Discussion Paper No. 442

Date: 27 November 2014

Authors: Max Nathan[1] and Anna Rosso[2] with Francois Bouet[3]

MAPPING INFORMATION ECONOMY BUSINESS WITH BIG DATA: FINDINGS FROM THE UK

[1] National Institute of Economic and Social Research, London School of Economics and IZA
[2] National Institute of Economic and Social Research
[3] Growth Intelligence

+44 (0)20 7222 7665
+44 (0)20 7654 1905
info@niesr.ac.uk
@NIESRorg
http://niesr.ac.uk
2 Dean Trench Street
London SW1P 3HE

A company limited by guarantee. Registered in England number 341010. Registered charity number 306083

Abstract

Governments around the world want to develop their ICT and digital industries. Policymakers thus need a clear sense of the size and characteristics of digital businesses, but this is hard to do with conventional datasets and industry codes. This paper uses innovative 'big data' resources to perform an alternative analysis at company level, focusing on ICT-producing firms in the UK (which the UK government refers to as the 'information economy'). Exploiting a combination of public, observed and modelled variables, we develop a novel 'sector-product' approach and use text mining to provide further detail on the activities of key sector-product cells. On our preferred estimates, we find that counts of information economy firms are 42% larger than SIC-based estimates, with at least 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.

JEL C55, C81, L63, L86, O38

Keywords quantitative methods, firm-level analysis, Big Data, text mining, ICTs, digital economy, industrial policy

NOVEMBER 2014

Acknowledgements

This research was funded by Nesta.

Many thanks to Tom Gatten, Prash Majmudar and Alex Mitchell at Growth Intelligence for data, and help with its preparation and interpretation. Thanks to Rosa Sanchis-Guarner for maps. For advice and helpful comments, thanks to Hasan Bakhshi, Theo Bertram, Siobhan Carey, Liam Collins, Steve Dempsey, Juan Mateos-Garcia, Jonathan Portes, Rebecca Riley, Chiara Rosazza-Bondibene, Brian Stockdale, Dominic Webber and Stian Westlake plus participants at workshops organised by Birmingham University, Google, NEMODE, NIESR and TechUK. 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 funder or the data providers. Any errors and omissions are our own.

Corresponding author details

Max Nathan, National Institute of Economic and Social Research, 2 Dean Trench St, London, SW1P 3HE. Email: m.nathan@niesr.ac.uk

1/ Introduction

Information and Communications Technologies (ICTs) - and the 'digital economy' they support - are of enduring interest to researchers and policymakers. National and local government are particularly keen to understand the characteristics and growth potential of 'their' digital businesses. Given the recent resurgence of interest in industrial policy across many developed countries (Rodrik 2004; Aiginger 2007; Harrison and Rodríguez-Clare 2009; Aghion, Dewatripont et al. 2012; Aghion, Besley et al. 2013), there is now substantial policy interest in developing stronger, more competitive digital economies. For example, the UK's industrial strategy (Cable 2012) combines horizontal interventions with support for seven key sectors, of which the 'information economy' is one (Department for Business Innovation and Skills 2012; Department for Business Innovation and Skills 2013). The desire to grow high-tech clusters is often prominent in the policy mix - for instance the UK's Tech City initiative, Regional Innovation Clusters in the US and elements of 'smart specialisation' policies in the EU (Nathan and Overman 2013).

In this paper we use novel 'big data' sources to improve our understanding of 'information economy' businesses in the UK - those involved in the production of ICTs. We also use this experience to critically reflect on some of the opportunities and challenges presented by big data tools and analytics for economic research and policymaking.

For policymakers, a solid understanding of these sectors, products and firms is necessary to design effective interventions. However, it is hard to do this using conventional administrative datasets and industry codes. Data coverage is often imperfect, industry typologies can lack detail, and product categories do not closely align with sector space. More broadly, real-world features of an industry tend to evolve ahead of any given industrial typology.

The UK Government is clear about these challenges:

> *Addressing the lack of clear and universally-agreed metrics will be an early priority for Government and industry. There will be a need for continual reassessment of the scope and definition of the information economy as it evolves. (BIS 2013, p11)*

We use an innovative dataset developed by Growth Intelligence (hence Gi), which deploys an unusual combination of public administrative data, observed information, and modelled variables from unstructured sources and developed using machine learning techniques. We use this off-the-shelf material to develop a novel 'sector-product' mapping of ICT firms. We also take raw text fragments derived by Gi from company websites, and use text mining to shed further light on key sector-product cells. We run these analyses on a benchmarking sample of companies that allows direct comparisons of conventional and big data-driven estimations. The differences are non-trivial: in our preferred estimates we find that the 'ICT production space' is around 42% larger than SIC-based estimates, with at least 70,000 more companies. We also find employment shares over double the conventional estimates, although this result is more speculative.

This approach delivers significant extra dimensionality and detail compared to simply using SIC codes, but it is not without limitations. This brings us to the second contribution of the paper, in which we draw on our experience to highlight opportunities and challenges for researchers working with similar big data methodologies. 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 use our work to discuss the substantial richness big data can bring to innovation research, and 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 defines key terms and issues. 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 respectively detail sample construction and identification steps. Sections 6 and 7 give descriptive results. Section 8 concludes.

2 / Context and key issues

Our research questions are: first, what is the true extent of ICT manufacturing and service activity in the UK, and what are the key characteristics of these businesses? Second, what are the differences between big data-driven estimates and those from conventional administrative datasets?

2.1 / The 'digital economy', the ‘information economy’ and ICT production

Governments in the UK and elsewhere are keen to grow their 'digital economies'. What does this mean in practice? The ‘digital economy’ is an economic system based on digital technologies (Negroponte 1996; Tapscott 1997). This is an ecosystem of sorts: an interlocking set of *sectors* (industries and firms), *outputs* (both supporting products and services, and the content these are used to generate), and a set of production and distribution *inputs* used at varying intensities by firms and workers across all sectors (OECD 2011; OECD 2013). We could also define a set of cross-industry *occupations* where such technological tools are essential to the main tasks (Brynjolfsson and Hitt 2003; Acemoglu and Autor 2011).

Our analysis focuses on the production side of this system, where we map both industries and outputs. We ignore inputs, for the simple reason that it is now hard to think of any economic activity where digital inputs *do not* feature, and given the pace of change in (say) internet tools and platforms, definition and measurement problems for digital inputs are severe (see Lehr (2012) and OECD (2013) for a discussion of these issues). And as discussed above, while policymakers are keen to improve ICT infrastructure such as broadband networks, they are also increasingly interested in helping sectors and firms to grow.

The standard OECD/UN definitions of digital activities comprise detailed product/service groups identified by an international expert panel: these are then aggregated into less detailed 4-digit standard industry code (SIC) bins (OECD 2011).[1] These SICs form the basis of most

[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 SIC2007 4-digit cells. See OECD (2011) for detail.

analysis. 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 supply-side activity 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 sectors delineated in the UK Department of Business' 'information economy strategy' (Department for Business Innovation and Skills 2012; Department for Business Innovation and Skills 2013). We refer to firms in these industries as 'information economy businesses'.

There is a live debate in the UK about exactly how broadly to define the information economy in industry terms. Some analysts prefer a very narrow definition, which concentrates purely on ICT manufacturing; conversely, some industry voices would like a much broader approach that includes manufacturing, services and related supply chain activity (such as wholesale, retail, installation and repair). This means that alternative mappings of the information economy need to take into account these differences of opinion. We 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]

[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 made for the information economy element in a UK context. 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 ICT manufacturing, code 26) and a broader set of SICs comprising manufacturing, services and supply chain activity including 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

Table 1. ICT products and services. List of SIC2007 codes.

ICT manufacturing	
26	**Manufacture of computer, electronic and optical products**
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**
58210	Publishing of computer games
58290	Other software publishing
61	**Telecommunications**
61100	Wired telecommunications activities
61200	Wireless telecommunications activities
61300	Satellite telecommunications activities
61900	Other telecommunications activities
62	**Computer programming, consultancy and related activities**
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**
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.

engineering activities), 71122 (Engineering related scientific and technical consulting activities), 71121 (Engineering design activities for industrial process and production).

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; 2013) and because many product categories bleed across sector boundaries (see below).

2.2 / Measuring ICT production activity

Ascribing activities to sectors is necessarily an imprecise process, particularly when conventional, administrative datasets are used. In the UK there are three principal issues.

The first issue is about data coverage. For firm-level analysis, the main UK administrative source is the Business Structure Database (BSD), which draws on sales tax, employment and company records as well as government business surveys (Office of National Statistics 2010; Office of National Statistics 2012). However, the BSD only includes firms paying UK sales tax and/or those with at least one employee on the payroll. Pooling across sectors, the BSD covers 99% of UK enterprises but for sectors with large numbers of start-ups and small young firms - such as the digital and information economies, or fields such as nanotech - coverage will be significantly poorer. The BSD is also limited in terms of information, only providing variables on age, employment count, industry, turnover and business address. Alternative sources such as Companies House provide much better coverage of economic activity, but contain important limitations of their own (see Section 3).

The second issue is about SIC codes. 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, for niche or rapidly-evolving parts of the economy, SICs can be too broad or aggregated to shine much light. For this reason, firm counts for ‘other’ or ‘not elsewhere classified’ based SIC cells are often much bigger than for others close by in sector space, even at the most detailed five-digit level. In the 2011 BSD, for example, the second largest ICT cell is 'Other information technology service activities'

(62090) which contains 22,444 enterprises (compared to 66,090 in 'Information technology consultancy activities', cell 62020).[3]

A third, related issue is that product categories both contain far more detail than sector cells, and these product categories 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).

Taken together, these issues mean that mapping the extent and characteristics of firms in the digital economy using conventional sources and industry information alone is challenging - because of the nature of these firms, constraints on conventional data sources and on purely sector-based classifications. 'Big data' sources and analytics have the potential to bring helpful clarity here.

2.3 / Big Data

'Big Data' is a complex concept that needs careful specification. A popular – but seemingly circular – definition says that big data is 'datasets too large for conventional analysis' (Dumbill 2013). Instead 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 real time, or close to it; c) have high dimensionality, including phenomena previously hard to observe quantitatively, and d) are much less structured than 'conventional' sources, such as administrative data or surveys.

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; Ginsberg, Mohebbi et al. 2009; Choi and Varian 2012); proprietary datasets, such as those derived from mobile phone networks (Di Lorenzo, Reades et al. 2012); and material derived from texts,

[3] In our main dataset, which is based on Companies House, the relevant counts for 2011 are 42,491 and 65,072 quasi-enterprises, respectively. Again, these are the two largest cells.

both historic (Dittmar 2011) and contemporary textual information taken from the Web, political speeches, social media or patent abstracts (Gentzkow and Shapiro 2010; Lewis, Newburn et al. 2011; Couture 2013; Fetzer 2014). Structured administrative datasets also take on 'big' features when linked together or enabled with API functionality, allowing researchers to 'call' the data more or less continuously. In the UK, virtual environments such as the Secure Data Service (SDS) and HMRC DataLab provide researchers with secure/monitored spaces for matching exercises,[4] and a number of government agencies are introducing API functions for data stored online.

In theory, these sources, tools and platforms 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. The speed, scale and dimensionality should allow us both better coverage of businesses, clearer and more detailed delineation of product / sector space, and richer information on business characteristics. In turn, this promises more reliable analysis, which should lead to development of more effective policies.

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

[4] See http://ukdataservice.ac.uk/get-data/secure-access and http://www.hmrc.gov.uk/datalab/ (both accessed 1 December 2013).

3 / The Growth Intelligence dataset

Our main dataset is 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 1.

3.1 / Companies House layer

The 'base layer' comprises all 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 overseas companies with a UK branch and some business partnerships. Registered companies are given a unique CRN number, and 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. For this reason, descriptive results should be interpreted with some caution.

