; Fetzer, 2014; Gentzkow and Shapiro, 2010; Lewis et al., 2011). Structured administrative datasets also take on 'big' features when linked together, or enabled with APIs that allow researchers to download online material. In the UK, virtual environments such as the Secure Data Service (SDS) and HMRC DataLab provide researchers with secure spaces for matching, and several government agencies are putting data online with API functionality.

In theory, big data should help us to develop much stronger measures of the extent and characteristics of digital economy businesses (and other nascent high-value sectors such as clean technology). Our dataset, for example, is built on an API-enabled 100% sample of active companies in the UK which is updated daily, and combines both public (administrative, structured) and proprietary (unstructured, modelled) layers which are matched to the base layer using firm names and other company-level details. These qualities of speed, scale and additional dimensions should help researchers to tackle the information economy evolution, measurement and mapping challenges described earlier.

Conversely, big data approaches may turn out to have important limitations for academic research. Einav and Levin (2013) discuss two of these: limits on access to proprietary datasets,

Please cite this article in press as: Nathan, M., Rosso, A., Mapping digital businesses with big data: Some early findings from the UK. Res. Policy (2015), http://dx.doi.org/10.1016/j.respol.2015.01.008

and the potentially limited relevance of much business data to public policy-focused research questions. Other issues include coverage (for instance, of companies not present in scraped/mined sources), reliability (when variables are probabilistic rather than directly observed, and when data is sampled), and overall quality (proprietary datasets may not be validated to the standards of administrative sources, or at all). Our experience highlights many of these pros and cons.

3. Data

Our main dataset is commercial company-level information provided by Growth Intelligence (growthintel.com). Growth Intelligence (hence Gi) is a London-based firm, founded in 2011, that provides predictive marketing software to private sector clients. The Gi dataset is unusual in the 'big data' field in that it combines structured, administrative data and modelled information derived from unstructured sources. The simplest way to describe the data is in terms of layers. This section provides a summary: more details are available in Appendix A.

3.1. Companies House layer

The 'base layer' is the population of active companies in the UK, which is taken from the Companies House website and updated daily. Companies House is a government agency that holds records for all UK limited companies, plus some business partnerships. (Sole traders are not covered, so to the extent that they work in ICT, our estimates are lower bounds.) Companies are required to file annual tax returns and financial statements, which include details of company directors, registered office address, shares and shareholders, company type and principal business activity (self-assessed by firms using SIC5 codes), as well as a balance sheet and profit/loss account. In some cases companies also file employee data (as part of the accounts, or when registering for small/medium-size status, which carries less stringent reporting requirements). Coverage of revenue and employment data in Companies House is limited – around 14% of the sample file revenue data, and 5% employment data, and these samples may be positively selected (as poor performers may try to avoid public filings). For this reason, descriptive results should be interpreted with some caution.

3.2. Structured data layers

Gi matches Companies House data to a series of other structured administrative datasets, such as patents, trademarks and US exports. Gi uses these structured datasets in two ways: to provide directly observed information on company activity (for example, patenting), and as an input for building modelled information about companies – for example, text from patent titles as an input to company sector/product classifications, which we discuss below.

3.3. Proprietary layers

This part of the Gi dataset is developed through 'data mining' (Rajaraman and Ullman, 2011). Gi develops a range of raw text inputs for each company, and then uses feature extraction to identify key words and phrases ('tokens'), as well as contextual information ('categories'). These are taken from company websites, social media, newsfeeds (such as Bloomberg and Thomson Reuters), blogs and online forums, as well as some structured data sources. Using workhorse text analysis techniques (Salton and Buckley, 1988), Gi assigns weights to these 'tokens', indicating their likelihood of identifying meaningful information about the company. Supervised learning approaches (Hastie et al., 2009) are then used to develop bespoke classifications of companies by sector and product type, a range of predicted company lifecycle 'events' (such as product launches, joint ventures and mergers/acquisitions) and modelled company revenue in a number of size bands. Tokens, categories and weights are used as predictors, alongside observed information from the Companies House and structured data layers.

3.4. Pros and cons of a big data approach

The Gi dataset should allow us to tackle the measurement challenges outlined in Section 2. First, compared to administrative data sources, the Gi data has greater coverage and provides substantially more information (thanks to the matched and modelled layers). Second, classifying companies by sector and product should allow us a more precise delineation of ICT producing companies. Specifically, SIC5 codes provide 806 sectors in which to place companies, but Gi's 145 sector and 39 product groups provide 5,510 possible sector-product cells, a more than six-fold increase. Being able to examine products, sectors and token-level information within sector-product cells affords additional detail than administrative sources and SICs cannot provide.

Conversely, there are some potential limitations in the Gi dataset. Most importantly, while our data is based on the population of UK companies, coverage of some elements is not comprehensive. This gives us 'sampled' elements to the dataset, but without an explicit process of random sampling to generate the data. To draw inferences from the data, therefore, we need to understand and work around coverage/non-response issues.[3]

First, coverage of online sources is imperfect. Many companies in the UK do not have a website, and not all websites can be successfully scraped due to site content or build; Gi estimates around 500,000 companies have websites and have scraped around 50% of these.[4] While 'non-scrapability' is likely random, having a website is not. Of course, a large number of companies without websites will be inactive or connected to an active enterprise that is online; we clean these 'untrue' companies out of our estimation sample (see Section 4). For the rest, Gi's modelled variables also draw on a range of online and offline sources for modelled data, which further helps deal with potential bias. Very few companies have no observed or modelled information at all: these comprise less than 0.1% of the raw data, and are dropped from our sample.

Second, while the company has conducted some validation exercises on its modelled variables (see Appendix A) Gi's core code is proprietary, which limits our availability to do forensic quality checking. However, we are able to conduct our own checks by comparing estimates derived from Gi's modelled data against those derived from directly observed information. Section 4 gives more details.

4. Building a benchmarking sample

Our raw data comprises all active companies in the UK as of August 2012, and comprises 3.07m raw observations, of which 2.88m have postcodes. From this we need to build a sample that a) corresponds as closely as possible to the underlying set of businesses, and b) allows comparisons between digital economy estimates based on SIC codes and those based on modelled big data. Our cleaning steps are as follows:

First, this 'benchmarking' sample can only include observations with both SIC codes and Gi classifications. Because around 21% of companies in the raw are missing SIC information it will therefore be smaller than the 'true' number of companies. In some cases, we

[3] We are grateful to a referee for highlighting this point.

[4] Sites which use predominantly Flash or are out of order/404 cannot be tokenised.

can crosswalk SIC fields from the FAME dataset to reduce losses. Overall, these steps reduce our sample from 2.88m to 2.85m observations.

Second, we drop all companies who are non-trading, those who are 'dormant' (no significant trading activity in the past 12 months), dissolved companies and those in receivership/administration. We keep active companies in the process of striking off, since a) most still operate and b) some will have failed to file returns but may re-emerge in the market under a different name. These steps reduce our sample to 2.556m companies.[5] We also drop holding companies from the sample, which reduces it to 2.546m observations.

Third, we build routines to identify groups of related companies, and reveal the underlying structure of businesses. Companies are legal entities, not actual firms, so this is a crucial step to avoid multiple counting in the underlying firm structure (for instance, if company A is part of company B, it may include some of B's revenue/employment in its accounts). This step is necessarily fuzzy, as we are creating 'quasi-enterprises'. We do this in two ways, both of which deliver very similar results. Our preferred approach is to group companies on the basis of name (same name), postcode of registered address (same location) and SIC5 code (same detailed industry cell).[6] Within each group thus identified, we keep the unit reporting the highest revenue (as modelled by Growth Intelligence). Note that for the purposes of benchmarking, we are required to do the industry matching on SIC code. This procedure gives us a benchmarking sample of 1.94m quasi-enterprise-level observations.[7]

We also test an alternative approach that exploits corporate shareholder information matched from FAME. The intuition is that if company A owns more than 50% of company B, A is likely to report B's revenue and employment. We drop B from the sample in these cases. This approach gives us a benchmarking sample of 1.823m observations. Headline results from this alternative approach are in line with our main results set out in Section 6.[8]

We validate our cleaning steps by comparing the size of a 'true' sample of all quasi-enterprises against counts of actual enterprises in a) the 2011 BSD and b) the 2012 UK Business Population Estimates (the most recent available at the time of writing). The BSD contains 2.161m enterprises, but excludes sole traders and many SMEs. Our 'true sample' of quasi-enterprises contains 2.460m observations as of August 2012, so the BSD figure is within 88% of this: acceptable given the differences in time and sample coverage. The BPE is a more helpful benchmark since it combines BSD enterprises with estimates for non-BSD businesses and sole traders (some of whom will be in our sample if they have registered a company). The BPE gives estimates up to January 2012; to make the comparison cleaner we estimate an August 2012 figure. We include companies, partnerships and sole traders with employees, plus 10% of other sole traders as a proxy for single-owner registered companies. This gives a January 2012 baseline of 2.36m enterprises. We then project the 2011–12 trend through to August. This gives a figure of 2.45m businesses, within 99% of our true sample estimate.[9]

We also test the robustness of our benchmarking sample structure. This is important to explore, as firms registering at Companies House assign themselves a SIC code. Companies doing novel activities not well covered in SICs might systematically select into 'not elsewhere classified' SIC bins rather than their 'true' classification. The set of information economy SICs contains quite a lot of these, which might lead to upwards bias. Conversely, self-assignment might lead to missing SICs for information economy firms, leading to undercounts.

Specifically, we compare across all five-digit SIC bins in Companies House with those in the 2011 BSD. Appendix B sets out the analysis. We find that the different population frames of the BSD and Companies House produce some differences in levels and internal structure, reflecting real differences in company and sector characteristics, such as firm age, industry structures and entry barriers. The overall distribution of Companies House and BSD SIC5 bins is well matched. Around the extremes, we find a number of 'not elsewhere classified' type bins where Companies House counts are higher than the BSD. These bins account for just over 10% of all the data, but only four out of 74 of these bins are in the information economy. Conversely, 21.5% of observations in the Companies House raw data lack SIC codes altogether. Taken together, this suggests that any Companies House processes (such as self-assignment) could be generating a small amount of upwards bias, but this is more than outweighed by the likely downwards bias produced by non-assignment.

5. Identifying ICT production activity

Our benchmarking sample comprises nearly 2m 'quasi-enterprises' classified with both SIC codes (based on company self-assessment), and Gi's sector and product categories (based on a range of observed and modelled information). We use this additional richness to develop alternative counts of information economy firms.

Our identification job is analogous to studies that seek to map a social/economic phenomenon through analysis of structured and unstructured information, both in data mining and in related fields such as bibliometrics. These studies have important differences, but share many of the same basic steps. Each begins with a given vocabulary or item set K_x describing the phenomenon X, and which is used to analyse a much larger item set, U_x, for which information about X is unknown. Items in K_x may map directly onto U_x, or common features – such as distinctive terms in both K_x and U_x – may be used to generate a mapping.

For instance, Gentzkow and Shapiro (2010) use speeches by members of the US Congress to analyse ideological 'slant' in the American media: they develop a core vocabulary of liberal and conservative politicians' most distinctive phrases, which is then mapped onto a similar vocabulary of newspaper op-ed pieces in order to estimate media affiliation. Working with patents data, Fetzer (2014) uses existing technology field codes to delineate broad spaces for 'clean' technology, then generates finer-grained technology vocabularies from patent titles and abstracts. These are

[5] Dropping non-trading companies removes 92,929 observations; dropping dormant companies removes 106,589 observations; dropping all but active and partially active companies removes 318,906 observations. Some companies may be in more than one of these categories, so sub-totals may not sum.

[6] We do not use the full company name, but we use the first if there is only one word in the name, or if the second word is some common acronym that refers to the status of the company (Limited, Ltd-Plc Company, LLP). We use the first and the second words if there are at least two words in the name, or if the third word is again an acronym as in the previous case.

