CENTRO STUDI LUCA D'AGLIANO

WWW.DAGLIANO.UNIMI.IT

CENTRO STUDI LUCA D'AGLIANO
DEVELOPMENT STUDIES WORKING PAPERS

N. 429

July 2017

Innovative Events

*Max Nathan**
*Anna Rosso***

* University of Birmingham
** University of Milan and Centro Studi Luca d'Agliano

ISSN 2282-5452

INNOVATIVE EVENTS

Max Nathan[1] and Anna Rosso[2]

[1] University of Birmingham. m.nathan@bham.ac.uk
[2] University of Milan and Centro Studi Luca d'Agliano. anna.rosso@unimi.it

Abstract [100 words]

Policymakers need to understand innovation in high-profile sectors like technology. This can be surprisingly hard to observe. We combine UK administrative microdata, media and website content to develop experimental measures of firm innovation – new product/service launches – that complement existing metrics. We then explore the innovative performance of technology sector SMEs – firms also of great policy interest – using panel fixed effects settings, comparing conventional and machine-learning-based definitions of industry space. For companies with event coverage, tech SMEs are substantially more launch-active than non-tech firms, with suggestive evidence of firm-city interactions. We use instruments and reweighting to handle underlying event exposure probabilities.

Keywords: innovation, ICT, data science
JEL: C55, L86, O81

Acknowledgements

Thanks to seminar/conference participants at ZEW Economics of ICT workshop 2017, 4th VPDE-BRICK workshop, SpazioDati, CTCS 2016, Uddevalla 2016, AAG 2016, University of Birmingham, CIRCLE and NIESR, and to Cathy Atkinson, Christian Catalini, Steve Dempsey, Rob Elliott, Bill Kerr, Francesco Rentocchini and Brandon Stewart for helpful comments. Many thanks to Tom Gatten, Prash Majmudar and Francois Bouet at Growth Intelligence for data. Aslam Ghumra, Neeraj Baruah and Christoph Stich provided outstanding technical assistance. 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. This research builds on projects funded by NESTA and by Google UK. It represents our views not theirs.

1/ Introduction

Researchers and policymakers are keen to understand innovative activity in 'high-value' industries. There is particularly sharp debate about the innovation contribution (or not) of digital technology / ICT industries and firms (Gordon, 2016; Mokyr, 2013; OECD, 2013; Cowen, 2011). Resolving such debates is hard because conventional innovation metrics have well-known limitations: in practice, most firms do not use formal IP protection, with one recent UK study finding that just 1.6% of firms file patents. More fundamentally, standard industry typologies also tend to lag real-world firm-level change, so that even classifying firms on the technological frontier is hard.[1] Both issues leave effective industrial policy harder to design. This paper makes two linked contributions to these challenges. First, we use a novel mix of UK administrative microdata, media and website content to develop experimental measures of firm innovation, specifically new product/service launches, which complement existing metrics (Hall and Harhoff 2012, Marisse and Mohnen 2010). We argue these give insight into 'downstream' innovative activity, and provide evidence to support this. Second, we use this platform to explore the innovative performance of firms often targeted in industrial strategies – digital technology small and medium-size enterprises (SMEs) – and run comparisons using conventional and machine-learning-based definitions of industry space.[2]

To do this, we exploit a cutting-edge dataset developed by the data science firm Growth Intelligence (GI), which uses machine-learning routines on company website and media content to model firms' lifecycle 'events' (for example, new product/service launches, mergers and acquisitions, new hiring, or joint ventures). We extensively refine these

[1] This number comes from Hall et al (2013). Hall et al (2014) provide an extensive review of firms' choices of formal vs. informal IP protection mechanisms. Nathan and Rosso (2015) discuss typology lags in detail.
[2] For background, see Aghion (2016) on innovation, entrepreneurship and growth policy, Mazzucato (2011) on the state's role in innovation, especially ICT; and Chatterji et al (2014) on high-tech clusters.

variables, using structural topic modelling to better align the reported GI data with underlying real-world activity. We run a series of checks for error rates in GI's web crawling and feature extraction routines on a data sample where errors are *a priori* most likely. We build a panel dataset matching these 'innovative events' variables to UK administrative firm level microdata 2014-15, plus patents and trademarks data. We focus on single plant small and medium-size enterprises (SMEs), which most cleanly links events to firms and locations. This dataset of 2.4m observations is the first firm-level resource of its kind that we are aware of.[3]

Our resulting innovation measures complement existing metrics such as patents, trademarks and surveys, providing high volume (for single-plant SMEs we find 10,349 UK launches in 2014/2015, versus 2,908 patent applications and 3,902 trademarks filed) and even cross-sector coverage. Daily launch data offers potential for high-frequency applications. Using a simple firm-level framework that incorporates knowledge production and absorptive capacity, we show positive links from past IP activity to current launches (and not vice versa), with cross-industry variation along expected lines.