3.2 / Structured data layers

Gi match Companies House data to a series of other structured administrative datasets. In this analysis we focus on two of these. Patents data is taken from the European Patent Office PATSTAT database. Patent titles and abstracts are obtained from the EPO API feed and combined with the raw data. We also use UK trademarks data, which is taken from the UK Intellectual Property Office (UK IPO) API feed.[5] Gi use these structured datasets in two

[5] Patents and trademarks matching is done on the basis of name and address information. We are grateful to the UK IPO for use of a recent patents-companies crosswalk, which we deploy alongside Gi matching.

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).

3.3 / Proprietary layers

This part of the Gi dataset is developed through 'data mining' (Rajaraman and Ullman 2011). Gi develop a range of raw text inputs for each company, then use 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 assign weights to these 'tokens' which indicate their likelihood of identifying meaningful information about the company. Supervised learning approaches (Hastie, Tibshirani 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. More information is provided in Appendix 1.

The Gi dataset is complex. For this proof of concept paper we use the Companies House 'base layer' plus a selection of Gi's modelled variables (in-house sector and product classification, plus modelled revenue); in addition to these off-the-shelf variables we also use 'raw' web tokens and token categories for exploratory text-mining exercises on parts of our sample.

3.4 / Pros and cons of a Big Data approach

The properties of the Gi dataset should allow it to deal with the three measurement challenges outlined in Section 2. First, compared to administrative data sources like the BSD, the Gi data has greater coverage of economic activity and provides substantially more information (thanks to the matched and modelled layers outlined above). Second, the additional dimensionality in company classification should allow us a more precise delineation of companies providing ICT products and services. Specifically, SIC5 codes provide 806 sectors

in which to place companies, but Gi's 145 sector and 39 product groups provide 5510 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. Third, because many of Gi's sources are available in real time or close to it, the company can regularly update its data and track switches in company characteristics, such as pivoting from one product type to another.

Conversely, there are some potential limitations in the Gi dataset. First, coverage of online sources is not perfect. Many companies in the UK do not have a website, for example, and not all websites can be successfully scraped due to site content or build. 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 1) 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 information 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 data are missing SIC information it will therefore be smaller than the 'true' number of companies. In some cases, we 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.[6] 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).[7] 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.[8]

[6] 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.

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

[8] 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

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.[9]

We validate our cleaning steps by constructing a 'true' sample of all quasi-enterprises, this time including all the companies dropped because of missing SIC codes. We then compare this 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 project the August figures based on smoothed 2011-2012 trend: this gives a figure of 2.45m businesses, within 99% of our true sample estimate.[10]

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

---

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.

[9] 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.

[10] The 2.36m figure 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.

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 2 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 consists of 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 now use basic industry-level information economy categories (from SIC codes), and exploit the additional richness and dimensionality in our 'big data' 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. While these studies have important differences, they 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, Porter et al (2008) deploy bibliometric analysis of academic publications to identify the contours of nanotech research. Specifically, they construct a 'core' set of nanotech publications that is then verified by experts, and use keywords for these publications to build a two-stage Boolean search algorithm that can be run on databases of academic papers or patents. 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 generate finer-grained technology vocabularies from patent titles and abstracts. These are 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 texts for companies in our benchmarking sample. In practice, we have a category-level starting definition of the information economy from the UN/OECD definition and their UK variants (see Section 2). However, in our data this is only available for industry sectors - and with some disagreement among policy actors about field boundaries. And rather than raw words and phrases, we are working with larger, off-the-shelf sector and product categories (see Section 3).

We therefore use this 'categorical vocabulary' as a starting point for our analysis. We are also able to compare estimates for the information economy done with conventional industry codes (based on company self-assessment), to those done with Gi's sector and product categories (based on a range of observed and modelled information).

5.1 / Mapping strategy

Our basic mapping steps are as follows. First, we take the sub-sample 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. Kicking out the long tail 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.[11] The review is rule-based: specifically, we look for sparse Gi sector or product cells where the sector or product 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 Gi sector cells 'computer network security' and 'e-learning', which features in the OECD product list, but exclude the product cell 'hardware tools machinery', which Gi use 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 principle product / service is *also* ICT-relevant.

[11] 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'.

5.2 / Identification

This 'sector-product' approach, built on a range of data sources, should provide a better mapping of information economy firms than using self-ascribed industry codes alone. Specifically, 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 is 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.

To make this analysis robust we also need to deal with some potential problems.

First, our starting categories are not completely fixed; as outlined in Section 2, there is some disagreement about which SIC codes should be used to delineate the information economy (recall that while some industry analysts want a very small set of SICs covering production, others believe that a wider set of ICT supply chain industries should be included). This means that the sector-product results (specifically, the set of false negatives) might be endogenous to the set of starting industry cells, rather than being driven by real differences in sector-product information. To deal with this, we 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, we might worry that our 0.2% threshold rule still identifies some irrelevant sector / product space (leading to false positives). We therefore experiment with tighter thresholds at 0.3% and 0.5% of long-listed observations.

Third, we might worry the sector-product approach may 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. Note that this objection also applies to SIC codes, as we discuss in Section 2. In our case, we are relying on the *combination* of sector-by-product information to provide extra dimensionality across the pooled sample, 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, providing additional descriptives.

6 / Results

6.1 / Headline counts and shares

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

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 Gi ICT sectors of our preferred estimates. Panel F and G use different threshold rules to identify Gi ICT products and sectors.

Panels A and B give sector-based and sector-product based estimates of information economy companies, based on SICs in Table 1 and GI sector-product cells respectively. Sector 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 in panel B are 42% higher compared to SIC-based definitions in Panel A. Overall, this difference in headline numbers – around 70,000 'missing' companies not in the SIC-based estimates but in the Gi-based estimates – 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. This also implies that the true numbers of information economy firms is likely to be higher than the counts here.

Other panels report robustness checks that explore some of the identification challenges discussed in section 5.2. Panels C and D show sector-based estimates when changing the starting set of SIC sectors. As we discuss in section 1, stakeholders disagree over the 'real' scope of the information economy, with some favouring broader or narrower definitions than BIS have chosen. Therefore in Panel C1 we look only at SICs covering ICT services, while in Panel D1 we use a broader definition of the information economy that also includes SIC codes in the wider ICT value chain.[12] 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). While this highlights the importance of how the set of 'information economy' businesses is initially defined, our main results survive – albeit with a smaller set of 'missing' companies unearthed.

Panel E tests the effectiveness of the sector-product approach as opposed to using sector-only information. We would expect the lack of granularity to produce higher estimates, which it

[12] 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).

does (305,177 versus 225,800 companies, almost 16% of the sample). (Using only the product dimension the share would be driven up to more than 50%.) [13]

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.

6.2 / 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. To illustrate the difference, Table 3 maps these quasi-enterprises back onto their SIC codes, for the 18 largest SIC cells.[14]

[13] Results available on request.

[14]We conduct the same exercise mapping back to SIC using different ICT SIC definitions. Results are available on request.

Table 3. SIC codes for 'additional' ICT producer companies, 18 largest cells.

Description	SIC2007	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 n.e.c.	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 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.

Note that some of these SIC bins (specifically, 33200 and 95110, 4.8% of the total) would be included in our 'broad-based' set of information economy SIC codes, as discussed in the previous section. 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 SIC codes 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.

6.3 / Internal structure of the ICT producer space

Next, we take a closer look at the internal structure of our Gi-based ICT producer estimates. Tables 4 and 6 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 principle product is (say) 'consultancy' might be in any of the sector cells in the sector table. In principle, then, we could construct a very large matrix of all 378 sector*product combinations.

Table 4: Total number of firms in the information economy by GI sectors

			Revenues (£)	
	Observations	*%*	*Mean*	*Median*
computer_games	2,585	1.14	1793241	3181.5
computer_hardware	3,514	1.56	2473394.4	83803
computer_networking	3,902	1.73	2135848.7	93784
computer_network_security	226	0.1	13223530	1027628
computer_software	23,455	10.39	1433080.5	35564
consumer_electronics	2,074	0.92	11125476	97584
design	10,049	4.45	753104.63	53798.5
e_learning	347	0.15	4496422.4	320504.5
electrical_electronic_manufacturing	17,319	7.67	3696466.6	93784
information_services	823	0.36	5018562.8	182405
information_technology	104,768	46.4	995039.69	38364
internet	2,954	1.31	6527924.2	195958
marketing_advertising	11,038	4.89	3695790.4	42077
mechanical_or_industrial_engineering	27,326	12.1	1145004.3	93784
semiconductors	183	0.08	64762995	1323417
telecommunications	15,237	6.75	16347362	78165
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.

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.

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.

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.

Finally, we use text mining on website information for a sub-sample of companies to uncover more information about the largest cells, 'information technology' and 'consultancy'.[15] As set out in Section 3, Growth Intelligence scrapes website text and uses machine learning to

[15] 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.

uncover key words and phrases (raw 'tokens'), and contextual information for each token ('token categories'). Gi reports 12 token categories of which we use four – organization, product, technical term and technology – most likely to describe the nature of the company, the technology used and the type of product.[16] Tokens in these categories are assigned a value 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. This raw token information needs to be cleaned: we harmonize the words that appear in the tokens, 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 some English stopwords following an existing vocabulary.[17]

In Figure 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 (26,408 companies).[18] 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%).

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

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

[18] This threshold can be modified to higher or lower frequency.

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

Source: Gi data

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-product cell 'information technology-consultancy' (Panel B), product cell 'consultancy' across all ICT sectors (Panel C) and for the sector cell 'information technology' across all ICT products (Panel D).[19]

[19] In the subsample of companies with tokens we have 3,716 companies doing IT and consultancy, 12,556 companies providing some consultancy service in any ICT sectors, and 4,296 in the information technology sector (any ICT products).

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%

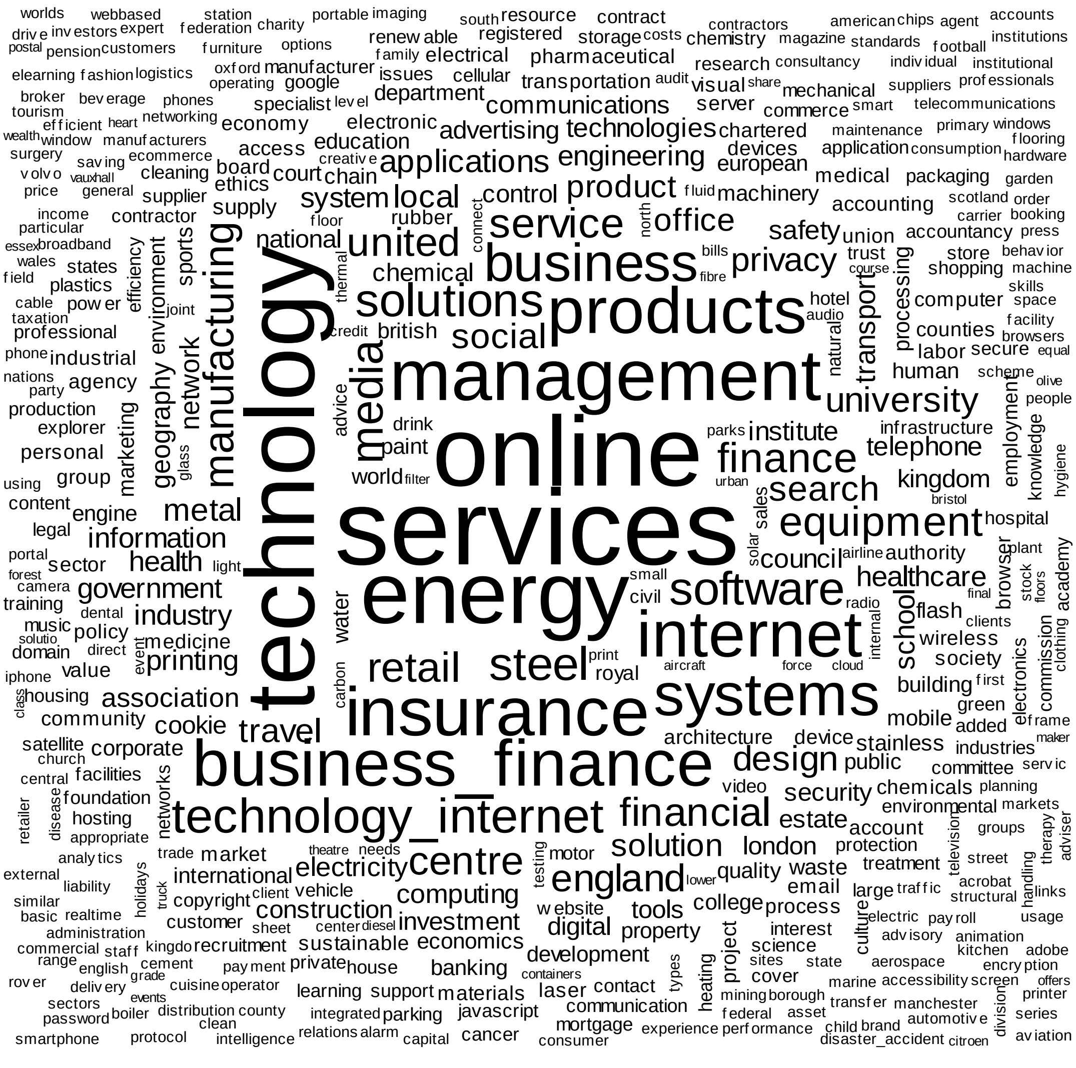
system	18,911	1%	2,727	1%	7,998	1%	3,663	1%
service	18,493	1%	3,410	2%	9,901	1%	3,872	2%
energy	18,013	1%	2,340	1%	9,108	1%	2,591	1%
products	17,627	1%	2,192	1%	7,179	1%	2,590	1%
applications	17,477	1%	2,977	2%	7,603	1%	3,593	1%
marketing	16,758	1%	1,404	1%	9,974	1%	1,614	1%
social	16,033	1%	2,384	1%	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 the IT (sector) and consultancy (products). Panel C companies doing consultancy. Panel D companies in the IT sector.