[7] We test the sensitivity of this approach by matching on postcode sector (that is, the first 4/5 digits of the postcode) rather than the full postcode. This less restrictive approach would reduce false negatives (related companies that are very closely co-located but not present at exactly the same address), but might increase false positives (similarly-named but non-related companies in the same industry and neighbourhood). Results show that company counts decline in almost the same proportions across all sectors. This is reassuring, as it implies that there is nothing systematic happening in our selection process. Details are available on request.

[8] Specifically, using SIC-based definitions we have 158,810 ICT producer companies (8.17%) compared to 225,800 companies (11.62%) using the 'sector-product' approach. See Table 2 for headline comparisons.

[9] The 2.36m total includes 1.34m companies, 448,000 partnerships, 297,000 'sole proprietorships and partnerships' with employees and 271,000 sole traders without employees. We also conduct sensitivity checks including 1) 5% of sole proprietors without employees (2.253m enterprises) and 2) basing on 2009–2011 trends (2.390m enterprises). Full results available on request.

then used to resample the patents data to provide an alternative mapping of the clean technology space.

Ideally, then, we would look for a rich word- or phrase-level objective vocabulary for information economy companies, K_{ie}, which we would then map onto a corpus of company-level texts for companies. In practice, we have a category-level item set for the information economy, which is expressed in our data with SIC codes (see Section 2). And rather than raw words and phrases, we are working with a 'categorical vocabulary' of off-the-shelf sector and product categories mined by Gi (see Section 3).

5.1. Mapping strategy

Our basic mapping steps are as follows. First, we take the subsample of companies with OECD/BIS ICT products and services SIC codes, as defined in Table 1. Next, we extract the corresponding Gi sector and product classifications for those companies: this provides a long-list of 99 Gi sectors and 33 Gi product groups. We treat this as a rough cut of the true set of ICT sectors and products/services.

Following this, we refine the cut. We first use a crude threshold rule to exclude 'sparse' Gi sectors and product cells, which might be marginal and/or irrelevant to ICT sector/product space. Sparse groups are defined as those present in less than 0.2% of the long-listed observations. Removing this group of sparse cells results in a shortlist of 16 sectors and 12 product groups, which account for the majority of ICT-relevant observations.

Next, we review the sparse Gi sector and product lists in detail to recover any marginal but relevant cells. By construction, each of these cells comprises less than 0.2% of the long-listed observations.[10] The review is rule-based: specifically, we look for sparse Gi sector or product cells where the name corresponds to 1) the OECD definition of ICT products and services, or 2) BIS modifications to this list. We use the detailed OECD guidance (OECD, 2011) and Gi metadata to guide marginal decisions: we include cells that have some correspondence to the OECD-specified SIC4 or CPC group, and exclude those where no such correspondence exists. For example, we recover the sector cells 'computer network security' and 'e-learning', which feature in the OECD product list, but exclude the product cell 'hardware tools machinery', which Gi uses to designate construction tools (such as mechanical hoists).

Finally, we use this set of sectors and products to resample sector-by-product cells from the whole benchmarking sample. This creates a set of companies in 'ICT' sectors whose principal product/service is also ICT-relevant.

5.2. Identification

This 'sector-product' approach, built on a range of data sources, provides an alternative mapping of information economy firms. It should allow us to deal with false negatives in our data (via incorrect SIC coding). It should also tackle false positives, by allowing us to identify the set of companies in 'ICT' sector contexts whose main outputs (products and services) are also ICT-related, disregarding those who are not involved in digital activity. This allows us to keep those companies in (say) the mobile telecoms industry who are actually making mobile phones, and exclude those who are involved in wholesale, retail or repairs.

We then run various robustness checks. First, as outlined in Section 2, there is some disagreement about which SIC codes should be used to delineate the information economy. Sector-product results might then be endogenous to the set of starting SIC cells, rather than being driven by real differences in sector-product information. We therefore reproduce the analysis with different SIC starting sets, both a very narrow set of ICT service industries and a broader set of manufacturing, service and supply chain industry bins.

Second, our 0.2% threshold rule might still identify some irrelevant sector/product space (leading to false positives). We experiment with tighter thresholds at 0.3% and 0.5% of long-listed observations. Third, the sector-product approach might collapse to a 'sector' or 'product' analysis, if one of the Gi vectors turns out to be uninformative. In this case false positives could be included in the final estimates. We test this by reproducing the analysis with Gi sector cells alone, and Gi product cells alone.

A final worry is that our off-the-shelf Gi categories are too high-level to always provide useable information (this objection also applies to SIC codes). In our case, we are relying on the combination of sector-by-product information: but analysis using only Gi sector or product typologies, or individual sector/product cells, may be less informative. We therefore use raw token information from company websites to look inside the largest sector and product cells.

6. Results

How do conventional and big data-based estimates of ICT production differ? Table 2, below, gives headline results. Panels A and

Table 2
ICT producer counts and shares: comparing SIC and big data estimates.

	Companies	%
A. SIC 07–manufacturing and services		
Other	1,783,973	91.83
Information economy	158,810	8.17
B. Gi sector and product–manufacturing and services		
Other	1,716,983	88.38
Information economy	225,800	11.62
C1. SIC 07 – ICT services only		
Other	1,789,405	92.11
Information economy	153,368	7.89
C2. Gi – ICT services only		
Other	1,761, 811	90.68
Information economy	180,972	9.32
D1. SIC 07 – services, manufacturing & supply chain		
Other	1,748,607	90.01
Information economy	194,176	9.99
D2. Gi – services, manufacturing & supply chain		
Other	1,708,549	87.94
Information economy	234,234	12.06
E. Gi sector		
Other	1,637,606	84.29
Information economy	305,177	15.71
F. Gi sector and product – manufacturing and services (0.3% threshold)		
Other	1,744,303	89.78
Information economy	198,480	10.22
G. Gi sector and product – manufacturing and services (0.5% threshold)		
Other	1,749,376	90.04
Information economy	193,407	9.96
Total/panel	1,942,783	100

Source: Gi and Companies House data.
Note: In Panel A, SIC-defined information economy includes sectors as reported in Table 1. Other includes all the other firms. Panel B defines the information economy using Gi ICT sector by ICT product "cells", starting from the initial SIC category including both ICT services and manufacturing. Panel C defines the information economy using SIC "cells", starting from the initial SIC category including only ICT services. Panel D defines the information economy using SIC "cells" including ICT services, manufacturing and supply chain sectors. Panel E shows the count if the information economy was only defined using Gi ICT sectors. Panel F and G use different threshold rules to identify Gi ICT products and sectors.

[10] We include the following sectors: e-learning', 'computer network security', 'information services', 'semiconductors'. We include the following products: 'software web application' and 'software mobile application', but we exclude: 'hardware tools machinery'.

Table 3
SIC codes for 'additional' ICT producer companies, 16 largest cells.

Description	SIC 2007	Observations	%
Other engineering activities (not including engineering design for industrial process and production	71129	12,520	17
Advertising agencies	73110	9,166	12
Specialised Design Activities	74100	7,596	10
Engineering related scientific and technical consulting activities	71122	4,872	6.5
Technical testing and analysis	71200	2,982	4
Repair of other equipment	33190	2,918	3.9
Engineering design activities for industrial process and production	71121	2,874	3.8
Other business support service activities not elsewhere classified	82990	2,583	3.4
Manufacture of electric motors, generators and transformers	33140	1,924	2.6
Repair of machinery	33120	1,849	2.5
Installation of industrial machinery and equipment	33200	1,845	2.4
Repair of computers and peripheral equipment	95110	1,778	2.4
Wholesale of electronic and telecommunications equipment and parts	46520	1,605	2.1
Manufacture of other electrical equipment	27900	1,424	1.9
Activities of head offices	70100	1,132	1.5
Electrical installation	43210	1,115	1.5
Management consultancy activities (other than financial management)	70229	819	1.1
Retail sale of computers, peripheral units and software in specialised stores	47410	773	1

Source: Gi and Companies House data.
Note: Firms in the information economy (Gi definition) but not in the SIC code definition. The percentage refers to the percentage of firms captured using Gi definition in each SIC code excluded from the official definition (only the most relevant are reported). The information economy is defined using Gi sectors and products.

B give alternative estimates of information economy companies. SIC coding identifies 158,810 ICT quasi-enterprises, 8.17% of our benchmarking sample. By contrast, the sector-product approach identifies 225,800 quasi-enterprises, around 11.62% of the economy. That is, our big data-driven estimates are over 40% higher compared to SIC-based definitions in Panel B. Overall, this difference in headline numbers – nearly 70,000 'missing' companies – suggests the precision gain is non-trivial.

By construction, our sample includes only those companies with SIC and Gi coding, so missing SIC codes are not driving the results. Other Panels report robustness checks that explore some of the identification challenges discussed in Section 5.2. Panels C and D show the effect of changing the starting set of SIC sectors. In Panel C1 we look only at SICs covering ICT services, while in Panel D1 we use a broader definition of the information economy including SIC codes in the wider ICT value chain.[11] Panels C2 and D2 give corresponding Gi-based estimates. If our main results were entirely driven by choice of the SIC starting categories, we would find alternative SIC (sector-based) counts converging to the Gi (sector-product) estimates in panel B. Even with the broadest starting set of SICs (Panel D1) we find 31,624 fewer companies than our baseline Gi estimates (Panel B) and 40,058 more companies in the corresponding Gi counts (Panel D2).

Panel E tests the effectiveness of the sector-product approach as opposed to using sector-only Gi information. We would expect the lack of granularity to produce higher estimates, which it does (305,177 versus 225,800 companies, almost 16% of the sample). (Using only the product dimension of Gi data, the share would be driven up to more than 50%.)[12]

The last two panels shows estimates using more conservative threshold rules to exclude sparse Gi sectors and products cells: 0.3% and 0.5% in panels F and G, respectively. Again, we would worry if the resulting counts approached the initial sector-based estimates in Panel A (indicating that the sector-product approach delivers little precision over SIC sectors). Information economy counts and shares drop as expected, but even in the most conservative specification (Panel G) we find 34,597 additional companies using sector-product cells compared to SIC sector codes.

6.1. *What kind of additional companies?*

Our sector-product method gives us a large number of companies that we would not treat as ICT producers using SIC codes alone. Table 3 maps these quasi-enterprises back onto their SIC codes, for the 18 largest SIC cells.

Note that some of these SIC bins (33200 and 95110, 4.8% of the total) would be included in our 'broad-based' set of information economy SIC codes, as discussed above. Another 8% (33190, 43210, 46250, 47410) also fit into 'value chain space'. However, more than 26% of the omitted companies classify themselves in the 'Other engineering activities', 'Engineering related scientific and technical consulting activities' and 'Engineering design activities for industrial process and production' bins (respectively, 71129, 71122, 71121); and another 20% define themselves in the advertising agency or specialised design sectors (such as 73110 or 74110). While these companies are in 'non-ICT' sector contexts, in other words, their principal products and services put them into the information economy.

6.2. *Internal structure*

Next, we take a closer look at the internal structure of our Gi-based ICT producer estimates. Tables 4 and 5 provide headline counts, shares and revenue information for the largest sector-product cells. Each table 'rotates' the cells to indicate sector information (Table 4) and product information (Table 5), so that companies in (say) the 'computer games' sector could have any of the principal outputs listed in the products table – and companies whose principal product is (say) 'consultancy' might be in any of the sector cells in the sector table. (Together, all of these combinations would form a 378-cell matrix too large to show here.)

More than 46% of companies in Table 4 are located in information technology, almost 15% in computer-related sector groups (computer software, hardware, games), around 20% in engineering and manufacturing sectors, and a further 7% in telecommunications.

[11] Panel C covers ICT services only (see Table 1). Panel D includes all the SICs in Table 1 plus 33120 (Repair of machinery), 33190 (Repair of other Equipment), 33140 (Repair of Electrical Equipment), 33200 (Installation of industrial machinery and equipment), 95110 (Repair of computer and peripheral equipment), 71129 (Other engineering activities), 71122 (Engineering related scientific and technical consulting activities), 71121 (Engineering design activities for industrial process and production).

[12] Results available on request.

Table 4
Total number of firms in the information economy by Gi sectors.

