G Model
RESPOL-3108; No. of Pages 20
ARTICLE IN PRESS

Research Policy xxx (2015) xxx–xxx

Contents lists available at ScienceDirect

Research Policy

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

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

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

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

ARTICLE INFO

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

JEL classification:
C55
C81
L63
L86
O38

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

ABSTRACT

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

1. Introduction

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

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

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

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

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

http://dx.doi.org/10.1016/j.respol.2015.01.008
0048-7333/

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

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

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

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

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

2. Framework

2.1. Definitions

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

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

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

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

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

2.2. Data challenges

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

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

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

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

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