Results show that the word that appears the most across panels A, B and C is 'technology', while for the IT sector alone it is ‘software’. The former represents 4% of the total number of words in the complete ICT producer space (Panel A), 7% in the 'IT-consultancy' sector-product cell, 5% in Panel C (consultancy products) and 6% in Panel D (IT sector), while 'software' in IT appears in 7% of cases. Even more interesting is that the distribution across panels within these information economy cells is very similar, and despite being relatively sparse, with some words appearing only 1% of the time, we observe a high density in the same words across all four panels. To understand how distinctive these words

are, we also look at the word distribution in the rest of the economy: we might worry, for example, that these are simply terms which appear on *any* company's website. Interestingly, we find that the most relevant words are not the same and actually the words that are denser in ICT production space are under-represented in other activity spaces (Figure 2).

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



Source: Gi data

7 / Characteristics of ICT and non-ICT businesses

This section provides more detailed information on companies' age, inflows, revenues and employment. Not all companies report revenue or employment data, so these latter analyses are done on suitable sub-samples. While some companies have no revenue or employees to report, there are also some holes in the Companies House data.[20] We perform a range of diagnostic checks to make sure the sub-samples are representative, but data limitations mean that revenue and employment information has to be interpreted with some care.

7.1 / Age

Table 7 reports the average age of ICT and non-ICT companies in the benchmarking sample.[21] 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.

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

In Table 8, we show the distribution of companies by age groups. This share can easily be interpreted as a survival rate as nothing is revealed about the actual turnover rate of companies.[22] Panel A uses SIC code definitions; panel B uses sector-product groups. In Panel

[20] Some companies will not file annual returns or accounts on time; others may file incomplete information; others may fail to declare revenue. Companies House may have limited resources to chase up offenders.

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

[22] 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.

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, usually defined as companies less than three years old, are slightly more common among ICT producers than in the rest of the economy

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

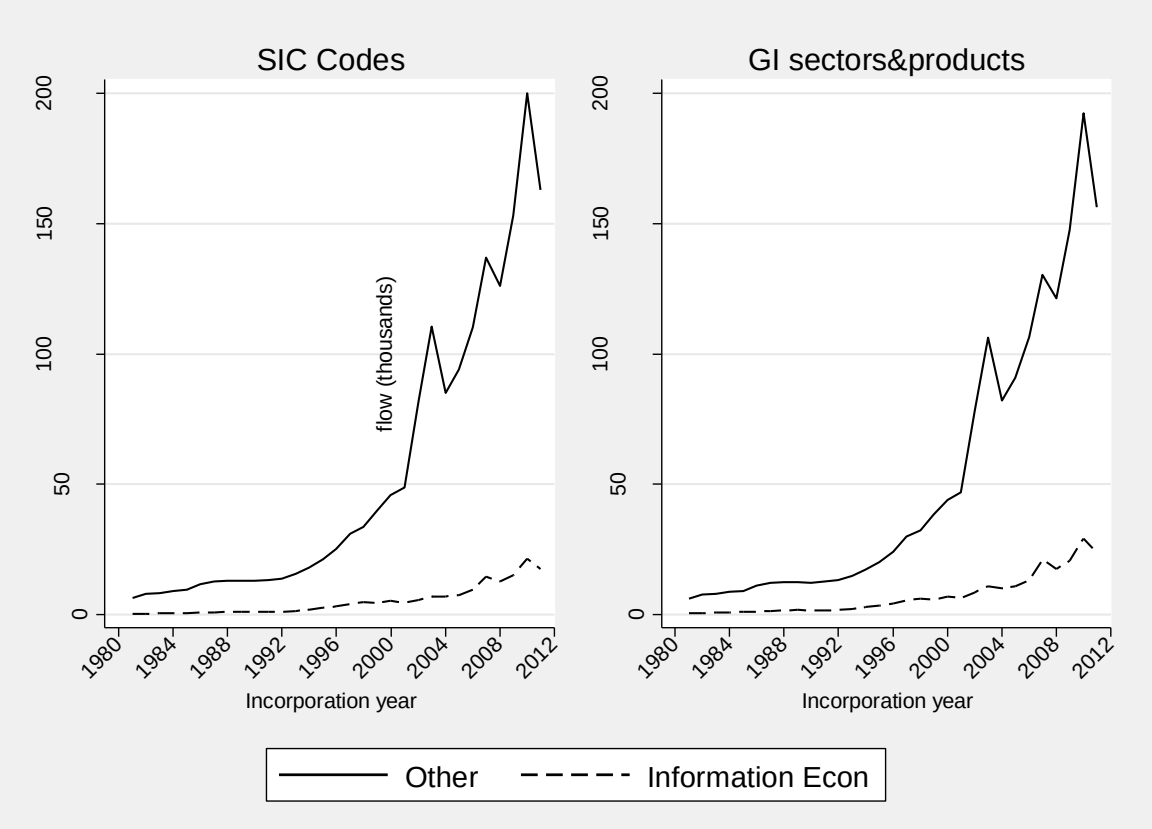
On the face of it, these findings are surprising. The popular image of the ICT industry is of start-ups and very young companies. Our evidence, however, suggests that there is no reason to think that the ICT companies are more ephemeral than the other companies. Our analysis of inflows, below, also tells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

7.2 / Inflows

Figure 3 shows the inflow of companies into the economy, comparing inflows of companies into ICT production (dashed line) with 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.[23]

Figure 3. Inflow of companies between 1991 and 2011

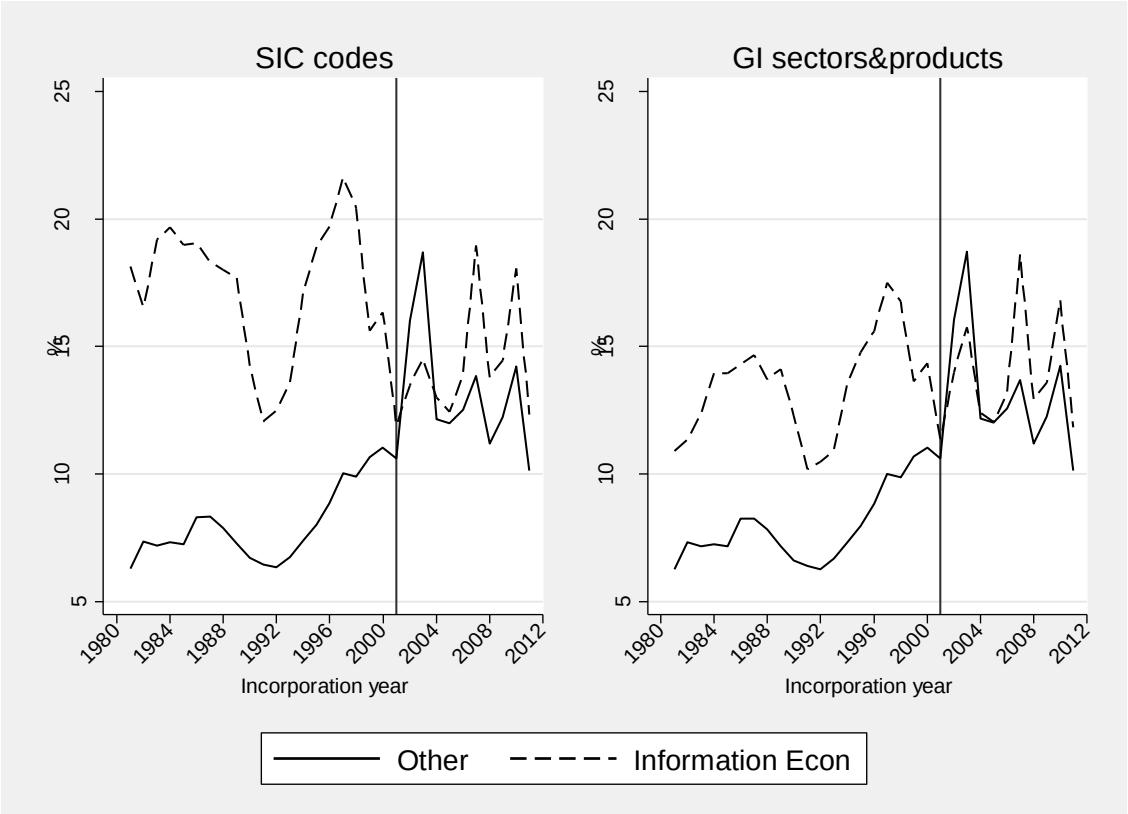


Source: Gi and Companies House data
Note: The graphs show the inflow of active companies in each year

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 Figure 4.

[23] Company reclassification may be more pronounced over longer periods: this will not be captured in SIC codes, which in Companies House are ascribed when companies are set up. Growth Intel's more up to date information may be buying us extra precision here.

Figure 4. Growth rate in the number of firms between 1980 and 2011



Source: Gi and Companies House data

Note: Growth rate as a percentage of number of firms entering the economy each year over the total existing firms

Two things are worth noting. First of all, 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.

7.3 / Revenue

As discussed in Section 3 and in the Appendix, 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: Mean and median revenues and revenue growth from Companies House

	A. Average Revenues						B. Average Annual Revenue Growth			
	Companies House		Growth Intel		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 to 2012. Growth Intel 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 to 2012.

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 companies who report revenues to Companies House over more than one year. The first column reports the average percentage growth, defined as

the within-firm growth of revenues 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: Growth Intel's revenues by sector

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

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

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, so we run checks to determine the shape of the bias.[24] 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.

[24] Full results are available on request.

Table 11. Employees per firm.

	Breakdown	Observations	Gi sector*product		SIC codes	
			Mean	Median	Mean	Median
2008-2012	Other	143989	31.86	5	34.79	5
	Information Economy		60.06	3	22.82	4
	Average		34.17	5	34.17	5
2010-2012	Other	75927	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

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 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.

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 product*sectors)	11.75	8.92
Other	88.25	91.08

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 information economy 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 2010-12. This is as we would expect, since underlying company counts are higher in our big data-driven definitions.