	Observations	%	Revenues	
			mean	median
Computer_games	2,585	1.14	1,793,241	3181.5
Computer_hardware	3,514	1.56	2,473,394.4	83,803
Computer_networking	3,902	1.73	2,135,848.7	93,784
Computer_network_security	226	0.1	13,223,530	1,027,628
Computer_software	23,455	10.39	1,433,080.5	35,564
Consumer_electronics	2,074	0.92	11,125,476	97,584
Design	10,049	4.45	753,104.63	53,798.5
e_learning	347	0.15	4,496,422.4	320,504.5
Electrical_electronic_manufacturing	17,319	7.67	3,696,466.6	93,784
Information_services	823	0.36	5,018,562.8	182,405
Information_technology	104,768	46.4	995,039.69	38,364
Internet	2,954	1.31	6,527,924.2	195,958
Marketing_advertising	11,038	4.89	3,695,790.4	42,077
Mechanical_or_industrial_engineering	27,326	12.1	1,145,004.3	93,784
Semiconductors	183	0.08	64,762,995	1,323,417
Telecommunications	15,237	6.75	16,347,362	78,165
Total	225,800	100	2,723,804	57,282

Source: Gi and Companies House data.
Note: Observations by sector when defining digital economy using Gi ICT products and sectors (manufacturing and services). Revenues are Gi modelled revenues.

Table 5 shifts the focus to products and services. Most of the companies are providing some kind of consultancy service (67%), offering software development (8.8%), care and maintenance (7%), web hosting (just under 3%) or some sort of broadband or software related services. This analysis of within-structure starts to give a sense of what firms in the information economy are offering in the product/service mix. The main impression is of technological diffusion outside computer hardware and software into other industries: notably engineering and manufacturing sectors, but also digitised consultancy and business services. As we discuss in Section 2, our analysis is likely capturing evolving activities, products and services that do not show up easily in administrative classifications. To build on this, we use text mining to uncover more information about the largest cells, ‘information

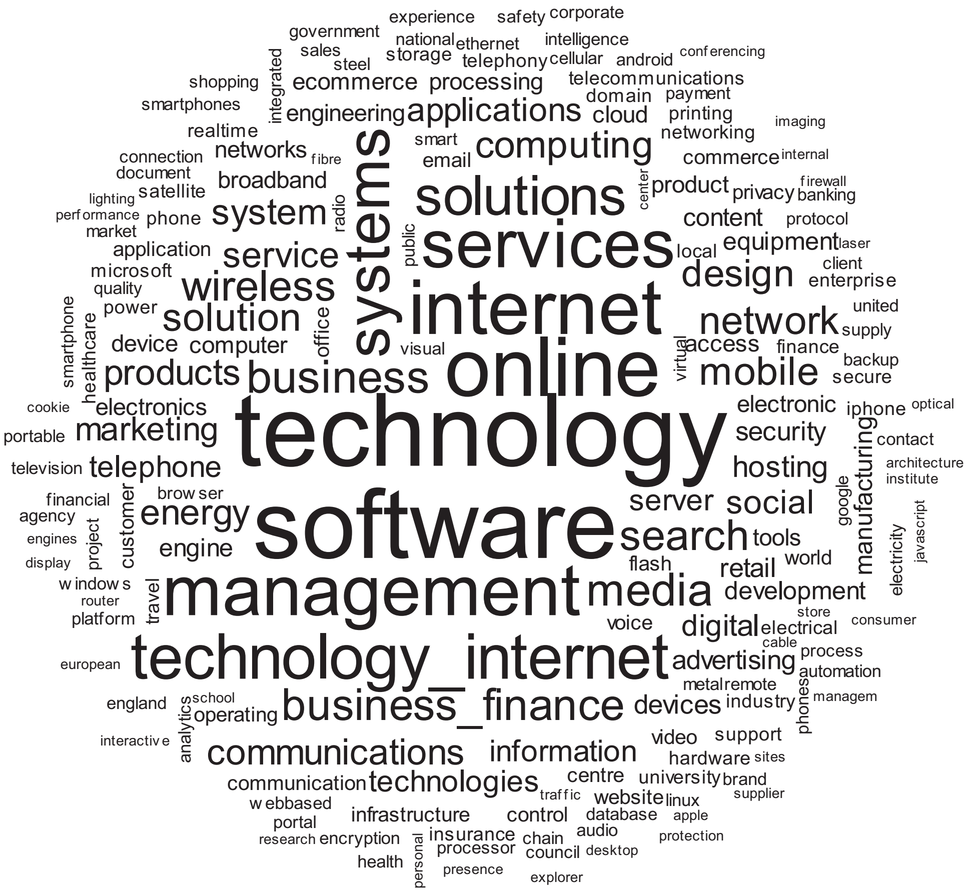

Fig. 1. Most frequent words in ICT producer activity space: web tokens.

Source: Gi data.

Table 5
Total number of firms in the information economy by Gi product.

	Revenues			
	Observations	%	Mean	Median
Advertising_network	1,663	0.74	3,163,943	341,687
Broadband_services	8,628	3.82	4,050,860	18,369
Care_or_maintenance	15,663	6.94	1,300,043	54,642
Consultancy	151,408	67.05	2,009,348	57,802
Education_courses	645	0.29	6,321,385	434,989
Electronics	15,180	6.72	12,953,757	174,866
Peer_to_peer_communications	1,300	0.58	13,120,439	0
Software_desktop_or_server	5,237	2.32	547,854	13,171
Software_mobile_application	31	0.01	2,953,207	1,426,606
Software_web_application	43	0.02	14,577,145	409,863
Custom_software_development	19,981	8.85	1,012,336	34,814
Web_hosting	6,021	2.67	1,392,615	34,765
Total	225,800	100	2,723,804	57,282

Source: Gi and Companies House data.
Note: observations by product when defining digital economy using Gi ICT products and sectors (manufacturing and services). Revenues are Gi modelled revenues.

technology' and 'consultancy'.[13] To do this we use raw text data (tokens) and contextual information (token categories) taken from websites and news feeds (see Section 3). Gi reports 12 token categories of which we use four – organization, product, technical term and technology.[14] Tokens are assigned values representing the relevance of the token for the company, ranging from 0 to 1: we include only tokens whose company relevance is above 0.2. We harmonise token content by putting all the words into lower case, removing punctuation, and removing words that may refer to legal status of the company: 'ltd', 'plc', 'llp', 'company'. We also remove stopwords.[15]

In Fig. 1, we report, in a word cloud, the most popular words across the whole set of information economy firms when the sector is defined using the Growth Intelligence classification as per Panel B in Table 2. For reasons of space, we only show the words that appear at least 2,000 times in the whole sample of the information economy. We end up with a list of 363 words where the total number of words is 1,839,014. The larger and darker the word is, the more frequent it appears in the sample of companies in the information economy that report token information. For example, the most frequent word is 'technology' which appears 70,139 (4% of the total number of words) in the sample, the word 'technology_internet' is very frequent and appears 40,286 times (2%).

In Table 6 we report a list of the most popular words (48% of total number of words) in the information economy with the total number of appearances, and the relative share given by the number of appearances over the total number of words (1,839,014) (Panel A). We also show the same information for the companies in the sector 'information technology' and product cell 'consultancy' (Panel B), 'consultancy' products across all ICT sectors (Panel C) and 'information technology' firms providing any ICT products (Panel D).[16]

The word that appears the most in Panels A–C is 'technology'; for the IT sector it is 'software'. It represents 4% of the total number of words in the complete ICT producer space (Panel A), 7% in the sample including only companies doing IT and consultancy, 5% in Panel C (consultancy) and 6% in Panel D (IT), while 'software' in IT appears 7% of the times. Note that the distribution across panels within these information economy cells is very similar, and despite relatively sparse some words appearing only 1% of the time, we observe a high density in the same words across all the four panels.

We might worry that these are simply terms which appear on any company's website. To understand how distinctive these words are, then, we also look at the word distribution amongst the sectors in the rest of the economy (Fig. 2). Interestingly, we find that the most relevant words are not the same: the words that are denser in ICT production space are under-represented in other activity spaces.

7. Characteristics of ICT and non-ICT businesses

This section provides descriptive analysis of companies' age, inflows, revenues and employment.

7.1. Age

Table 7 reports the average age of ICT and non-ICT companies in the benchmarking sample.[17] Using SIC codes, ICT companies around almost three years younger than non-ICT firms; using sector-product definitions the difference shrinks slightly. Notably, median differences between ICT and non-ICT firms are substantially smaller; the median ICT firm is now about a year younger than its non-ICT counterpart, whichever definition is used.

In Table 8, we show the distribution of companies by age groups. This share can easily be interpreted as a survival rate.[18] Panel A uses SIC code definitions; panel B uses sector-product groups. In Panel B, around 66% of 'ICT' companies are under 10 years old, 33% under five years, 14.4% under three years old and around 1% less than a year old. This compares with 64.6%, 30.6%, 13.8% and 2.2%, respectively, in the rest of the economy. Analysing the distribution using SIC codes (Panel A) shows very similar patterns. Start-ups, defined here as companies less than three years old, are slightly more common in amongst ICT producers than in the rest of the economy.

On the face of it, these findings are surprising: the popular image of the ICT industry is of start-ups and very young companies. Our evidence, however, suggests that there is no reason to think that the ICT companies are more ephemeral than the other companies. Our analysis of inflows, below, also tells a similar story.

7.2. Inflows

Fig. 3 shows the inflow of our companies into the economy, comparing inflows of companies into ICT production (dashed line) with

[13] We have run some statistical tests in order to check how different the sample of tokens is in comparison to the whole sample of companies (benchmarking sample), both in terms of within sectoral distribution (share of ICT companies) and in terms of characteristics to conclude that the information economy sector when defined using SIC codes is around 8% (similarly to the whole sample). When defined using Gi definition the information economy is slightly overrepresented in the token sample, it is likely to be the case as Gi algorithms puts more weight to the presence of web tokens when assigning a company to a sector. Sectors/products where token information is better (in particular it is likely that ICT sectors do have a better internet coverage) are likely to be larger. In terms of characteristics, ICT companies in the token sample are likely to be older, and have higher revenues. All the differences are statistically significant.

[14] The full list of token categories is: Company, Contact Details, Entertainment Event, Location, Operating System, Organization, Person, Position, Product, Technical Term, Technology, TV Show.

[15] http://jmlr.org/papers/volume5/lewis04a/a11-smart-stop-list/english.stop, accessed 15 December 2013.

[16] In the subsample of companies with tokens we have 3716 companies doing IT and consultancy, 12,556 companies providing some consultancy service in any ICT sectors, and 4296 in the information technology sector (any ICT products).

[17] We report estimates only for our preferred definition, panels A and B of Table 2.

[18] We have looked at companies that dissolved in year 2012, which have dropped from the selected sample. We have looked at the distribution of companies by incorporation year and by sector and also in this case, the distribution over time is similar in the ICT sectors and in the rest of the economy. This also implies that the average age is similar and it is actually higher for the digital economy sectors when using Gi definition.

companies in the rest of the economy (solid line), from 1980 to 2012. The number of ICT companies entering the economy every year has always been much smaller, but it is interesting to see that when using Growth Intelligence's classification we are able to capture a higher level of inflow over the whole period considered but in particular after the year 2000.

We also estimate the growth rate, defined as the percentage of the yearly inflow over the total existing companies and compare it across the two sectors. Results are shown in Fig. 4.

The growth rate of ICT companies has been higher than the rate in the rest of the economy in the period before the dot-com bubble which happened in year 2000, and this is even more evident when using the SIC codes. The reason why the rate is smoother in the Gi-based classification may be related to the fact that when using our alternative definition we are also capturing companies that have been in the economy for a longer period and started to produce products or provide services that we would include in the ICT definition. After the dot-com bubble, the information economy started to follow the cycles of the rest of the economy, and the growth rate even started to be lower than the rate in other sectors.

7.3. Revenue

Regular Companies House data provides relatively limited information on company revenues. Only 13.9% of the companies in our sample have reported revenues in the period between 2010 and 2012 and even a smaller percentage (8.4%) have filed revenues every year over the same period. We therefore supplement this information with Gi's modelled revenue data, which covers all of the companies in the dataset.