We then use this platform to explore the innovative performance of UK digital technology SMEs, a closely related area of interest. Innovation and growth policies often seek to target small/young tech firms, given their supposed high productivity and job growth potential (Decker et al., 2016; Coutu, 2014). We adapt a two-stage specification developed by Combes et al (2008) to recover the role of tech firm status from firm fixed effects estimates. We pay careful attention to the fact that events exposure is not random, and that events are reported,

[3] Existing datasets of news events such as GDELT and Events Registry are designed for country-level analysis, especially politics / current affairs. Proprietary firm-level datasets such as Mattermark (US) and Beauhurst (UK) are restricted to small numbers of 'high-potential' businesses. Crunchbase is a global wiki-type dataset for the tech sector with good US coverage but limited coverage for other countries, as well as significant quality concerns due to its self-reported nature (Motoyama and Bell-Masterson, 2014).

not directly observed; control for a range of factors affecting launch decisions, timing and wider conditions, and test different definitions of sector space.

The paper provides a fresh contribution to well-established literatures on innovation in firms – where patents and surveys remain the dominant metrics – by developing new ways to view firms' innovative activity based on web scraping and natural language processing.[4] This allows us broader coverage than the pioneering studies of media coverage and innovation, such as Katila and Ahuja (2002) and Fosfuri et al (2008), which are restricted to a few hundred firms in single sectors. We are also able to advance on rich data papers such as Hall et al (2013), who combine conventional administrative data, patents, trademarks and innovation surveys for 8,600 UK firms. Our paper further contributes to a growing empirical literature that uses 'big data' sources and data science techniques.[5] Unusually for this field, we combine commercial big data and large, high quality administrative data sets. The latter provides a clear sampling frame that helps understand implicit sampling issues in the former, substantively aiding inference and interpretation (Einav and Levin, 2014).

2/ Data

Modelled company 'lifecycle events' are at the heart of our empirical approach. Each event derives from content taken from 3,740 online news sources (including major sources such as Reuters or Yahoo news, as well as industry sources such as IT Briefing and PRWeb. Our raw

[4] See Arora et al (2017), Hall and Harhoff (2012), and Mairesse and Mohnen (2010) for recent reviews of patents and surveys in innovation research. Trademarks are increasingly used alongside these metrics; see Block et al (2015). Gentzkow et al (2017) review economic applications of natural language processing.

[5] For reviews see Varian (2014) and Einav and Levin (2014). Guzman and Stern (2015) provide an example of 'nowcasting' and 'placecasting' entrepreneurship, using cross-validation on a very large sample of US data.

data consists of 318,899 observations covering financial years 2014 and 2015 (specifically, August 2013 to November 2014 inclusive). Figure 1 provides two examples for product / service launches, the event type we focus on. GI match raw events observations to a UK-wide company register, Companies House, using firm names and contextual information. They then use supervised learning to classify the activity described as one of several event types. We focus on 'product/service launch'; other types include 'alliance/joint venture', 'contract awarded', 'management change' and 'merger/acquisition'. Classification uses both event text and information from company websites. Nathan and Rosso (2015) give more detail.[6]

Figure 1 about here

The intuition behind using 'events' is that one can exploit how companies describe themselves or their activities – and how these are reported by others – to understand things that companies do or that happen to them. Ideally, each event observation represents a distinct thing that happened to some firm, or an action that the firm does. In practice, we need to substantially clean the data to get closer to this ideal. Cleaning details are provided in Appendix A1 and summarised here.

We first remove duplicates and control for 'farmed' content.[7] We next run a quality check on GI's syntax parsing and matching routines, for a sample of 5,000 'hard cases' where *a priori*, ascription errors are most likely to occur. Specifically, we sample observations ascribed to

[6] Text fragment for illustration. GI use the full page of content to assign text to a subject company and to classify the activity. Where text describes more than one subject company, as in mergers or joint ventures, GI assign to pair or n-groups. GI also filter to remove results from irrelevant domains (for example, mentions of companies in celebrity magazines, or results from sites that largely or wholly deal in markets outside the UK).