7.5 / Location

To get a sense of how the information economy is distributed across the UK, we geo-code individual companies into Travel to Work Areas (TTWAs). TTWAs are designed to represent functional labour markets, and are generally considered to the best available approximation of a local economy.[25] Our analysis is using 'quasi-enterprises' rather than individual plants, and using the registered addresses of those companies. This needs to be borne in mind in interpreting the results. First, in most cases the registered address of a company will also be their trading address, but not in all cases.[26] Bundling companies into TTWAs minimises the chances of putting companies in the 'wrong' part of the country. Second, using registered addresses is also likely to lead to a more big-city-centric distribution - since London and large urban cores are more likely to contain company headquarters than TTWAs with smaller cities, or rural areas.

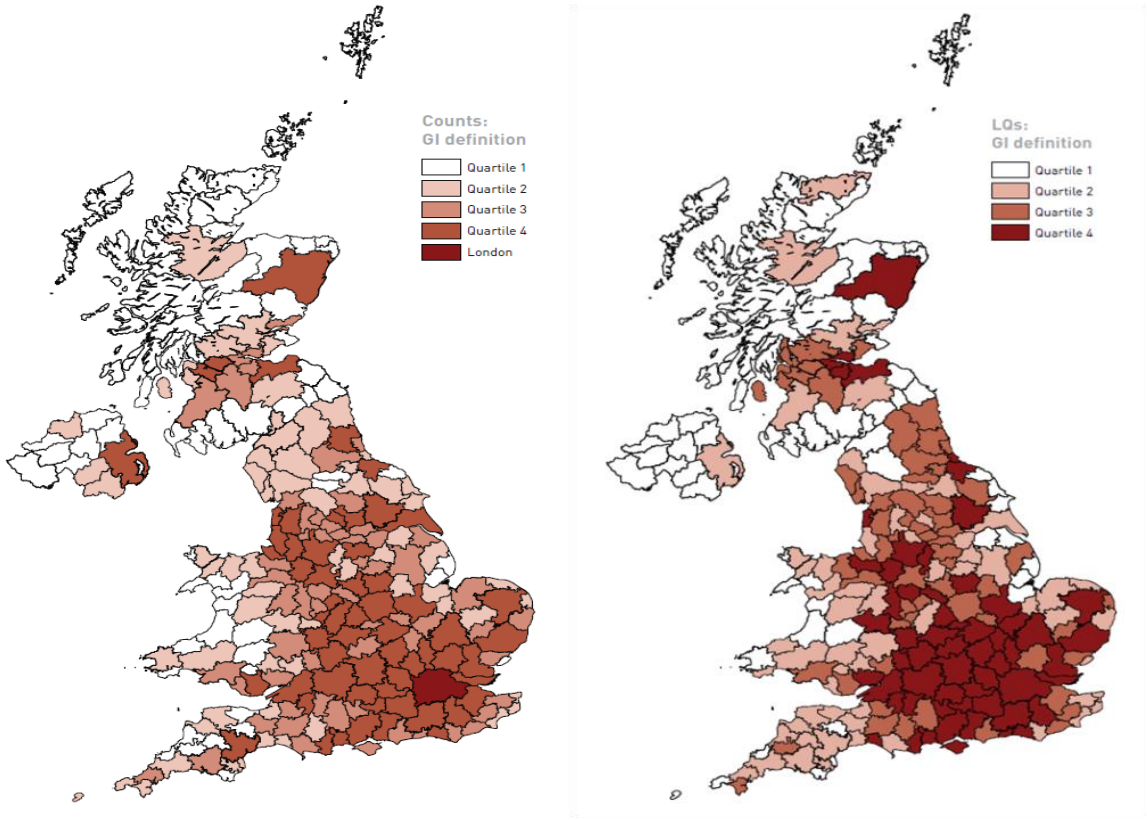
Geo-coding also slightly shrinks our benchmarking sample, from 1.94m to 1.936m companies. This is because not all company addresses provided to Companies House include postcodes, and because some companies provide PO Box addresses (where the postcodes are not assigned to a particular geography).

[25] Formally, at least 75% of those living in a given TTWA also work in that TTWA, and vice versa.

[26] Gi have collected experimental trading address postcodes for a sub-sample of 316,884 companies, using postcode data from company websites and phone directories. This data is very noisy and should be treated only as a fuzzy estimate (issues include false positives for common company names and 'false missings' if websites are non-scrapable or not provided). 257,358 companies in our benchmarking sample (13.6%) have trading address data. Of these, 216,349 (84.07%) have only one trading address. Identical or co-located registered and trading addresses for same-named companies are very likely to represent the same company. In 97,629 cases (45.31%) the full trading postcode is the same as the registered address; for 111,183 cases (51.39%) the trading address is in the same 3/4-digit postcode sector as the registered address; for 149,426 companies (69.35%) the trading address is in the same 2-digit postal area as the registered address.

We first look at the distribution of companies around the country (figure 5). The left hand map maps the UK's Travel to Work Areas and shows banded counts of information economy firms, using the Gi-based sector-product measure. We have divided the counts into quantiles, each of which represents 25% of the observations, plus a separate London band.

Figure 5. Information economy company counts and LQs by TTWA (sector*product)



Note: counts are quartiles plus London.

The information economy is very spiky, with a lot of co-location in London, Manchester and the Greater South East. Using our preferred sector*product measure, the 10 TTWAs with the most digital economy companies are London (58,248 companies), Manchester (7,582), Guildford and Aldershot (6,172), Birmingham (5,384), Luton and Watford (4,578), Reading and Bracknell (4,091), Bristol (3,862), Crawley (3,827), Wycombe and Slough (3,483), and Brighton (3,376). Underneath this group are another 40-odd TTWAs with 1,000-3,000 information economy companies, followed by a very long tail: over 60% of the areas on the map have less than 500 companies, and 25% have under 100.

Using SIC codes the top 10 TTWAs are very similar, although counts are smaller: London (43,802 companies), Guildford and Aldershot (4,825), Manchester (4,604), Birmingham (3617), Luton and Watford (3,592), Reading and Bracknell (3,405), Crawley (2,841), Bristol (2,803), Wycombe and Slough (2,670) and Brighton (2,668). Overall, around 80% of companies are in urban areas - defined as TTWAs with a city of at least 125,000 people - although this share will be higher than plant-level analysis, which would look at trading locations as well as registered addresses.

Next, we use location quotients to get a sense of where the information economy is most locally clustered (in the sense of co-location). Location quotients compare the local area share of a group *i* to its national share.[27] Location quotients over 1 indicate local clustering; under 1 suggests dispersion. Results are shown in the right hand panel of Figure 5.

Looked at this way, the spatial footprint of the information economy is rather different. Using our preferred Growth Intelligence-based metrics, the 10 areas with the highest location quotients are Basingstoke (1.84), Reading (1.78), Newbury (1.68), Milton Keynes (1.54), Swindon (1.51), Luton and Watford (1.43), Guildford (1.41), Middlesbrough (1.38), Wycombe and Slough (1.374) and Stevenage (1.372). Just outside this are Brighton (1.35), Coventry (1.34) and Cambridge (1.33).[28] Using LQs, then, highlights the importance of the digital economy to cities in the Greater South East, especially in the crescent of high-value activity that runs around the West of London, but also highlights some perhaps unexpected hotspots, both in the North East (Middlesbrough, 1.38 and Hartlepool, 1.21) and in Eastern Scotland (Livingston and Bathgate, 1.20 and Aberdeen , 1.11).

Why don't we find places like London, Manchester and Birmingham in these lists too? Partly because these are large cities with diverse economies. However, we can use more detailed geocoding to look at very local clustering within these cities, particularly for young firms. Evidence suggests that large cities can act as 'nurseries' for start-ups (Duranton and Puga 2001), and we see some confirmation of this in our data. Table 13 shows the 30 postcode

[27] Formally, LQ*ia* = (p*ia* / pa) / (pi / p), where p*ia* / pa is the local population share of *i* in area *a*, and p*i* / p is *i*'s national population share. An LQ of above 1 indicates concentration, or local shares above the national shares; scores below 1 indicate dispersion, or local shares below the national share.

[28] Using SIC codes to define the top 10 information economy clusters, we find a broadly similar pattern orientated around the Greater South East.

sectors with the largest counts of information economy start-ups (defined as companies up to three years old), and the counts of all start-ups and all information economy firms.[29]

Table 13. Start-ups by postcode sector: top 30 areas.

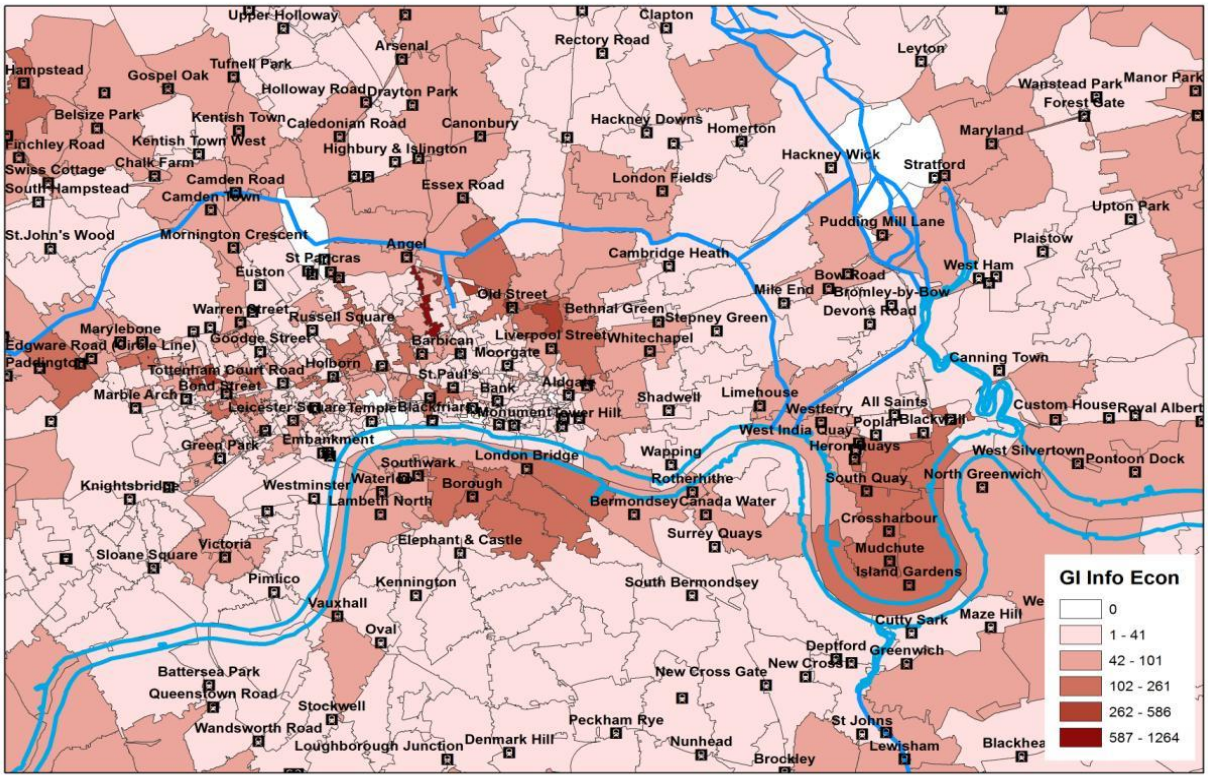
		#IE startups		#IE firms	
Area	**#startups**	**SIC**	**Gi sector-product**	**SIC**	**Gi sector-product**
EC1V4	1557	194	242	1059	1277
BN36	368	200	218	784	860
N120	615	111	139	356	457
EC1V2	288	92	106	533	590
HP11	190	95	96	488	518
SW191	254	77	86	363	413
CV12	293	75	85	389	440
W1B3	370	70	83	289	345
BH121	486	77	80	328	340
EC2A3	292	69	77	276	348
SO237	93	57	59	271	294
E145	147	47	53	216	238
DA144	162	42	52	145	189
EC1N8	249	37	51	194	268
NG27	100	41	50	250	298
BN11	263	39	48	315	390
FY45	109	28	47	218	293
BH11	91	38	47	340	379
E149	216	46	45	244	267
W1G9	557	32	44	216	300
BN32	261	34	42	194	253
RG78	186	33	41	247	328
N31	309	36	40	280	335
SW193	116	38	40	341	375
FY42	66	34	39	135	158
IG26	200	33	36	158	189
SW64	261	26	35	117	161
CV48	83	32	35	150	159
SW192	169	25	33	183	223
WD61	158	30	33	188	224
TW33	96	32	33	137	150

Note: data are sorted by # information economy startups (sector-product).