Table 9 sets out these two sources together. We can see from Panel A that the sub-sample of companies reporting revenues is similar to the full sample in terms of information economy shares. For this sub-sample, non-ICT companies have higher average and median revenues, but on Growth Intelligence's measures the gaps between the two groups narrow substantially. When shifting to modelled revenue, ICT firms have lower average revenue but rather higher median revenue than non-ICT firms. In Panel B, we look at 2010–2012 revenue growth for the companies who report revenues over more than one year. The first column reports the average percentage growth, defined as the within-firm growth of revenues

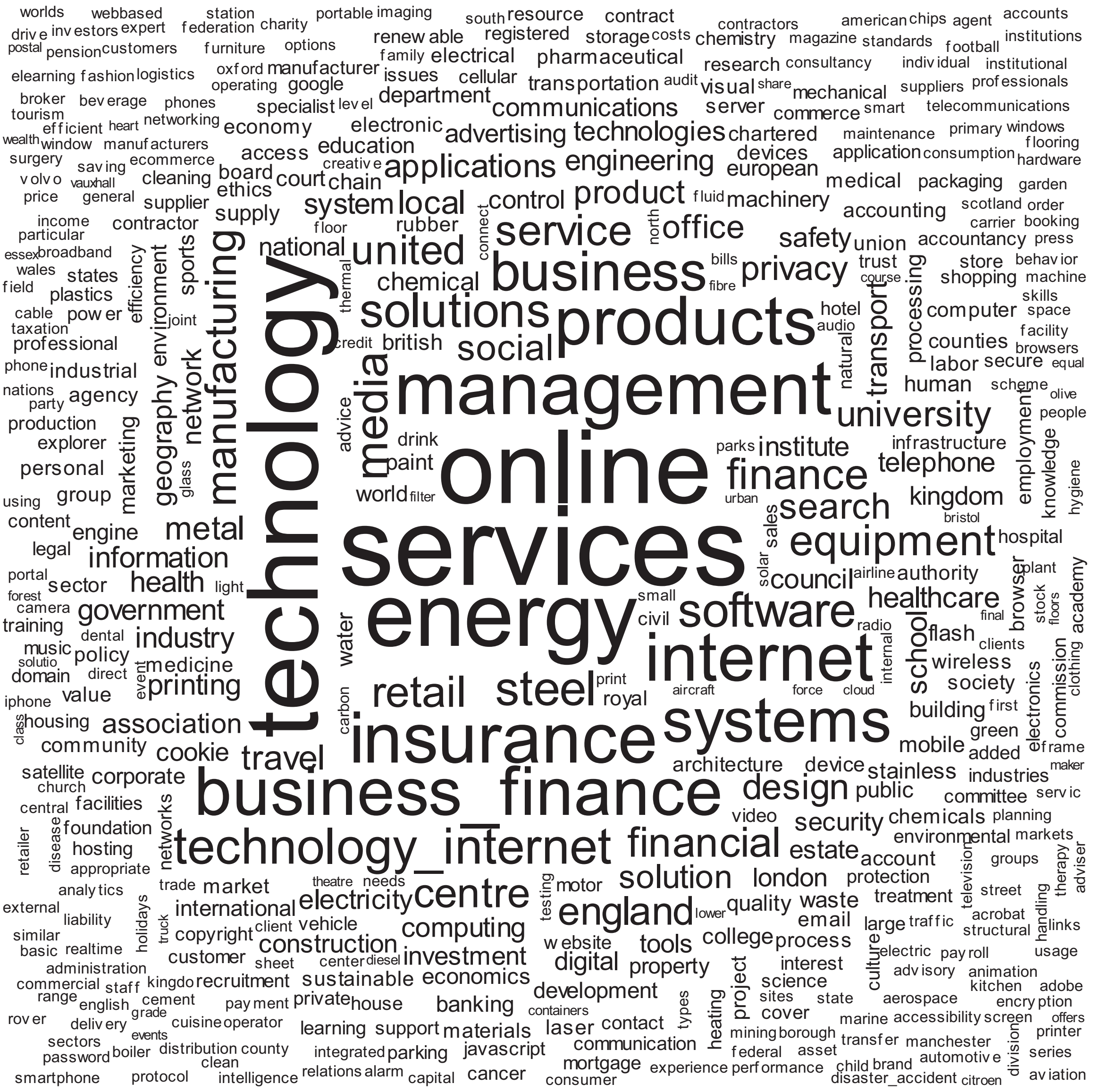

Fig. 2. Most frequent words in the rest of the economy: web tokens.

Source: Gi data.

Table 6
Word distribution within sectors.

	A. ICT MF and services		B. IT & consultancy		C. Consultancy		D. IT	
	Words appearances	Relative share	Words appearances	Relative share	Words appearances	Relative share	Words appearances	Relative share
Technology	70,139	4%	13,874	7%	37,708	5%	16,002	6%
Software	66,063	4%	13,767	7%	35,036	4%	16,485	7%
Online	54,668	3%	7,106	4%	26,175	3%	8,465	3%
Internet	49,843	3%	6,114	3%	21,090	3%	7,423	3%
Management	47,312	3%	11,209	6%	32,027	4%	12,602	5%
Services	43,136	2%	9,658	5%	27,194	3%	10,701	4%
Technology_internet	40,286	2%	4,960	3%	18,349	2%	6,397	3%
Systems	38,195	2%	6,152	3%	17,657	2%	7,280	3%
Solutions	33,726	2%	7,599	4%	20,273	2%	8,816	4%
Business	26,851	1%	6,134	3%	18,135	2%	6,859	3%
Media	26,474	1%	3,073	2%	15,083	2%	3,835	2%
Business_finance	25,406	1%	3,581	2%	15,603	2%	4,028	2%
Search	23,731	1%	2,406	1%	10,365	1%	2,871	1%
Wireless	23,018	1%	2,032	1%	7,007	1%	2,858	1%
Solution	22,178	1%	4,678	2%	12,647	2%	5,557	2%
Mobile	21,694	1%	3,226	2%	11,079	1%	3,992	2%
Network	20,883	1%	3,656	2%	11,435	1%	4,275	2%
Computing	20,540	1%	5,251	3%	10,746	1%	6,214	3%
Design	19,387	1%	1,341	1%	7,845	1%	1,655	1%
Communications	18,990	1%	2,145	1%	11,230	1%	2,363	1%
System	18,911	1%	2,727	1%	7,998	1%	3,663	1%
Service	18,493	1%	3,410	2%	9,901	1%	3,872	2%
Energy	18,013	1%	2,340	1%	9,108	1%	2,591	1%
Products	17,627	1%	2,192	1%	7,179	1%	2,590	1%
Applications	17,477	1%	2,977	2%	7,603	1%	3,593	1%
Marketing	16,758	1%	1,404	1%	9,974	1%	1,614	1%
Social	16,033	1%	2,384	1%	9,507	1%	2,753	1%
Server	14,044	1%	2,522	1%	6,186	1%	3,467	1%
Technologies	14,002	1%	3,627	2%	8,418	1%	4,157	2%
Digital	13,656	1%	1,274	1%	5,877	1%	1,618	1%
Telephone	13,574	1%	0	0%	6,135	1%	1,210	0%
Information	13,263	1%	3,957	2%	8,748	1%	4,552	2%
Total	884,371	48%	146,776	74%	463,318	57%	174,358	70%

Source: Gi data.
Note: Word appearance refers to the number of time the word appears in the sample of companies reporting token. Relative share is computed as the number of appearances over the total number of words in the sample. Panel A reports words in the tokens in all the companies in the information economy defined including both manufacturing and service sectors. Panel B reports the words in the tokens of the companies in IT (sector cell) and consultancy (product cell). Panel C companies doing consultancy. Panel D companies in the IT sector.

averaged over the sample. On the sector-product basis, growth is higher for ICT companies (22%) than the rest of the economy (15%) – with similar results for SIC-based definitions. Median differences are rather smaller.

Table 10 takes a higher-level view of modelled revenue across the whole benchmarking sample. Average revenues for ICT firms run at around 40% of the non-ICT average for SIC definition but slightly higher on the sector-product. Looking at medians, non-ICT firms have slightly lower modelled revenue than ICT firms using both SIC and sector-product cells. Again, levels differences between means and medians are substantial, suggesting the presence of outliers.

7.4. *Employment*

Under Companies House rules, companies are only obliged to report employment data in specific cases: in our raw data, only 100,359 companies provide this information. As with revenue, this will be a selected sub-sample. We would expect companies with employees to be older and have higher revenues than those without, and this turns out to be the case: those in the employment 'set' are on average twice as old, and report average modelled revenues around 2/3 higher than the non-employment 'set'. These caveats should be borne in mind in what follows. On the other hand, tests of industrial structure suggest very similar shares of ICT and non-ICT companies and the spatial distribution of the companies across the UK is very similar, with three out of the top five locations being shared.

First we look at employees per firm. Table 11 shows average and median employees per company. As not all companies report employment in every year, we smooth the data across three and

Table 7
Age of companies, mean and median years of activity.

	Other		Information economy	
	Mean	Median	Mean	Median
SIC 07 – manufacturing and services	10.3	6.5	7.7	5.4
GI sector and product	10.3	6.5	8.4	5.7

Source: Gi and Companies House data.
Note: Age defined as years of activity since the company was incorporated.

Table 8
Distribution of companies by age groups.

	%	
	Other	Information economy
A. SIC 07 – manufacturing and services		
Up to 1 year old	2.04	2.14
Up to 3 years	13.71	16.33
Up to 5 years	30.55	35.48
Up to 10 years	64.57	67.31
B. GI sector and product		
Up to 1 year old	2.18	1.00
Up to 3 years	13.84	14.44
Up to 5 years	30.66	33.06
Up to 10 years	64.61	66.06

Source: Gi and Companies House data.
Note: Each entry represents the share of companies within each age group.

Please cite this article in press as: Nathan, M., Rosso, A., Mapping digital businesses with big data: Some early findings from the UK. Res. Policy (2015), http://dx.doi.org/10.1016/j.respol.2015.01.008

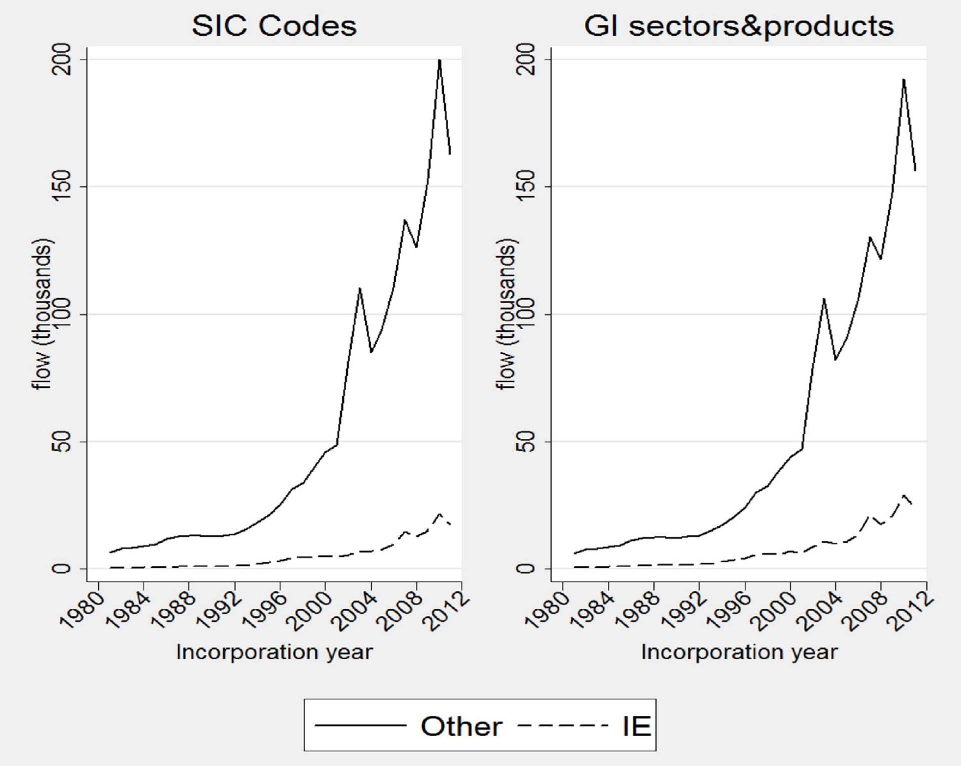

Fig. 3. Inflow of companies between 1991 and 2011. *Note*: The graphs show the inflow of active companies in each year.
Source: Gi and Companies House data.

five-year periods. Average employment counts for ICT businesses differ substantially between SIC and Gi-based definitions. Using SIC codes, non-ICT businesses are somewhat larger and ICT firms, and a little bigger than the average firms. Using sector-product definitions, ICT firms employ rather more people on average than companies in the wider economy and the average firm, especially in the 2008–2012 period. However, median differences are much smaller, with non-ICT firms consistently reporting higher worker counts. That suggests outliers explain much of the mean differences.