[29] Postcode sectors comprise the first four / five digits of a postcode, for seven / eight-digit postcodes respectively.

Overall, the distribution of these young firms is highly uneven: over 32% of postcode sectors have no start-ups at all, and 56% have less than 10. The remaining areas contain over 93% of all start-ups. Five of the top 10 postcode sectors are in Central London, of which three are in East London (EC1 or EC2). The rest of the top 10 are in Brighton (BN36), Coventry (CV12), High Wycombe (HP11) and Poole (BH121). Figure 6 maps this postcode-sector activity across central London. We can see some of the familiar geography of Tech City (Nathan and Vandore 2014), but also other hotspots around London Bridge and Canary Wharf, as well as parts of the West End. (Remember that our mapping does *not* include digital content activity, so broader 'digital economy' counts will be rather higher than this.)

Figure 6. Information economy company counts in Central London.



Note: geographies are postcode sector.

7.5 / Patenting

Information on companies' patenting activity provides a useful insight into IP and ideas generation. This section gives some headline descriptive findings. Our patents data covers European Patent Office (EPO) applications and is matched onto company data, using name

and company/applicant/inventor location information.[30] We're interested in patents where the applicant is based in the UK, or where at least one of the inventors involved is UK based. The overall match rate from patents to companies is 65.4%, which is satisfactory. A number of patents will not match because the applicant is an individual rather than a company; where the applicant name field has errors; or when applicants are not in our benchmarking sample but may be in the wider Companies House data.

The resulting matrix comprises 63,860 'raw' patents filed by 8,869 companies between 1978 and 2012; 108,316 inventors are named, of whom 85,498 are resident in the UK. Patents are organised by 'priority year', that is, the year in which they first entered the EPO or other patents office). A number of patents have more than one applicant: so to avoid double-counting the analysis is done using weighted patents, where weighting patents with the number of applicants.

Table 14 looks at company-level patent counts, pooled across years. Average counts are very small (Panel A) explained by the fact that most UK companies do not patent at all (see Hall et al (2013) for more on this). However, information economy companies tend to patent more than non-information economy businesses, whichever measure is used. In both cases, the differences are statistically significant once weighted patents are used. For the subset of companies with at least one patent (Panel B), information economy companies again patent more than those outside the information economy, but differences are not statistically significant.

[30] See OECD OECD (2009). OECD Patent Statistics Manual. Paris, OECD. for an overview of the use of patent data in economic analysis, and discussion of the EPO and other patent filing systems. At this stage we do not look at IPC filings, restrict to granted patents or weight patents by citations. All of these steps are feasible for future analysis.

Table 14. Average patent counts at company level, all years.

Company type	A. All companies			B. Companies that patent		
	Obs	Average patent count		Obs	Average patent count	
		Weighted	Unweighted		Weighted	Unweighted
Other	1,785,805	0.0199	0.0351	7,812	4.56	8.03
Information economy (sector)	156,977	0.0397	0.0561	1,057	5.89	8.32
All	1,942,782	0.0215	0.0368	8,869	4.71	8.06
Different?		**	*N*		*N*	*N*
Other	1,716,983	0.0198	0.035	7,341	4.64	8.2
Information economy (Gi sector-product)	225,799	0.0343	0.0501	1,528	5.06	7.41
All	1,942,782	0.0215	0.0368	8,869	4.71	8.06
Different?		**	*N*		*N*	*N*

Notes: two-tailed T-test, * = significant at 10%, ** 5%, *** 1%.

While information economy companies are higher-patenting, they are in the minority both in the wider data and in the patenting sub-sample. So the majority of patenting is done *outside* the information economy, as the analysis below will highlight.

Next, we look at the distribution of patents across technology fields, using the OST7 classification developed by Schmoch (2008). Table 15 gives the breakdown. Looking across all patents (Panel A), we can see that about 70% of activity is covered by the first four classes (electrical engineering and electronics, instruments, chemicals and pharma/biotech) with mechanical engineering taking the next largest share. By contrast, information economy companies' patenting is heavily orientated towards electrical engineering and electronics, followed by instruments (panels B and C). The spread across classes is more even when Gi measures are used.

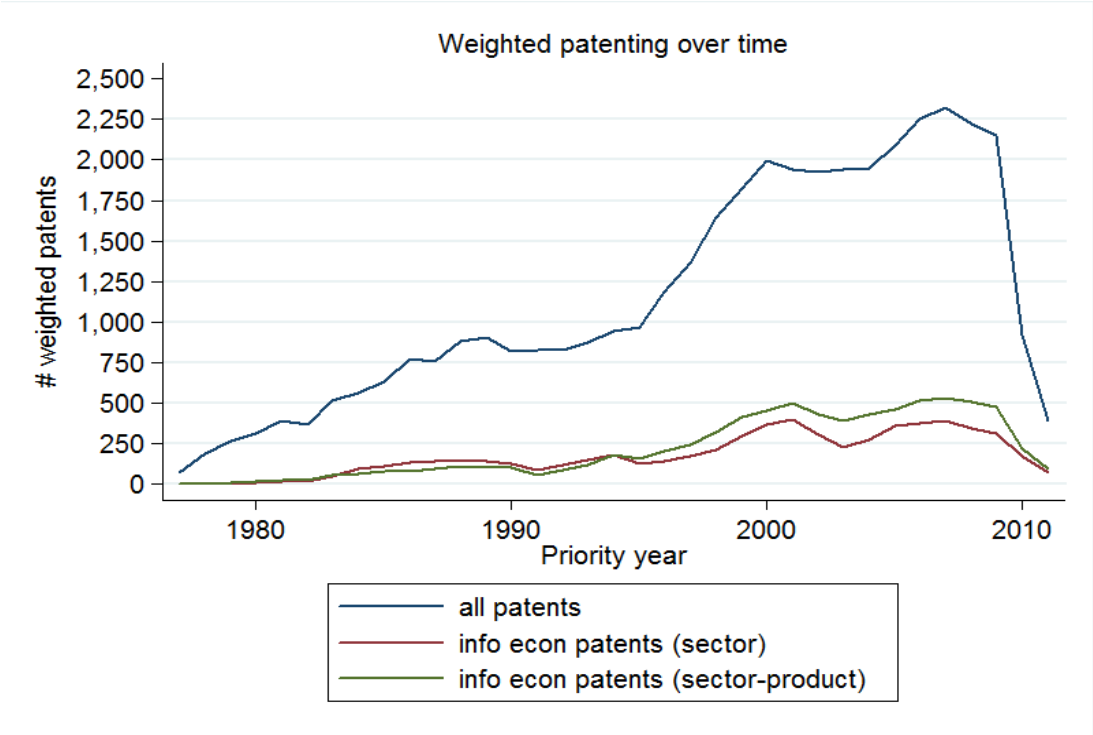
Table 15. Breakdown of unweighted patent types, all years.

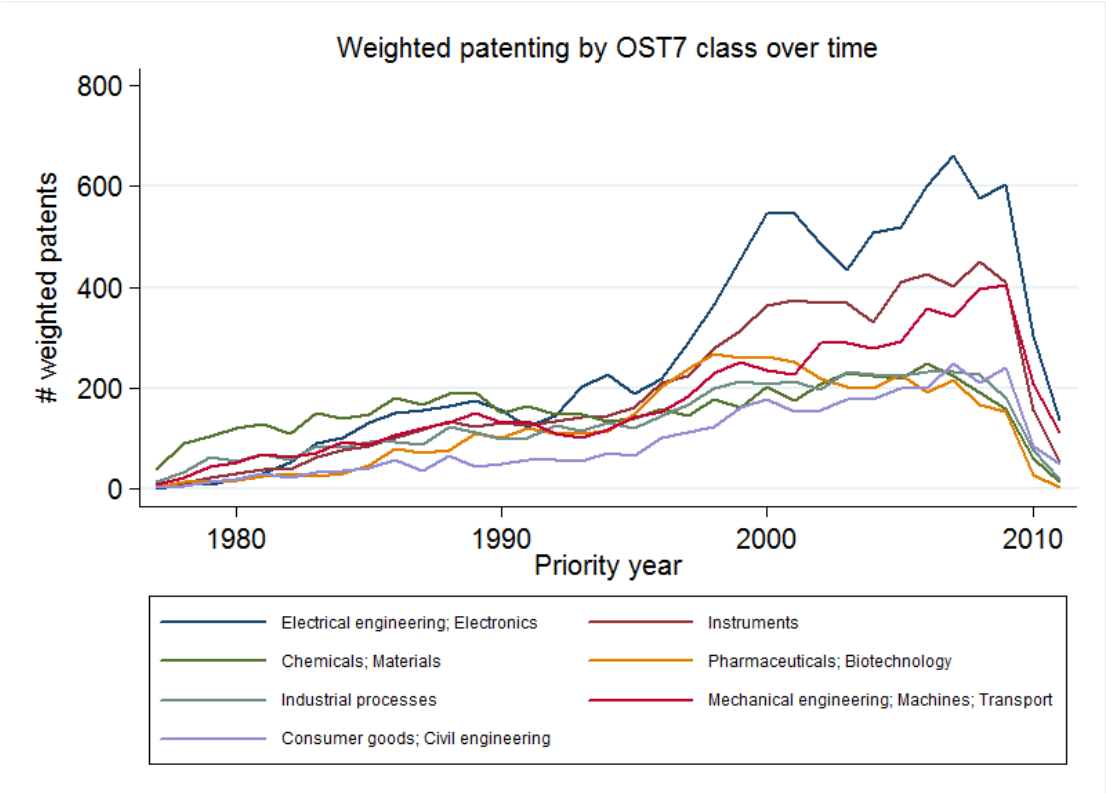
ISI-OST-INPI 7-fold technology class	Obs.	Percent	Cumulative
A. Pooled			
Electrical engineering; Electronics	13,359	20.93	20.93
Instruments	10,398	16.29	37.22
Chemicals; Materials	10,958	17.17	54.38
Pharmaceuticals; Biotechnology	9,605	15.05	69.43
Industrial processes	6,959	10.9	80.33
Mechanical engineering; Machines; Transport	8,050	12.61	92.94
Consumer goods; Civil engineering	4,508	7.06	100
Total	*63,837*	*100*	
B. Information economy (sector)			
Electrical engineering; Electronics	6,307	73.75	73.75
Instruments	1,720	20.11	93.86
Chemicals; Materials	49	0.57	94.43
Pharmaceuticals; Biotechnology	36	0.42	94.86
Industrial processes	118	1.38	96.23
Mechanical engineering; Machines; Transport	168	1.96	98.2
Consumer goods; Civil engineering	154	1.8	100
Total	*8,552*	*100*	
C. Information economy (sector-product)			
Electrical engineering; Electronics	7,445	67.96	67.96
Instruments	2,117	19.32	87.28
Chemicals; Materials	168	1.53	88.82
Pharmaceuticals; Biotechnology	100	0.91	89.73
Industrial processes	347	3.17	92.9
Mechanical engineering; Machines; Transport	459	4.19	97.09
Consumer goods; Civil engineering	319	2.91	100
Total	*10,955*	*100*	

Note: patents are unweighted. Distributions A-C are different at 1% (two-tailed test).

Using Gi definitions, we can see that information economy firms undertake more than half of electrical engineering and electronics patenting, and around a fifth of instruments patenting (these fall to 45% and 16% when SIC-based definitions are used). However, note that this analysis is done on unweighted patents, so does not take into account the number of applicants per patent. Weighted distributions will differ depending on the extent of co-patenting across technology fields, and inside / outside the information economy.

Figure 7. Patenting over time, overall and by OST7 technology class.



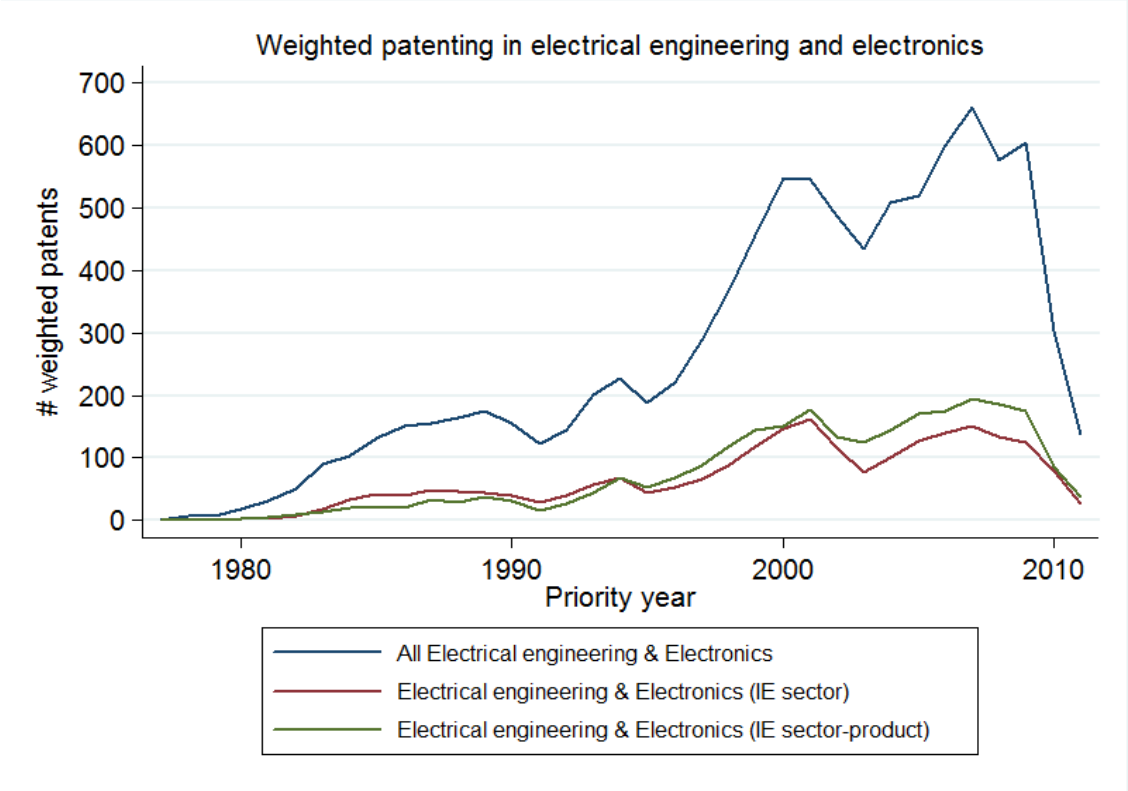


NB Patent counts are weighted by #applicants.

We then turn to patenting over time (Figure 7). The top panel shows the overall distribution, weighted by applicants on the patent, plus the information economy trend. We can see a rapid growth in patenting overall, while information economy activity rises much more gently. The bottom panel looks at patenting across OST7 technology fields. We can see very strong growth in electrical engineering and electronics patents, some increase in instruments and in mechanical engineering, then weaker growth in other fields.

The rapid growth in electronics and electrical engineering is partly driven by a spike in software patenting, which in turn is partly driven by changes in US IP legislation in the early 1990s (Li and Pai 2010).

Figure 8. Weighted and unweighted patenting in electronics and electrical engineering.



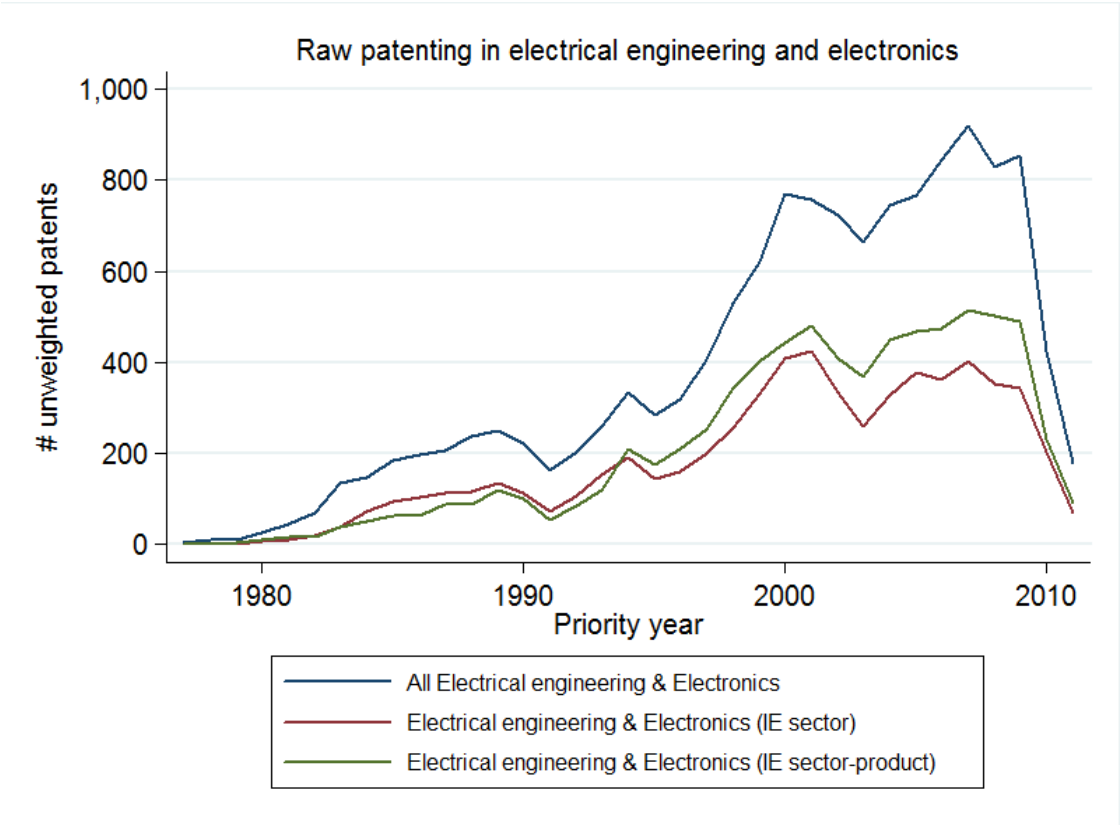


Figure 8 focuses on this technology field in more detail, and sets out the information economy share of activity. The top panel shows raw patent activity, the bottom panel activity with patents weighted by applicants. The raw patents analysis shows a very high share of information economy activity in overall patenting, which corresponds with the breakdowns in Table 16. However, once we control for the number of applicants on each patent, the information economy share drops significantly. This suggests a substantial amount of co-patenting by information economy businesses, which does not apply so much to other firms patenting in this technology field.[31]

7.6 / Trademarking

As with patents, trademarks (TMs) also provide some indication of firms' intellectual property holdings and the innovative activity underlying this(Mendonca, Pereira et al. 2004). There are also differences; patents typically indicate investments in technical knowledge, while trademarks are more closely associated with marketing strategy (Sandner and Block

[31] We test for the presence of patents which have an 'information economy' applicant and at least non-information economy co-applicant. We find 257 occurrences (0.4% of all patents).

2011). Specifically, while patents are granted for ideas developed, trademarks can be granted against future IP - for example, a name or slogan that may be used in the future for a product that does not yet exist. As measures of *innovation*, therefore, TMs are not clear-cut; as broader indicators of strategic IP activity they are very useful.

At this stage in the analysis our trademarks data is a single slice of 14,637 trademarks live in 2012-2014, taken from the UK IPO journal and matched to companies in the benchmarking sample. The overall match rate is 61.5%, for 5559 companies holding at least one TM. Even taking non matches into account, this implies that the majority of firms in our benchmarking sample do not use TMs at all, a finding echoed in Hall et al (2013). Firms in the sub-sample are on average 12 years older than those outside, and have significantly higher average revenues.

Trademarks are classified using 46 'NICE' classes,[32] and can be listed in multiple classes (although over 80% of our trademarks have three or fewer classes). For simplicity, we organise TMs into four mutually exclusive groups covering manufacturing, food and drink, services, and hybrid (covering at least one of the previous three classes). We also identify technology-oriented trademarks within these groups.[33]

Table 16 shows trademarking activity for 2012-2014, across the full benchmarking sample (Panel A) and the subset of firms with at least one live TM (Panel B). As noted above, the majority of firms hold no trademarks, so counts in the pooled sample are very low. Counts in the sub-sample are higher, with the average firm holding just over 1.6 trademarks (Panel B). Notably, overall holdings inside the information economy are always significantly smaller than outside, and smaller than the underlying sample average. This compares to patenting, where firms in the information economy hold more patents than non-IE counterparts (and the average firm).

[32] See http://oami.europa.eu/ec2 (accessed 13 May 2014).

[33] 'Manufacturing' covers NICE classes 1-28, 'food and drink' 29-34, 'services' 35-46. Within these, 'manufacturing tech' covers NICE classes 7 ("machines and machine tools") and 9 ("scientific instruments, audio, video, computers"); 'services tech' covers NICE classes 38 ("Telecommunications"), 41 ("Education, training, entertainment"), 42 ("Scientific and technological services including software") . We find no technologically orientated NICE classes in the food and drink group.

Table 16. Trademarking activity, all years.

Company type	A. All firms		B. Firms with trademarks	
	Obs	Total trademarks	Obs	Total trademarks
Other	1,785,805	0.00459	4,954	1.66
Information Economy (sector)	156,978	0.00518	605	1.34
All	1,942,783	0.00464	5,559	1.62
Different?		***		**
Other	1,716,983	0.00462	4,730	1.68
Information Econ (sector-product)	225,800	0.00483	829	1.31
All	1,942,783	0.00464	5,559	1.62
Different?		**		**

Notes: two-tailed T-test, * = significant at 10%, ** 5%, *** 1%.

Table 17 provides a summary breakdown of trademark groups. The top panel shows that 'manufacturing' trademarks comprise around 39% of marks, followed by 'crossover' TMs (just under 26%), services (just under 24%) and food and drink (around 11%). Technology-orientated TMs comprise 37.8% of the sample: the breakdown here is rather different, with crossover and services trademarks dominating.

Table 17. Breakdown of trademark types, all years.

TM group	Obs	Percent	Cumulative
manufacturing	3,540	39.27	39.27
food and drink	1,028	11.4	50.67
services	2,135	23.68	74.35
crossover	2,312	25.65	100
Total	9,015	100	

Of which 'tech'	Obs	Percent	Cumulative
manufacturing	*781*	*23.18*	*23.18*
food and drink	.	.	.
services	*1,049*	*31.13*	*54.3*
crossover	*1,540*	*45.7*	*100*
Total	*3,370*	*100*	

Notes: Categories are mutually exclusive. 'Manufacturing' covers NICE classes 1-28, 'food and drink' 29-34, 'services' 35-46. 'Manufacturing tech' covers NICE classes 7 ("machines and machine tools") and 9 ("scientific instruments, audio, video, computers"). 'Services tech' covers NICE classes 38 ("Telecommunications"), 41 ("Education, training, entertainment"), 42 ("Scientific and technological services including software").