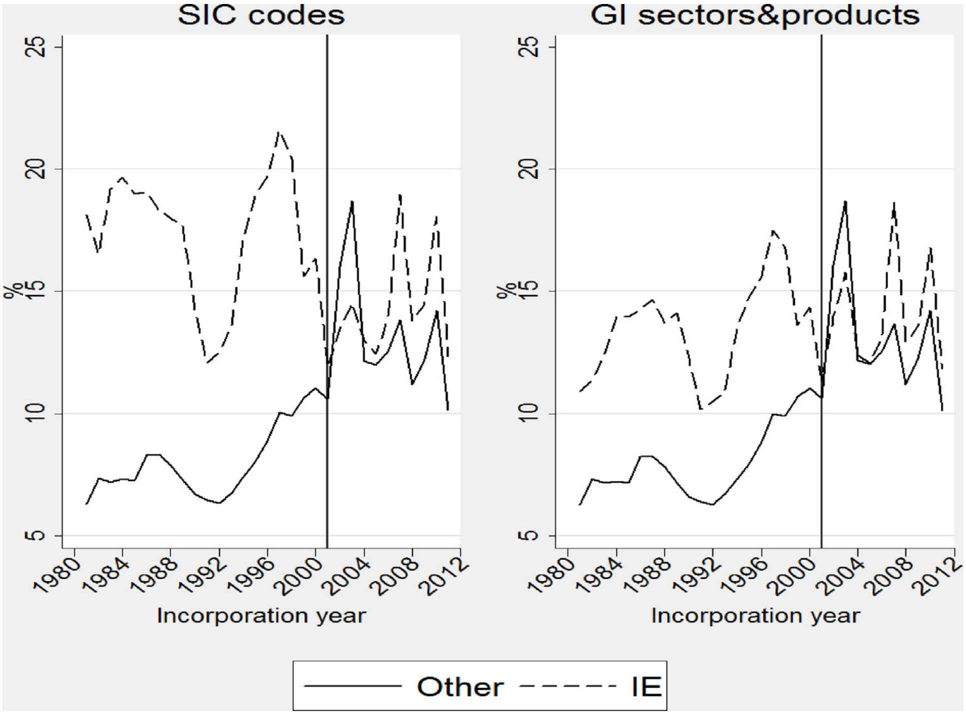

Fig. 4. Growth rate in the number of firms between 1980 and 2011.
Note: Growth rate as a percentage of number of firms entering the economy each yearover the total existing firms.
Source: Gi and Companies House data.

Table 9
Mean and median revenues and revenue growth from Companies House.

	A. Average revenues						B. Average Annual Revenue Growth			
	Companies House		Gi		Obs	Sector distribution	Companies House		Obs	Sector distribution
	Mean	Median	Mean	Median			Mean	Median		
SIC 07 – manufacturing and services										
Other	21,640,058	125,281	25,780,253	70,196	254,025	0.94	0.16	0.02	154,442	0.94
Information economy	11,658,404	97,669	13,142,859	83,073	17,593	0.06	0.23	0.05	9,402	0.06
GI sector and product										
Other	21,605,718	124,241	25,864,831	68,469	245,940	0.91	0.15	0.02	149,791	0.91
Information economy	15,130,138	106,640	16,311,935	91,240	25,678	0.09	0.22	0.05	14,053	0.09

Source: Gi and Companies House data.
Note: Companies House average revenues are averaged over the period 2010–2012. Gi revenues are computed over the same sample. For the Companies House dataset if for each company there is more than one observation, only the most recent is kept. Average annual revenue growth is computed on a smaller sample, as information for at least two consecutive years is need. The years considered are the same as above, 2010–2012.

Table 10
Gi modelled revenues by sector.

	Gi (mean and median) revenues			
	SIC 07 – manufacturing and services		GI sector and product	
	Mean	Median	Mean	Median
Other	4,945,056	45,975	4,948,276	44,611
Information economy	1,820,333	47,071	2,723,804	57,282

Source: Gi and Companies House data.
Note: Gi modelled revenues.

Table 11
Average employees per firm.

	Breakdown	Observations	Gi		SIC codes	
			Mean	Median	Mean	Median
2008–2012	Other	143,989	31.86	5	34.79	5
	Information economy		60.06	3	22.82	4
	Average		34.17	5	34.17	5
2010–2012	Other	75,927	22.35	4	23.42	4
	Information economy		32.92	3	17.99	3
	Average		23.16	4	23.16	4

Notes: Sub-sample of companies filing employment information to Companies House.

Next, we turn to ICT firms' share of all employment (for which we have information). Table 12 shows that shifting from SIC-based definitions of digital businesses to Gi definitions shifts ICT firms' employment share substantially upwards, from around 3.5% to nearly 12% of all jobs in 2008–2012, and from 3.7% to 8.92% in 2010–2012. This is as we would expect, since underlying company counts are higher in our big data-driven definitions.

8. Discussion

This paper uses innovative 'big data' resources to perform an alternative analysis of the digital economy, focusing on ICT producing firms in the UK (so-called 'information economy' businesses). Exploiting a combination of public, observed and modelled variables, we develop careful cleaning routines and develop a novel 'sector-product' mapping approach, using text mining to provide further detail. We argue that this can provide advantages over SIC codes and conventional datasets, which tend to lag rapidly evolving real-world features of these industries.

Our big data-driven estimates suggest that the count of information economy firms is around 42% larger than SIC-based estimates, with almost 70,000 more companies. We also find employment shares over double the conventional estimates, although this result is more speculative. The largest sector-product cells are in information technology (sectors) and consultancy (products); text analysis suggests software, Internet tools, system management and business/finance are particular strengths of companies in these cells. More broadly, ICT hardware, games, ICT-related engineering/manufacturing, telecoms, care and maintenance are key activities across the ICT production activity space. ICT firms are slightly younger than non-ICT firms, with a slightly higher share of start-ups; while their average revenues are lower, on some measures revenue growth for ICT firms is higher than for their non-ICT counterparts. Defined on a sector-product basis, ICT firms employ more people on average than non-ICT firms (although median differences are much smaller).

We thus find a set of companies that is larger, more established and perhaps more resilient than popular perceptions. Our analysis also suggests diffusion of digital platforms and prod-

Table 12
ICT and non-ICT employment shares.

Category	Share of all employment (%)	
	2008–2012	2010–2012
Information economy (SIC codes)	3.54	3.70
Other	96.46	96.30
Information economy (Gi)	11.75	8.92
Other	88.25	91.08

Notes: Sub-sample of companies filing employment information to Companies House.

Please cite this article in press as: Nathan, M., Rosso, A., Mapping digital businesses with big data: Some early findings from the UK. Res. Policy (2015), http://dx.doi.org/10.1016/j.respol.2015.01.008

ucts out of computer hardware and software into other parts of the economy, notably business services and engineering/high-end manufacturing. This is consistent with specific industry studies (see e.g. Nathan and Vandore, 2014), and supports our case that big data can shine a light on real-world economic shifts that are moving ahead of current administrative data and classifications.

Our results are robust to multiple validations of the core dataset and a series of robustness checks. Some care has to be taken with the revenue and employment findings, since these derive from non-random sub-samples, but Gi is able to provide some workarounds for these (such as modelled revenue).

Our experiences so far with the Growth Intelligence dataset also provides us with some valuable lessons on the pros and cons of using ‘frontier’ data for innovation research. Gi data has excellent reach and granularity and, as we have shown, provides rich detail on fast-changing parts of the economy. However, like other commercial products such as FAME, the Gi dataset is not free to academic researchers and there is no automatic right to access. Similarly, Gi’s proprietary layers are based on non-public code, ultimately limiting what validation can be done. This may limit wider replicability of the results by other teams and in other country contexts. These constraints are not unique to big data, however.

Other issues derive directly from the use of core big data tools and analytics. Web and news-based information on companies is extremely rich but is not always comprehensive, and needs to be supplemented from other sources. Data providers may throttle data drawn from APIs, which places some constraints on speed of draw-down and thus the ‘real-time’ character of some unstructured sources: in some such cases, paying for direct access to the full dataset may be a more sensible solution. At a more basic level, the use of learning routines to generate probabilistic variables is ideal for exploring aggregate patterns in very large datasets, but can become noisy when researchers wish to look at smaller blocs of the data, or when they are working with relatively few observations to start with. In this case, we shifted to using raw data for small-cell analysis.

Together, these imply broader issues for researchers and policymakers. First, researchers should carefully consider the advantages and limitations of ‘off the shelf’ big datasets, and consider developing their own bespoke information as a complement. Second, government and universities need to develop researcher capacity to generate, as well as analyse, unstructured and other frontier data resources. Third, there is a clear need for secure sharing environments where proprietary and public data can be pooled, explored and validated. In the UK, the Secure Data Service provides one potential model for such platform. Finally, and linked to this, there is a need for structured partnership projects to incentivise researchers and data providers to work together.

We suggest various avenues for future research. One is exploring co-location and clusters. Another is to use modelled events as predictors of future observed behaviour. A third is to look at determinants of growth or lifecycle events. In the last two cases, the analysis would need to be done for the sub-sample of companies that can be ‘panellised’ in the data, and would benefit from merging with administrative datasets. More broadly, this company-level data could be combined with worker-level information to explore how ICTs are changing patterns of labour use and workforce organisation.

Acknowledgements

This paper is based on research projects funded by Google and by NESTA. Many thanks to Francois Bouet, Tom Gatten, Prash Majmudar and Alex Mitchell at Growth Intelligence for data, and help with its preparation and interpretation. For advice and helpful comments, thanks also to Hasan Bakhshi, Theo Bertram, Siobhan Carey, Steve Dempsey, Juan Mateos-Garcia, Jonathan Portes, Rebecca Riley, Chiara Rosazza-Bondibene, Brian Stockdale, Dominic Webber and Stian Westlake plus participants at Birmingham University, Google, RSA, Imperial College, NARSC, NIESR, NESTA, NEMODE and TechUK workshops. Thanks also to Martin Kenney, Maryann Feldman, Francesco Lissoni and two anonymous referees. This work includes analysis based on data from the Business Structure Database, produced by the Office for National Statistics (ONS) and supplied by the Secure Data Service at the UK Data Archive. The data is Crown copyright and reproduced with the permission of the controller of HMSO and Queen’s Printer for Scotland. The use of the ONS statistical data in this work does not imply the endorsement of the ONS or the Secure Data Service at the UK Data Archive in relation to the interpretation or analysis of the data. This work uses research datasets that may not exactly reproduce National Statistics aggregates. All the outputs have been granted final clearance by the staff of the SDS-UKDA. The paper gives the views of the authors, not the funders or the data providers. Any errors and omissions are our own.

Appendix A. The Growth Intelligence dataset

19. The Growth Intelligence dataset

Growth Intelligence (Gi) is a London-based company, founded in 2011, that provides business intelligence services to largely private sector clients. The Gi dataset combines public administrative data, structured data and modelled data derived from unstructured sources. The dataset is best described in terms of layers.