Table 18 looks at company trademarking in these technologically orientated NICE classes. In contrast to the whole set of TMs, here we can see significantly higher trademarking by information economy firms both in the full benchmarking sample (Panel A) and in the sub-set of trademark-holding companies (panel B).

Table 18. Trademarking activity in technology-orientated TMs, all years.

Company type	A. All firms		B. Firms with trademarks	
	Obs	Total trademarks	Obs	Total trademarks
Other	1,785,805	0.00152	4,954	0.547
Information Economy (sector)	156,978	0.0042	605	1.09
All	1,942,783	0.00173	5,559	0.606
Different?				***
Other	1,716,983	0.0015	4,730	0.543
Information Economy (Gi sector-product)	225,800	0.00355	829	0.967
All	1,942,783	0.00173	5,559	0.606
Different?				***

Notes: two-tailed T-test, * = significant at 10%, ** 5%, *** 1%.

The overall information economy share of these trademarks is 23.8% (on the Gi sector-product measure), versus 19.6% when the analysis is done on a sectoral basis. While information economy firms hold more of these trademarks than non-IE counterparts, they are a minority of firms in the overall benchmarking sample.

8 / Discussion

Governments around the world want to develop their ICT and digital industries. To do this effectively, policymakers need a clear sense of the size and characteristics of these businesses – which as we have shown, is hard to do with conventional datasets and definitions. This paper uses innovative ‘big data’ resources to perform an alternative analysis, focusing on ICT producing firms in the UK ('information economy' businesses). Exploiting a combination of public, observed and modelled variables, we develop a novel ‘sector-product’ mapping approach and use text mining to provide further detail on the activities of key sector-product cells. We argue that this provides greater precision and richness than relying on SIC codes and conventional datasets.

Overall, we find that the 'ICT production space' is around 42% larger than SIC-based estimates, with at least 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 UK's 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). Patent and technologically-orientated trademark holdings are higher for information economy businesses than for non-information economy firms, although the differences are not always statistically significant. Information economy businesses are highly clustered across the country, with very high counts in the Greater South East, notably London (especially central and east London), as well as big cities such as Manchester, Birmingham and Bristol. Looking at local clusters, we find hotpots in Middlesbrough, Aberdeen, Brighton, Cambridge and Coventry, among others.

We thus find a set of companies that is larger, more established and perhaps more resilient than popular perceptions. These results derive from the many affordances of our dataset, and from the careful cleaning and identification procedures we have employed. Some care has to be taken with the revenue and employment results, 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. The Gi dataset has excellent reach and granularity and, as we have shown, provides significant extra information on fast-changing parts of the economy. We also highlight some challenges. Like other commercial products such as FAME, which we also use here, 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, so while validation is possible it is limited by the relative lack of metadata. 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 information drawn from APIs, which places some constraints on speed of draw-down and thus the 'real-time' character of some unstructured sources. 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.

Taken together, these suggest a number of 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.

The Gi dataset suggests various avenues for future research. One is exploring co-location and clusters in more detail. 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.

References

Acemoglu, D. and D. Autor (2011). Skills, Tasks and Technologies: Implications for Employment and Earnings. Handbook of Labor Economics. C. David and A. Orley, Elsevier. Volume 4, Part B: 1043-1171.

Aghion, P., T. Besley, et al. (2013). Investing for Prosperity: Skills, Infrastructure and Innovation. Report of the LSE Growth Commission. London, Centre for Economic Performance / Institute for Government.

Aghion, P., M. Dewatripont, et al. (2012). Industrial Policy and Competition. NBER Working Paper 18048. Cambridge, Mass, NBER.

Aiginger, K. (2007). "Industrial Policy: A Dying Breed or A Re-emerging Phoenix." Journal of Industry, Competition and Trade 7(3): 297-323.

Anyadike-Danes, M. (2011). Aston University matching of BSD and FAME data: report to BIS. Birmingham, Aston Business School.

Askitas, N. and K. F. Zimmermann (2009). "Google Econometrics and Unemployment Forecasting." Applied Economics Quarterly (formerly: Konjunkturpolitik) 55(2): 107-120.

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

Baron, A., P. Rayson, et al. (2009). "Word frequency and key word statistics in historical corpus linguistics." Anglistik: International Journal of English Studies 20(1): 41-67.

Brynjolfsson, E. and L. M. Hitt (2003). "Computing Productivity: Firm-Level Evidence." The Review of Economics and Statistics 85(4): 793-808.

Cable, V. (2012). Industrial Strategy. Speech to Imperial College London. London.

Choi, H. and H. Varian (2012). "Predicting the Present with Google Trends." Economic Record 88: 2-9.

Couture, V. (2013). Valuing the Consumption Benefits of Urban Density. mimeo. Toronto, University of Toronto.

Department for Business, I. a. S. (2012). Business Population Estimates for the UK and Regions, 2012: Methodology and Quality Note. BIS Enterprise Directorate. London, BIS

Department for Business Innovation and Skills (2012). Industrial Strategy: UK sector analysis. London, BIS.

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

Department for Business Innovation and Skills, Department for Culture Media and Sport, et al. (2010). Impact Assessment for the Digital Economy Bill. London, BIS.

Di Lorenzo, G., J. Reades, et al. (2012). "Predicting personal mobility with individual and group travel histories." Environment and Planning B: Planning and Design 39(5): 838-857.

Dittmar, J. E. (2011). "Information Technology and Economic Change: The impact of the printing press." The Quarterly Journal of Economics 126(3): 1133-1172.

Dumbill, E. (2013). What is big data? An introduction to the big data landscape, O'Reilly Strata.

Duranton, G. and D. Puga (2001). "Nursery Cities: Urban Diversity, Process Innovation and the Life Cycle of Products." American Economic Review 91(5): 1454-1477.

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

Fetzer, T. (2014). Measuring Legislator Productivity: A New Approach. mimeo. London, LSE.

Foord, J. (2013). "The new boomtown? Creative city to Tech City in east London." Cities 33(August): 51-60.

Gentzkow, M. and J. M. Shapiro (2010). "What Drives Media Slant? Evidence From U.S. Daily Newspapers." Econometrica 78(1): 35-71.

Ginsberg, J., M. H. Mohebbi, et al. (2009). "Detecting influenza epidemics using search engine query data." Nature 457(7232): 1012-1014.

Hall, B. H., C. Helmers, et al. (2013). The Importance (or not) of Patents to UK Firms. National Bureau of Economic Research Working Paper 19089. Cambridge, MA, NBER.

Harrison, A. and A. Rodríguez-Clare (2009). Trade, Foreign Investment, and Industrial Policy for Developing Countries. National Bureau of Economic Research Working Paper Series No. 15261. Cambridge, MA, NBER.

Hastie, T., R. Tibshirani, et al. (2009). The Elements of Statistical Learning: Data mining, inference and prediction. Berlin, Springer.

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

Lehr, W. (2012). Measuring the Internet: The Data Challenge. OECD Digital Economy Papers 194. Paris, OECD.

Lewis, P., T. Newburn, et al. (2011). Reading the Riots: Investigating England's summer of disorder. London, LSE / The Guardian.

Li, X. and Y. Pai (2010). The Changing Geography of Innovation Activities: What do Patent Indicators Imply? The Rise of Technological Power in the South. X. Fu and L. Soete. Basingstoke, Palgrave MacMillan: 69-88.

Mendonca, S., T. S. Pereira, et al. (2004). "Trademarks as an indicator of innovation and industrial change." Research Policy 33(9): 1385-1404.

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

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

Nathan, M. and E. Vandore (2014). "Here be startups: exploring London's 'Tech City' digital cluster." Environment and Planning A 46(10): 2283-2299.

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

OECD (2009). OECD Patent Statistics Manual. Paris, OECD.

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

OECD (2013). Measuring the Internet Economy: A contribution to the research agenda. OECD Digital Economy Papers 226, OECD Publishing.

Office of National Statistics (2009). UK Standard Industrial Classification of Economic Activities 2007 (SIC 2007): Structure and explanatory notes. Basingstoke, Palgrave Macmillan.

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

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

Porter, A., J. Youtie, et al. (2008). "Refining search terms for nanotechnology." Journal of Nanoparticle Research 10(5): 715-728.

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

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

Salton, G. and C. Buckley (1988). "Term-weighting approaches in automatic text retrieval." Information Processing & Management 24(5): 513-523.

Sandner, P. G. and J. Block (2011). "The market value of R&D, patents, and trademarks." Research Policy 40(7): 969-985.

Schmoch, U. (2008). Concept of a Technology Classification for Country Comparisons. Final report to WIPO. Karlsruhe, Fraunhofer Institute for Systems and Innovation Research.

Tapscott, D. (1997). The Digital Economy: Promise and peril in the age of networked intelligence. New York, McGraw-Hill.

Varian, H. R. (2014). "Big Data: New Tricks for Econometrics." Journal of Economic Perspectives 28(2): 3-28..

APPENDICES

Appendix 1 / The Growth Intelligence dataset

Growth Intelligence (Gi) is a London-based company, founded in 2011, that provides a tool for Predictive Lead Generation to 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.

A1.1 / 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.[34] 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 extreme cases 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.

[34] 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.

A1.2 / Structured data layers

Gi match Companies House data to a series of other structured administrative datasets. In this analysis we focus on two of these. Patents data is taken from the European Patent Office PATSTAT database, and is matched to companies using name/applicant information as well as inventor location data. Patent titles and abstracts are obtained from the EPO API feed and combined with the raw data. We also use UK trademarks data, which is taken from the UK Intellectual Property Office (UK IPO) API feed. Again, this information is matched to Companies House using company name / address information. Growth Intel use 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. We discuss these modelled data layers below.

A1.3 / Modelled data layers

This part of the Gi dataset is developed through data mining (Rajaraman and Ullman 2011). Gi develop a range of raw text inputs for each company, then use feature extraction to identify key words and phrases ('tokens'), as well as contextual information ('categories'). Gi assign 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, Rayson 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.[35] 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.

For company classifications, Gi use 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.

[35] The TF-IDF approach is the workhorse method in the field (Salton and Buckey, 1988); an alternative is to use the Pearson chi^2 score (see Gentzkow and Shapiro (2010) for a recent example).

A similar supervised learning setting is for modelled 'events' data (for example, predicted product launches or joint ventures). In this case, the main inputs are tokens derived from news sources. Gi focus on industry news sources and relevant mainstream media, removing irrelevant sources (such as celebrity magazines) and some non-local sources.

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 signals in the Gi dataset.

Appendix 2 / Comparing estimates from the BSD and Companies House

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

A2.1/ 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 A1 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.

Table A1. 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	1,876,217	

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

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 A2 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%, 62020) and 'other information technology service activities' (17.96%, 27.7%, 62090).

Table A2. Information economy: shares and counts for component bins, 2011.

SIC5 sector name	BSD			CH		
	Freq.	Percent	Cum.	Freq.	Percent	Cum.
mf of electronic components	588	0.47	0.47	1,037	0.67	0.67
mf of loaded electronic boards	360	0.29	0.76	241	0.16	0.83
mf of computers and peripheral equipment	826	0.66	1.42	791	0.51	1.34
mf of telephone and telegraph equipment	1,342	1.07	2.49	700	0.45	1.8
mf of other communications equipment	163	0.13	2.62	199	0.13	1.93
mf of consumer electronics	614	0.49	3.12	487	0.32	2.25
mf of electronic measures and tests	1,578	1.26	4.38	1,050	0.68	2.93
mf of electronic industrial process control equipment	259	0.21	4.59	512	0.33	3.26
mf of non-electronic equipment not for ipc	185	0.15	4.73	42	0.03	3.29
mf of non-electronic ipc equipment	92	0.07	4.81	20	0.01	3.3
mf of optical precision instruments	123	0.1	4.91	128	0.08	3.38
mf of photographic and cinematographic equipment	88	0.07	4.98	64	0.04	3.43
mf 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	124,971	100		153,858	100	

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

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

A2.2/ 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.[36] 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 A3):

[36] 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 A3. 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	104,285	4.82	52.99	47.01
2004	93,431	4.32	57.31	42.69
2005	105,061	4.86	62.17	37.83
2006	132,971	6.15	68.33	31.67
2007	163,062	7.54	75.87	24.13
2008	150,699	6.97	82.84	17.16
2009	171,379	7.93	90.77	9.23
2010	164,360	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	114,892	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	115,940	6.18	55.28	44.72
2007	144,991	7.73	63.01	36.99
2008	135,701	7.23	70.24	29.76
2009	165,044	8.8	79.03	20.97
2010	216,961	11.56	90.6	9.4
2011	176,397	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.