19. Companies House layer

The ‘base layer’ of the Gi dataset comprises all active companies in the UK, which is taken from the Companies House API and updated daily. Under the Companies Act 2006, all limited companies in the UK, and overseas companies with a branch or place of business in the UK need to be registered with Companies House.[19] Some business partnerships (such as Limited Liability Partnerships) also need to register. There is a charge of around £100 to do this. Sole traders and business partnerships which are not LLPs do not need to register at Companies House, although they will need to file tax returns with HMRC. When they register, companies are asked to choose the Standard Industrial Classification (SIC) code which best reflects their principal business activity. Dormant and non-trading companies are also asked to include SIC information.

All registered companies must file a) annual company returns as well as b) annual financial statements (statutory accounts). Returns cover details of directors and company secretary, registered office address, shares and shareholders, as well as company type and principal business activity. There is a small charge for filing the return, which must be done within 28 days of the anniversary of incorporation. There are financial penalties for not filing the return on time: in the extreme Companies House can dissolve the company and prosecute the directors. Statutory accounts must be filed with Companies House, in addition to tax returns with HMRC. Accounts must include a balance sheet, a profit and loss account, a directors’ report and an auditors’ report. The balance sheet shows the value of company assets; the profit and loss accounts shows sales, running costs and subsequent profit / loss. Accounts must be compiled by nine months after the end of the financial year. As with returns, there are financial penalties for late filing, and possible criminal penalties for non-filing.

[19] See www.companieshouse.gov.uk for more information.

A number of companies are exempted from full filing. Limited companies that are 'small' can send abbreviated accounts consisting only of the balance sheet, and in some cases can apply for exemption from auditing. Small firms must meet two or more of the following: less than £6.5m turnover; less than £3.26m on the balance sheet; fewer than 50 employees. Some 'dormant' limited companies can also claim partial or full exemption from filing. Dormant companies are those defined as having no 'significant accounting transactions' during the accounting period in question.

Companies must inform Companies House about changes to limited companies, including directors/secretaries joining or leaving; changes to the company name, registered address or accounting dates, and where records are kept. Limited companies can request to be closed/dissolved, providing they have not traded within the last three months; not changed company name within that period; are not subject to current/proposed legal proceedings, and have not made a disposal for value of property or rights. There is a £10 charge for the striking off application. Once Companies House has accepted the application, a notice is placed in the London/Edinburgh/Belfast Gazette giving at least three months' notice of the intent to remove the company from the Register.

Companies are legal entities, and company-level observations may not always reflect the actual underlying business. We perform a number of cleaning steps to recover 'true' enterprises. These steps are discussed in detail in Section 4 of the main paper.

Structured data layers

Gi matches Companies House data to a series of other structured administrative datasets. Gi uses these structured datasets in two ways: to provide directly observed information on company activity (for example, patenting), and as an input for building modelled information about companies (for example, text from patent titles as an input to company sector/product classifications). We discuss these modelled data layers below.

20. Modelled data layers

This part of the Gi dataset is developed through data mining (Rajaraman and Ullman, 2011). Gi develops a range of raw text inputs for each company, and then uses feature extraction to identify key words and phrases ('tokens'), as well as contextual information ('categories').[20] Gi assigns weights to these 'tokens' based on likelihood of identifying meaningful information about the company. Machine learning approaches are then used to develop classifications of companies by sector and product type, predicted lifecycle 'events' and modelled company revenue. Tokens, categories and weights are used as predictors, alongside observed information from the Companies House and structured data layers.

Tokens and token categories are extracted from a range of textual sources, including company websites, news media and news feeds, blogs, plus patents and trademarks text fields. In the language of text analysis, these 'documents' form a complete 'corpus' about the universe of companies (Baron et al., 2009). Growth Intelligence use an approach based on Text Frequency-Inverse Document Frequency (TF-IDF) weights to identify the most distinctive words in each company's document set.[21] Informally, a given word will have a high TF-IDF for a given company if it a) appears in relatively few documents across the corpus, and b) appears many times when present in a given document.

Table B1
Information economy counts and shares: BSD vs Companies House 2011.

Enterprise/QE type	Freq.	Percent
BSD		
Other	2,036,557	94.22
Information economy mf + services	124,971	5.78
Total	2,161,538	
Companies House		
Other	1,722,359	91.81
Information economy mf + services	153,858	8.20
Total	187,217	

Source: BSD, Companies House.
Notes: BSD = enterprises, CH = quasi-enterprises.

For company classifications, Gi uses a supervised learning setting (see Hastie et al., 2009) for an overview of these approaches). The basic idea is to take a randomly sampled training set of observations where classifications are known, then use this to develop a machine-learnt algorithm that can accurately predict company type on the basis of observed information (but where classification is not known). Once validated on another random subsample, the tool is then used to classify the rest of the data.

Modelled revenue is generated using a machine-learnt regression. In this case reported revenue in Companies House data is used in the training set, with predictors drawn from other observed financial information, events and sector classification.

Appendix B. Comparing Companies House and BSD structures

22. Comparing Companies House and BSD structures

The benchmarking exercise in this paper involves taking raw Companies House (CH) data and cleaning it to produce 'quasi-enterprises'. We need to be confident that our estimates are accurate. To do this, we validate the level and structure of our data against the main UK administrative source, the Business Structure Database (BSD). Information in the BSD is extremely reliable and is checked against multiple sources (ONS, 2013). Firms enter the BSD when they have at least one employee on the payroll and/or have revenues high enough to charge VAT (sales tax). We look at levels and shares of SIC5 cells in CH and the BSD, across all sectors and for the 'information economy'.

There are a number of issues we need to test. First, our own cleaning steps may produce inaccuracies; in the main paper we run through a series of sensitivity tests on these. Second, the Companies House sampling frame may produce some structural peculiarities: legal entities are not necessarily active enterprises, and in sectors with low entry barriers (such as many parts of the information economy) we may see higher numbers than in the BSD. Our cleaning steps remove inactive companies so should mitigate this, but some underlying structural differences may persist. These reflect real characteristics of firms and industries, but we need to understand their nature. Third, Companies House processes may produce structural inaccuracies, particularly as firms assign themselves to an SIC code. Newly registering companies are – in most cases – very young, so may not understand the SIC system and/or fully know their main activity yet. This may lead companies to file in specific categories other than their 'true' categories. Specifically, companies might be more likely to file in uninformative 'not elsewhere classified' type SIC cells. The information economy set of SICs contains a number of these, which may bias up counts. Alternatively, companies may not provide SIC information at all. This plausibly affects

[20] Gi uses multiple techniques for matching online information to companies, including direct matches from web URLs; whois records, and Companies House numbers reported on websites.

[21] The TF-IDF approach is the workhorse method in the field (Salton and Buckley, 1988); an alternative is to use the Pearson chi2 score (see Gentzkow and Shapiro (2010) for a recent example).

Table B2
Information economy: shares and counts for component bins, 2011.

SIC5 sector name	BSD			CH		
	Freq.	Percent	Cum.	Freq.	Percent	Cum.
Manufacturing of electronic components	588	0.47	0.47	1037	0.67	0.67
Manufacturing of loaded electronic boards	360	0.29	0.76	241	0.16	0.83
Manufacturing of computers and peripheral equipment	826	0.66	1.42	791	0.51	1.34
Manufacturing of telephone and telegraph equipment	1,342	1.07	2.49	700	0.45	1.8
Manufacturing of other communications equipment	163	0.13	2.62	199	0.13	1.93
Manufacturing of consumer electronics	614	0.49	3.12	487	0.32	2.25
Manufacturing of electronic measures and tests	1,578	1.26	4.38	1,050	0.68	2.93
Manufacturing of electronic industrial process control equipment	259	0.21	4.59	512	0.33	3.26
Manufacturing of non-electronic equipment not for industrial process control	185	0.15	4.73	42	0.03	3.29
Manufacturing of non-electronic industrial process control equipment	92	0.07	4.81	20	0.01	3.3
Manufacturing of optical precision instruments	123	0.1	4.91	128	0.08	3.38
Manufacturing of photographic and cinematographic equipment	88	0.07	4.98	64	0.04	3.43
Manufacturing of magnetic and optical media	26	0.02	5	33	0.02	3.45
Publishing of computer games	111	0.09	5.09	254	0.17	3.61
Other software publishing	1,823	1.46	6.54	3,313	2.15	5.77
Wired telecomms activities	780	0.62	7.17	1,581	1.03	6.79
Wireless telecomms activities	657	0.53	7.69	1,413	0.92	7.71
Satellite telecomms activities	130	0.1	7.8	372	0.24	7.95
Other telecomms activities	5,208	4.17	11.97	7,658	4.98	12.93
Ready-made interactive leisure, entertainment software	623	0.5	12.46	2,459	1.6	14.53
Business and domestic software development	17,842	14.28	26.74	18,540	12.05	26.58
Information technology consultancy activity	66,090	52.88	79.62	65,319	42.45	69.03
Computer facilities management activities	207	0.17	79.79	2,212	1.44	70.47
Other information technology service activities	22,444	17.96	97.75	42,614	27.7	98.17
Data processing hosting and related activities	2,812	2.25	100	2,819	1.83	100
Total	1,24,971	100		1,53,858	100	

Source: BSD, Companies House.
Notes: BSD = enterprises, CH = quasi-enterprises.

companies with novel products and services, such as information economy firms, and would lead to undercounts.

Headline comparisons

The 2011 BSD contains 2.161m enterprises, but excludes sole traders and many SMEs. Our ‘true sample’ of quasi-enterprises contains 2.460m observations as of August 2012 when firms without SICs are included, so the BSD figure is within 88% of this: acceptable given the differences in time and sample coverage.

Table B1 shows the headline estimates for the two datasets. The 2011 BSD contains 2.161m enterprises, of which 5.78% (124,971 enterprises) are ‘information economy’ businesses.

In Companies House, around 1.9m ‘quasi-enterprises’ are present in 2011. Quasi-enterprises are companies that have gone through our cleaning steps (see Section 4 of the main report). 8.2% of our sample (153,858 quasi-enterprises) is in the information economy.

Table B2 gives more detail on the internal structure of the set of information economy firms, reporting counts and shares at SIC5 level. We can see that SIC bins have different shares in the two datasets. Typically these differences in shares are small, although there are some exceptions. One group consists of sectors where both counts and shares are low, such as ‘manufacturing of telephone and telegraph equipment’ (1.07% of the BSD set, 0.45% of the CH set, SIC 26301). The other group consists of larger cells, such as ‘business and domestic software development’ (14.28% of the BSD set, 12.05% of the CH set, SIC 62012); ‘information technology consultancy’ (52.88%, 42.45%, 62,020) and ‘other information technology service activities’ (17.96%, 27.7%, 62,090).

What might explain these differences? The rest of the Appendix tests possible channels.

22. *Age structures*

There are structural differences between the BSD and Companies House (Anyadike-Danes, 2011). The BSD covers 99% of businesses in the UK. But by definition, the BSD excludes firms that do not pay VAT and/or do not have employees on PAYE. For this reason it will tend to select older and more established firms than CH. Similarly, in sectors with low entry barriers – such as many information economy sectors – CH will tend to report larger numbers of observations than the BSD, but coverage in the BSD may be ‘skewed’ towards more established organisations.[22] Looking at the age structure of firms in the BSD and CH, we can see that the BSD coverage is orientated towards older firms than CH (Table B3).

Around 52% of BSD firms appear in the last 10 years (and about 17% of start-ups, defined as firms three years old or less). In contrast, 67% of CH observations are founded in the last 10 years and 21% of CH observations are start-ups. These differences are also noticeable in the information economy (Table B4). The differences are smaller for the set of firms 10 years old or less, but greater for start-ups.

We know that information economy sectors are typically characterised by low entry barriers, high levels of innovation and a lot of young firms (Department for Business Innovation and Skills, 2013). So counts/shares of such firms are likely to be higher in CH, even if estimates of sector-level employment/turnover will not differ much.