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 start-ups. These differences are also noticeable in the information economy (Table A4). The differences are smaller for the set of firms 10 years old or less, but greater for start-ups:

Table A4. Age structure for information economy sectors, BSD vs Companies House 2011.

Birth year	Freq.	Percent	Cum.	Inverse
BSD				
2002	6,962	3.92	42.1	57.9
2003	8,199	4.61	46.71	53.29
2004	8,989	5.06	51.76	48.24
2005	9,903	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	177,821	100		
Companies House				
2002	5,364	3.49	29.34	70.66
2003	6,577	4.27	33.61	66.39
2004	6,748	4.39	38	62
2005	7,288	4.74	42.73	57.27
2006	9,120	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	153,858	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.

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.

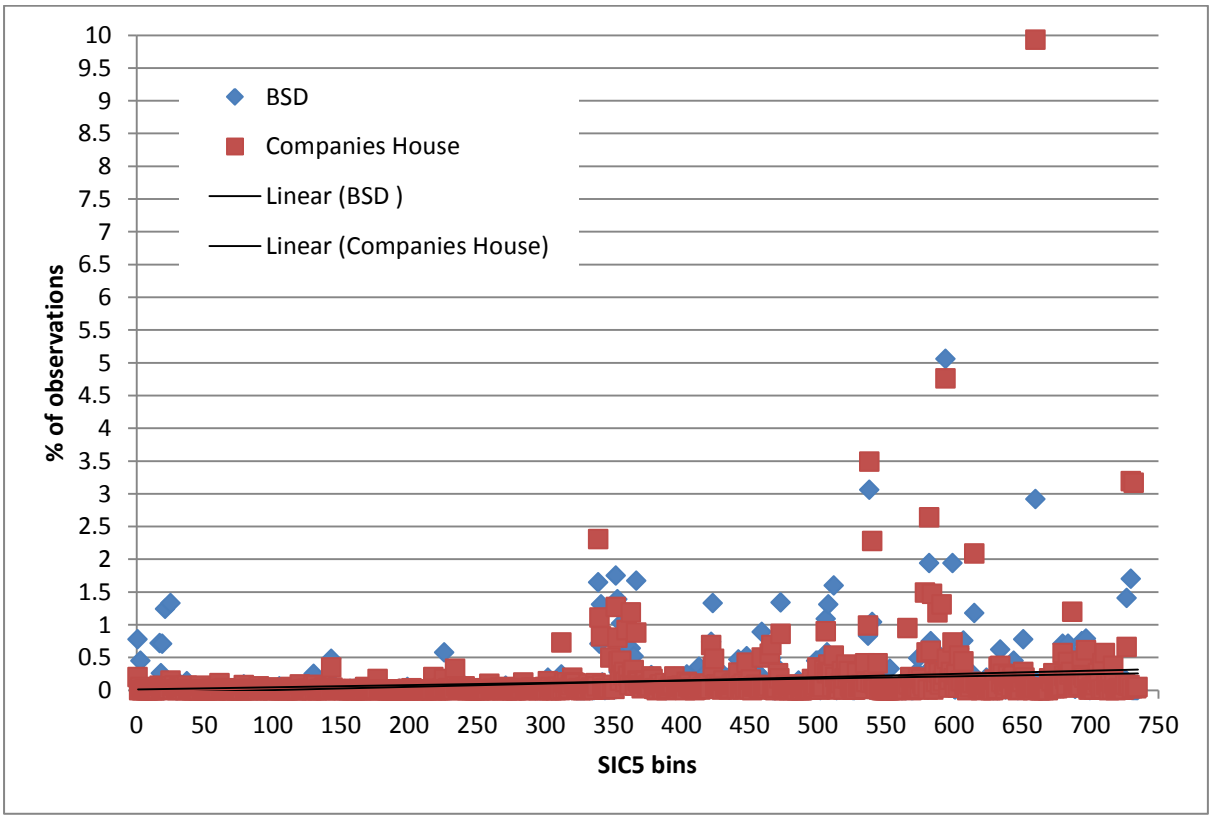
A2.3/ Sectoral distribution in the BSD and CH

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 are present 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), gathering wild growing products (19). Third, holding companies are present in the BSD but not CH because our cleaning kicks them out. In the BSD they comprise 14,281 observations, or 0.66% of the sample.

Figure A1 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.

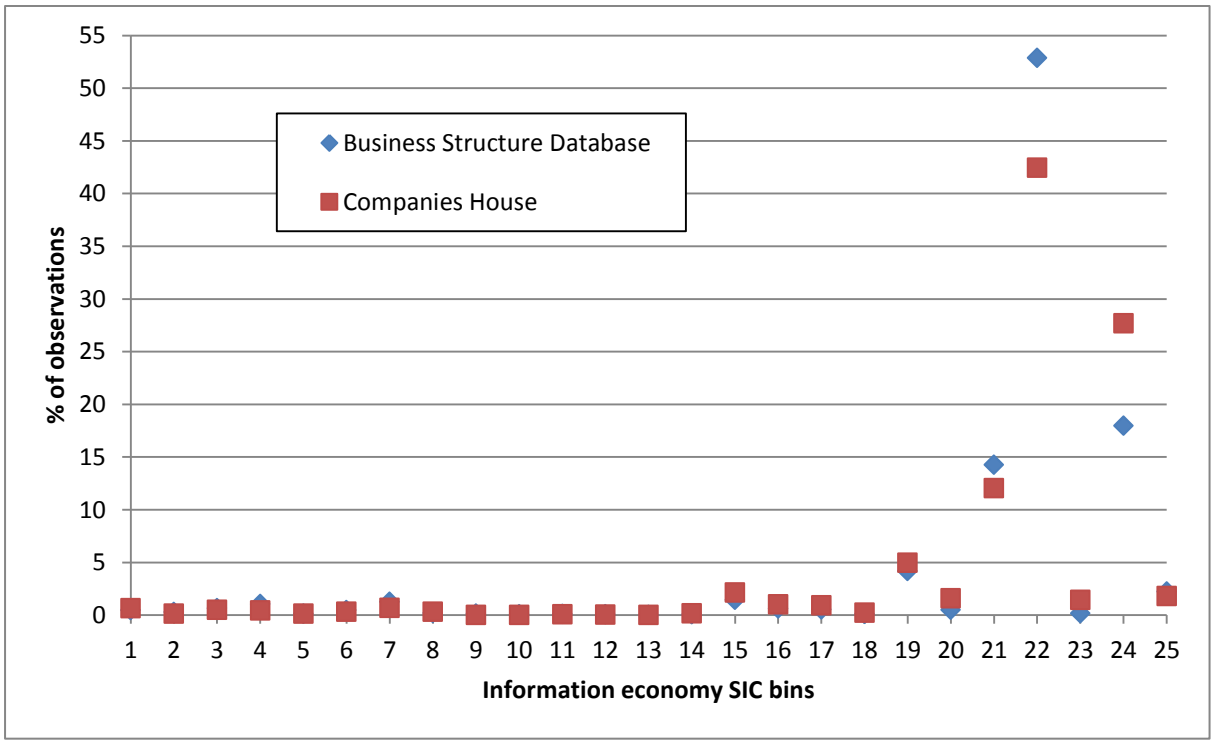
Figure A1. Comparing BSD and CH shares, all SIC5 sectors, 2011.



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

For the information economy, we can see that the matching is generally pretty 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%, 62020) and 'other information technology service activities' (17.96%, 27.7%, 62090).

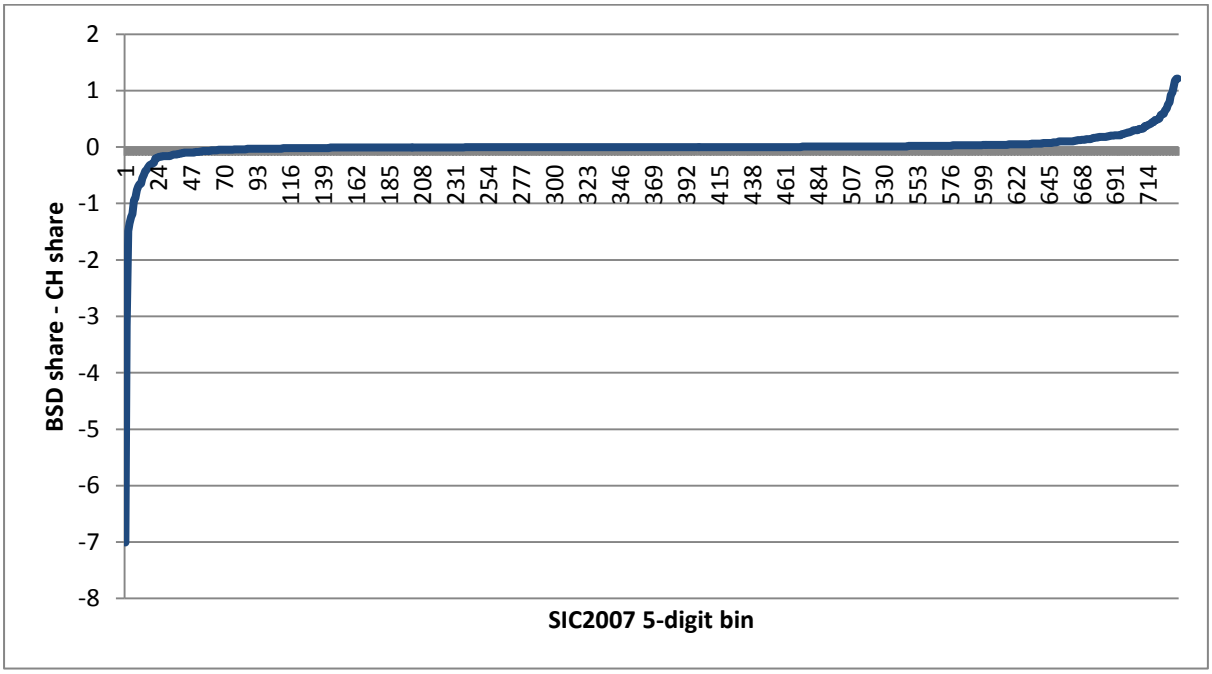
Figure A2. Comparing BSD and CH shares, info economy sectors, 2011.



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

We can see that in most cases, CH and BSD % differences are minimal / zero (Figure A3):

Figure A3. Comparing BSD and CH differences, 2011.



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

A2.4/ Exploring the extremes

We now look at the c. 10% of SIC bins where the differences are most pronounced (tables 4 and 5, 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).[37]

2.4.1 / CH > BSD shares

First we look at the bins where sector shares are higher in CH than the BSD. Results are given in Table A5. 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 these differences. Four of these bins are 'information economy' sectors (see highlights key). In particular, there are far more CH firms in 62090, 'other information technology service activities', than in the BSD. 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

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

age structure. The relevant firms in Companies House are much younger than their BSD counterparts.

Table A5. 5% of SIC5 bins with largest CH-BSD differences, 2011.

SIC2007 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. Yellow = information economy SIC5 bin.

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.

2.4.2/ BSD > CH shares

Results are set out in Table A6. 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.

Table A6. 5% of SIC5 bins with largest BSD-CH differences, 2011.

SIC2007 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
general 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. Blue = DE only, yellow = IE only, green = 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.

A2.5/ 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.

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.