Sectoral distribution in the BSD and CH

Next we look at levels and shares for all 735 SIC5 bins, for both datasets. Manual examination reveals some trivial differences. First, around 29 CH observations have invalid SIC codes (0.0016% of the CH sample). Second, some sectors arepresent in CH but absent in the BSD, for example households as employers (including 59,194 residential property management companies, 3.17% of the CH sample); space transport (22 observations); growing citrus fruits (2), oleaginous fruits (1), and ‘gathering wild growing products’ (19). Third, holding companies are present in the BSD but not CH because

[22] In practice, these comparisons understate the true differences, since the BSD/IDBR ‘birth’ variable measures time of entry into the dataset rather than true birth year of the business.

Table B3
Age structure for all sectors, BSD vs Companies House 2011.

Birth year	Freq.	Percent	Cum.	Inverse
BSD				
2002	97,427	4.51	48.17	51.83
2003	1,04,285	4.82	52.99	47.01
2004	93,431	4.32	57.31	42.69
2005	1,05,061	4.86	62.17	37.83
2006	1,32,971	6.15	68.33	31.67
2007	1,63,062	7.54	75.87	24.13
2008	1,50,699	6.97	82.84	17.16
2009	1,71,379	7.93	90.77	9.23
2010	1,64,360	7.6	98.37	1.63
2011	35,152	1.63	100	0
Total	2,161,538	100		
Companies House				
2002	85,071	4.53	32.93	67.07
2003	1,14,892	6.12	39.05	60.95
2004	89,635	4.78	43.83	56.17
2005	98,829	5.27	49.1	50.9
2006	1,15,940	6.18	55.28	44.72
2007	1,44,991	7.73	63.01	36.99
2008	1,35,701	7.23	70.24	29.76
2009	1,65,044	8.8	79.03	20.97
2010	2,16,961	11.56	90.6	9.4
2011	1,76,397	9.4	100	0
Total	1,876,217	100		

Source: BSD, Companies House.
Notes: BSD = enterprises, CH = quasi-enterprises. BSD enterprises measured by oldest local unit year of entry into the IDBR. CH QE age measured by year incorporated.

Table B4
Age structure for information economy sectors, BSD vs Companies House 2011.

Birth year	Freq.	Percent	Cum.	Inverse
BSD				
2002	6962	3.92	42.1	57.9
2003	8199	4.61	46.71	53.29
2004	8989	5.06	51.76	48.24
2005	9903	5.57	57.33	42.67
2006	11,270	6.34	63.67	36.33
2007	17,135	9.64	73.31	26.69
2008	13,363	7.51	80.82	19.18
2009	13,574	7.63	88.45	11.55
2010	16,840	9.47	97.92	2.08
2011	3,691	2.08	100	0
Total	1,77,821	100		
Companies House				
2002	5364	3.49	29.34	70.66
2003	6577	4.27	33.61	66.39
2004	6748	4.39	38	62
2005	7288	4.74	42.73	57.27
2006	9120	5.93	48.66	51.34
2007	14,304	9.3	57.96	42.04
2008	12,309	8	65.96	34.04
2009	14,665	9.53	75.49	24.51
2010	20,969	13.63	89.12	10.88
2011	16,740	10.88	100	0
Total	1,53,858	100		

Source: BSD, Companies House.
Notes: BSD = enterprises, CH = quasi-enterprises. BSD enterprises measured by oldest local unit year of entry into the IDBR. CH QE age measured by year incorporated.

our cleaning removes them. In the BSD they comprise 14,281 observations, or 0.66% of the sample.

Fig. B1 scatters the full set of bins for both datasets and illustrates each bin's share. The overall distribution of CH and the BSD is fairly close – see the two best fit lines – although this hides some differences (in particular 'other business support activities not elsewhere classified' (9.93% of CH, 2.92% of the BSD, SIC 82990) and 'Other business services not elsewhere classified' (3.17% of CH, 1.7% of the BSD, SIC 96090). We discuss other cases below in 6.1.

Table B5
5% of SIC5 bins with largest CH–BSD differences, 2011.

SIC 2007 5-digit category	%BSD	%CH	BSD–CH
Other business support activities nec	2.92	9.93	−7.01
Residents property management	0	3.17	−3.17
Other business services nec	1.7	3.19	−1.49
Buying and selling of own real estate	0.14	1.49	−1.35
Other information technology service activities	1.04	2.28	−1.24
Activities of head offices	0.12	1.31	−1.19
Management of real estate on fee/contract basis	0.53	1.47	−0.94
Other professional, scientific and technical activities nec	1.18	2.09	−0.91
Financial intermediation nec	0.19	0.95	−0.76
Other letting and renting of own / leased real estate	1.94	2.64	−0.7
Development of building projects	1.65	2.31	−0.66
Other human health activities	0.55	1.2	−0.65
Other building completion and finishing	0.64	1.19	−0.55
Other manufacturing nec	0.24	0.73	−0.49
Information technology consultancy activities	3.06	3.49	−0.43
Construction of commercial buildings	0.71	1.11	−0.4
Other amusement and recreation activities nec	0.21	0.57	−0.36
Other information service activities	0.09	0.41	−0.32
Renting and operating of housing association real estate	0.27	0.58	−0.31
Other accommodation	0.02	0.31	−0.29
Other sports activities	0.13	0.41	−0.28
Other food activities	0.06	0.26	−0.2
Other retail sale not in stores, sales or market	0.49	0.69	−0.2
Educational support activities	0.04	0.22	−0.18
Sound recording and music publishing activities	0.1	0.27	−0.17
Other telecomms activities	0.24	0.41	−0.17
Business and domestic software development	0.83	0.99	−0.16
Motion picture production	0.23	0.39	−0.16
Technical and vocational secondary education	0.1	0.26	−0.16
Other construction installation	0.28	0.44	−0.16
Other publishing activities	0.13	0.29	−0.16
Specialists medical practice activities	0.08	0.24	−0.16
Repair of other equipment	0.04	0.19	−0.15
Manufacture of other fabricated metal products nec	0.19	0.33	−0.14
Video production activities	0.05	0.18	−0.13
Non-life insurance	0.07	0.2	−0.13
Hospital activities	0.04	0.17	−0.13

Source: BSD, Companies House.
Notes: BSD = enterprises, CH = quasi-enterprises. Shaded = information economy SIC5 bin.

For the information economy, we can see that the matching is generally good – although there are three exceptions. As highlighted above these are 'business and domestic software development' (14.28% of the BSD set, 12.05% of the CH set, SIC 62012); 'information technology consultancy' (52.88%, 42.45%, 62,020) and 'other information technology service activities' (17.96%, 27.7%, 62,090).

Fig. B2 illustrates. We can see that in most cases, CH and BSD % differences are minimal/zero (Fig. B3)

23. *Exploring the extremes*

We now look at the approximately 10% of SIC bins where the differences are most pronounced (Tables B5 and B6, below). Specifically, we take the 37 bins at each end of the distribution above – the tails – where BSD–CH differences are greatest (in one direction or the other).[23]

CH > BSD shares. First we look at the bins where sector shares are higher in CH than the BSD. Results are given in Table B5. A large number of the bins are 'other' or 'not elsewhere classified' (NEC) – type sectors. While we do not directly observe the assignment process, this is consistent with CH processes generating some of the differences. Four of these bins are 'information economy' sectors (see highlights). In particular, there are far more CH firms in 62090, 'other information technology service activities', than in the BSD.

[23] Specifically, we are looking at (74/735) × 100 = 10.07% of the whole.

Please cite this article in press as: Nathan, M., Rosso, A., Mapping digital businesses with big data: Some early findings from the UK. Res. Policy (2015), http://dx.doi.org/10.1016/j.respol.2015.01.008

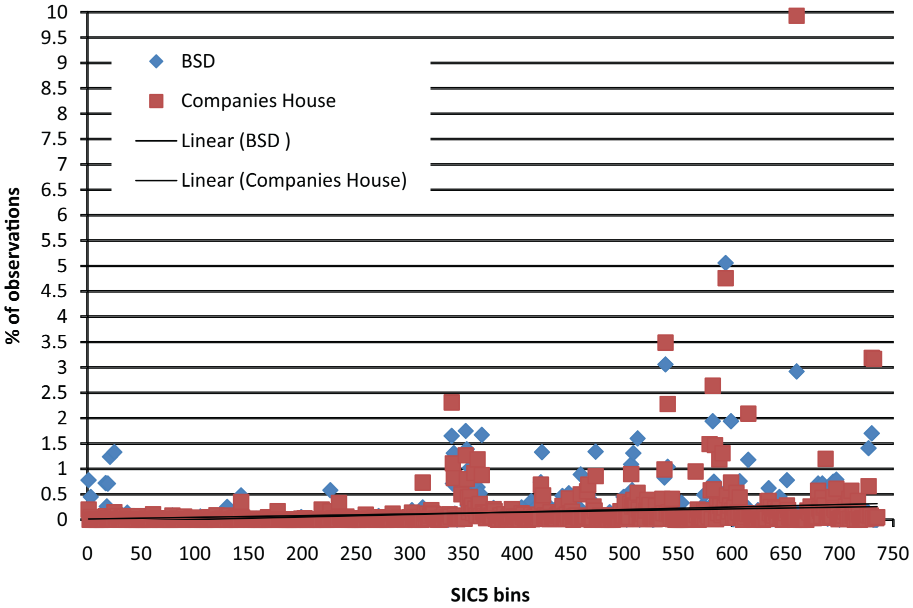

Fig. B1. Comparing BSD and CH shares, all SIC5 sectors, 2011. *Notes*: BSD = enterprises, CH = quasi-enterprises.
Source: BSD, Companies House.

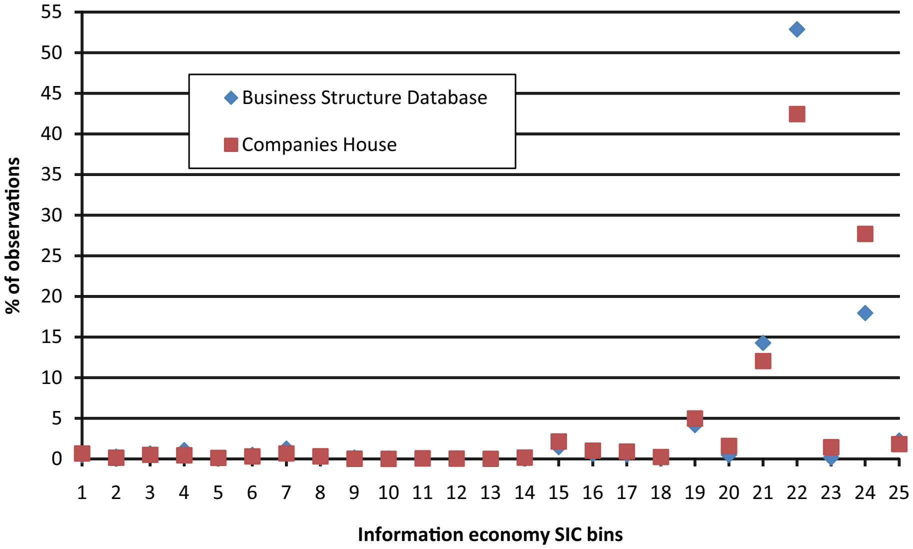

Fig. B2. Comparing BSD and CH shares, info economy sectors, 2011. *Notes*: BSD = enterprises, CH = quasi-enterprises.
Source: BSD, Companies House.

In the BSD, firms in the 62090 bin are slightly older than the BSD, DE and IE averages, and a lot older in terms of age structure. The relevant firms in Companies House are much younger than their BSD counterparts.

However, real estate and construction sector bins also exhibit large BSD–CH differences. We can speculate about the reasons for this. For instance, it is possible that CH shares are generally higher for sectors that have low entry barriers and lots of small players. In addition, retail and construction may both involve extensive use of temporary contracts and/or freelancing rather than PAYE employment.

BSD > CH shares. Results are set out in Table B6. This is a harder group to summarise. Only six bins are 'NEC' sectors. Notably, none of the bins is in our information economy sector set. Seven of the bins are agricultural sectors that likely exhibit large economies of scale and entry barriers. As before, we can speculate about the likely common characteristics of firms in these cells: many might tend to be labour-intensive (pubs and bars, speciality retail, solicitors, barristers), exhibit large economies of scale (construction of domestic buildings, freight shifting) or both.

Again, this suggests that industry-specific characteristics (age structure, entry barriers, economies of scale, input choices) might explain at least some BSD > CH differences. It is also consistent with CH self-assignment producing some of the differences.

Discussion

Overall, comparison of the BSD and Companies House shows that the majority of sectors are well matched. However, the bins where there are differences account for a non-trivial share of observations.

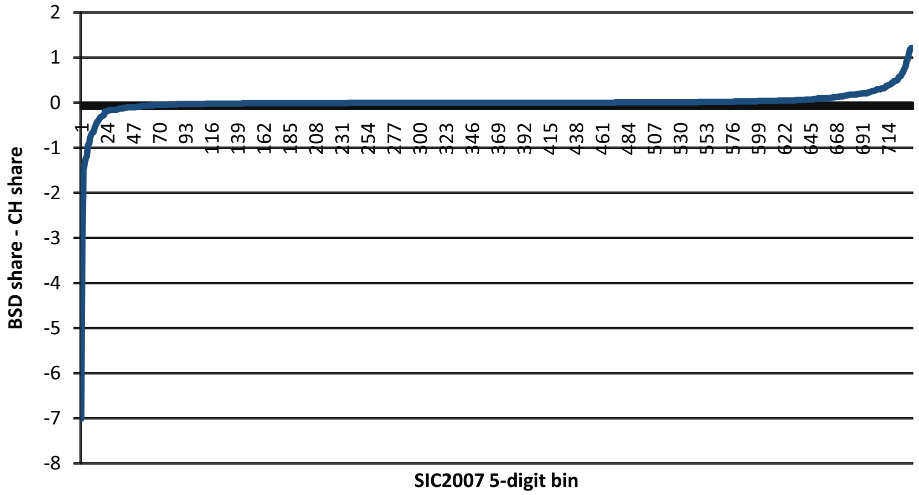

Fig. B3. Comparing BSD and CH differences, 2011. *Notes*: BSD = enterprises, CH = quasi-enterprises.
Source: BSD, Companies House.

Table B6
5% of SIC5 bins with largest BSD–CH differences, 2011.

SIC 2007 5-digit category	%BSD	%CH	BSD–CH
General cleaning of buildings	0.45	0.22	0.23
Security and commodity deal contracts	0.28	0.05	0.23
Raising of other cattle and buffaloes	0.26	0.02	0.24
Temporary employment agency activities	0.62	0.37	0.25
Painting	0.54	0.28	0.26
Wholesale of other machinery and equipment	0.36	0.1	0.26
Activities of religious organisations	0.41	0.14	0.27
Aeneral medical practice activities	0.71	0.43	0.28
Management consultancy other than financial	5.06	4.76	0.3
Activities auxiliary to financial intermediation nec	0.49	0.19	0.3
Other social work activities nec	0.75	0.45	0.3
Construction of other civil engineering projects	0.8	0.5	0.3
Unlicensed restaurants and cafes	0.58	0.26	0.32
Solicitors	0.6	0.28	0.32
Specialised design activities	0.76	0.44	0.32
Activities of other holding companies	0.33	0	0.33
Unlicensed carriers	0.45	0.08	0.37
Licensed clubs	0.42	0.05	0.37
Other sale of new goods in specialised stores	0.89	0.5	0.39
Growing of vegetables, roots and tubers	0.45	0.05	0.4
Machining	0.58	0.17	0.41
Barristers at law	0.45	0.01	0.44
Child day-care	0.51	0.07	0.44
Electrical installation	1.75	1.27	0.48
Freight transport by road	1.34	0.86	0.48
Construction of domestic buildings	1.31	0.82	0.49
Landscape service activities	0.78	0.28	0.5
Joinery installation	1.02	0.45	0.57
Growing of cereals	0.78	0.2	0.58
Plumbing, heating and air-con	1.39	0.8	0.59
Raising of dairy cattle	0.72	0.07	0.65
Raising of horses	0.71	0.03	0.68
Hairdressing and other beauty equipment	1.41	0.66	0.75
Maintenance and repair of motor vehicles	1.67	0.88	0.79
Take-away shops and mobile food stands	1.31	0.39	0.92
Retail sale with food, beer predominating	1.33	0.36	0.97
Pubs and bars	1.6	0.53	1.07

Source: BSD, Companies House.
Notes: BSD = enterprises, CH = quasi-enterprises.

The analysis above confirms that the different sampling frames of the BSD and CH produce some differences in levels and internal structure, even after cleaning Companies House data to make quasi-enterprises. In part these reflect real differences in company and sector characteristics, such as firm age, industry structures and entry barriers. This is not a cause for concern, but implies that we need to take care in making direct comparisons.

We have also tested whether Companies House processes create any sampling bias for information economy analysis. The overall distribution of CH and BSD SIC5 bins is well matched. However, in the bins where differences are most pronounced, we find a number of 'not elsewhere classified' bins where Companies House counts are higher than their BSD counterparts, four of which are in the information economy. That is consistent with self-assignment 'pushing' some firms into particular bins rather than their 'true' location. In turn, this suggests that information economy counts might be higher than true in CH data.

How large a problem is this? Overall, around 10% of observations in the raw CH data are in NEC bins. Conversely, over 20% of observations lack any SIC coding. Again, this is consistent with CH rules leading to non-assignment, and as we have discussed, plausibly biases information economy counts down in our benchmarking sample. Comparing these two magnitudes suggests that information economy counts and shares in our benchmarking sample are more likely to be lower bounds, not upper bounds.

References

Aghion, P., Besley, T., Browne, J., Caselli, F., Lambert, R., Lomax, R., Pissarides, C., Van Reenen, J., 2013. Investing for prosperity: skills, infrastructure and innovation. In: Report of the LSE Growth Commission. Centre for Economic Performance/Institute for Government, London.

Aiginger, K., 2007. Industrial policy: a dying breed or a re-emerging phoenix. J. Ind. Compet. Trade 7, 297–323.

Anyadike-Danes, M., 2011. Aston University Matching of BSD and Fame Data: Report to BIS. Aston Business School, Birmingham.

Askitas, N., Zimmermann, K.F., 2009. Google econometrics and unemployment forecasting. Appl. Econ. Quart. (Formerly: Konjunkturpolitik) 55, 107–120.

Audretsch, D., Feldman, M., 1996. R&D spillovers and the geography of innovation and production. Am. Econ. Rev. 86, 630–640.

Bakhshi, H., Mateos-Garcia, J., 2012. The Rise of the Datavores. NESTA, London.

Baron, A., Rayson, P., Archer, D., 2009. Word frequency and key word statistics in historical corpus linguistics. Angl. Int. J. English Stud. 20, 41–67.

Block, F., Keller, M., 2011. State of Innovation: The US Government's Role in Technology Development. Paradigm, Boulder, CO.

Bloom, N., Sadun, R., Van Reenen, J., 2012. Americans do IT better: US multinationals and the productivity miracle. Am. Econ. Rev. 102, 167–201.

Bresnahan, T.F., Brynjolfsson, E., Hitt, L.M., 2002. Information technology, workplace organization, and the demand for skilled labor: firm-level evidence. Quart. J. Econ. 117, 339–376.

Cable, V., 2012. Industrial strategy. Speech to Imperial College London, 11 September.

Choi, H., Varian, H., 2012. Predicting the present with Google trends. Economic Rec. 88, 2–9.

Couture, V., 2013. Valuing the Consumption Benefits of Urban Density, Mimeo. University of Toronto, Toronto.

Department for Business Innovation and Skills, 2012. Industrial Strategy: UK Sector Analysis. BIS, London.

Please cite this article in press as: Nathan, M., Rosso, A., Mapping digital businesses with big data: Some early findings from the UK. Res. Policy (2015), http://dx.doi.org/10.1016/j.respol.2015.01.008

Department for Business Innovation and Skills, 2013. Information Economy Strategy. BIS, London.

Dittmar, J.E., 2011. Information technology and economic change: the impact of the printing press. Quart. J. Econ. 126, 1133–1172.

Einav, L., Levin, J.D., 2013. The data revolution and economic analysis. In: National Bureau of Economic Research Working Paper Series No. 19035. NBER, Cambridge, MA.

Fetzer, T., 2014. Measuring Legislator Productivity: A New Approach. Mimeo, LSE, London.

Foord, J., 2013. The new boomtown? Creative city to Tech City in east London. Cities 33, 51–60.

Gentzkow, M., Shapiro, J.M., 2010. What drives media slant? Evidence from U.S. daily newspapers. Econometrica 78, 35–71.

Ginsberg, J., Mohebbi, M.H., Patel, R.S., Brammer, L., Smolinski, M.S., Brilliant, L., 2009. Detecting influenza epidemics using search engine query data. Nature 457, 1012–1014.

Harrison, A., Rodríguez-Clare, A., 2009. Trade, foreign investment, and industrial policy for developing countries. In: National Bureau of Economic Research Working Paper Series No. 15,261. NBER, Cambridge, MA.

Hastie, T., Tibshirani, R., Friedman, J., 2009. The Elements of Statistical Learning: Data Mining, Inference and Prediction. Springer, Berlin.

King, G., 2013. Restructuring the Social Sciences: Reflections from Harvard's IQSS. Institute for Quantitative Social Science, Cambridge, Mass.

Lehr, W., 2012. Measuring the internet: the data challenge. In: OECD Digital Economy Papers 194. OECD, Paris.

Lewis, P., Newburn, T., Taylor, M., McGillivary, C., 2011. Reading the Riots: Investigating England's Summer of Disorder. LSE/The Guardian, London.

Lorenzo, D., Reades, G., Calabrese, J., Ratti, F.C., 2012. Predicting personal mobility with individual and group travel histories. Environ. Plann. B Plann. Des. 39, 838–857.

Lucas, R., 1988. On the mechanics of economic growth. J. Monetary Econ. 22, 3–42.

Mazzucato, M., 2011. The Entrepreneurial State. Demos, London.

Moretti, E., 2012. The New Geography of Jobs. Haughton Mifflin Harcourt, Boston.

Nathan, M., Overman, H., 2013. Agglomeration, clusters, and industrial policy. Oxford Rev. Econ. Policy 29, 383–404.

Nathan, M., Rosso, A., 2013. Mapping the Digital Economy with Big Data. NIESR, London.

Nathan, M., Vandore, E., 2014. Here Be Startups: Exploring London's 'Tech City' digital cluster. Environ. Plan. A 46, 2283–2299.

Negroponte, N., 1996. Being Digital. Vintage, London.

OECD, 2011. Guide to Measuring the Information Society. OECD, Paris.

OECD, 2013. Measuring the internet economy: a contribution to the research agenda. In: OECD Digital Economy Papers 226. OECD Publishing.

Office of National Statistics, 2009. UK Standard Industrial Classification of Economic Activities 2007: Structure and Explanatory Notes. Palgrave Macmillan, Basingstoke.

Office of National Statistics, 2010. Business Structure Database: User Guide Social and Economic Micro Analysis Reporting Division. Office for National Statistics, Newport.

Office of National Statistics, 2012. Guide to the Business Population and Demographic Statistics Publications. ONS, Newport.

Rajaraman, A., Ullman, J.D., 2011. Data Mining: Mining of Massive Datasets. Cambridge University Press, Cambridge.

Rodrik, D., 2004. Industrial Policy for the Twenty-First Century, CEPR Discussion Paper 4767. Centre for Economic Policy Research, London.

Romer, P., 1990. Endogenous technological change. J. Politic. Econ. 98, 71–102.

Salton, G., Buckley, C., 1988. Term-weighting approaches in automatic text retrieval. Inf. Process. Manage. 24, 513–523.

Tapscott, D., 1997. The Digital Economy: Promise and Peril in the Age of Networked Intelligence. McGraw-Hill, New York.

Varian, H.R., 2014. Big data: New tricks for econometrics. J. Econ. Perspect. 28, 3–28.

Please cite this article in press as: Nathan, M., Rosso, A., Mapping digital businesses with big data: Some early findings from the UK. Res. Policy (2015), http://dx.doi.org/10.1016/j.respol.2015.01.008