[7] Recent structural changes to the media industry – notably, the rise of online platforms – may be reducing levels of quality and scrutiny, for example through 'content farming' and 'churnalism' (Viner, 2016; Gentzkow and Shapiro, 2010; Davies, 2009). The first leads to duplicate reported events; the second alter the distribution of event activity. Both may be particularly prevalent in the ICT sector (Lafrance, 2016). We identify duplicate observations events using all available variables except the source and time. Within each group we just keep one event, so that we are not selecting events on the basis of the quality of the source.

the largest ICT/tech companies by revenue (such as Google, Facebook and Microsoft), or to the largest media companies by market share (Reuters, PA, PR Newswire). These are cases where company names are likely to feature as context as well as subject, so that website content might be especially error-prone. Analysis using title and text fragment fields suggests around 16% error rates. Note that our focus on single plant SMEs removes these hard cases from the data, minimising the ascription error rate on the rest of the sample. However, to the extent that mis-ascription 'gives' events to large tech and media firms that actually belong to SMEs, we have a lower bound on the true level of event activity for our firms of interest. Full details are given in Appendix A1.

A further, crucial cleaning issue is that in its raw form, event data does not reflect the importance of the underlying (real world) event. For example, a major product launch is likely to be reported hundreds of times; in the raw data each is reported as a distinct event. We use structural topic modelling (STM) to deal with this. Topic modelling algorithms cluster text fragments that talk about the same topic in different ways, using different text but similar content words (Roberts et al., 2016). We exploit this approach to cluster raw events with similar content – that are likely to refer to the same real-world event – into single reported instances. STM substantially reduces the count of event observations, to 202,912. Again, full details are given in Appendix A1.[8]

2.1 / Panel

We combine cleaned events data with other sources. Our 'base layer' is the Business Structure Database (BSD) (Office of National Statistics, 2016). This high quality

[8] Future analysis could also exploit datasets like Prodcom, to give a sense of product/service quality.

administrative microdata covers 99% of UK enterprises, and gives a clearly defined sampling frame. Our 'bridging layer' is the Companies House dataset, an open dataset of UK-registered companies that provides unique company identifiers. Gi data comes pre-matched to Companies House; we work with the UK Data Service Secure Lab to match an anonymised version of this data to firms in the BSD. Finally, we use various matching routines[9] to link US, European and other patents data (from Orbis, application years 1950-2015) and UK trademarks data (from the UK Intellectual Property Office, 2012 - 2015) to Companies House, and thus to the BSD. The build is described in detail in Appendix A2.

We restrict the sample to single plant SMEs, allowing us to cleanly ascribe events to single firms and locations. This has the effect of substantively reducing the total count of event observations, from 202,912 to 81,323.

2.2 / Descriptive analysis

The panel comprises 2.406m observations for 1.22m single plant SMEs in the financial years 2014-2015 (that is, 2013/2014 and 2014/2015). Firms are mobile across industries (given by 4-digit SIC2003 codes) and across locations (here, Travel to Work Areas). 2.88% of firms are SIC movers, 2.24% of firms are TTWA movers. We cap the maximum number of events per firm to 1,000 to remove outliers (this drops four observations). Table A1 in Appendix A3 provides summary statistics, and Appendix A2 describes variables in detail.

Table 1 about here

[9] Bureau Van Dijk identifiers; or firm name and full postcode. An alternative approach would be the automated method developed by Autor et al (2016) which exploits internet search results.

Table 1 gives more detail on events coverage. Around 1% of firms have event coverage, and around a third of these have product / service launches. Firms with event coverage have around 2.5 events, of which around 1.5 are product/service launches. This distribution is broadly stable across the two sample years.

Event exposure is not random. Table 2 shows the mean characteristics of firms with and without reported events, averaged over 2014 and 2015. Using rank-sum tests, we find significant mean differences between the two samples.[10] Firms with reported events are on average older, bigger in terms of employment and revenue, with higher employment, revenue and revenue productivity growth, and are more IP-active (with more patents and trademarks filed). However, they are less likely to be foreign owned, and firms without events are more likely to be 'gazelles'.[11]

Table 2 about here

3/ Launches as an innovation measure

Innovation in firms is a multi-stage process. The knowledge production function paradigm pioneered by Griliches (1979) links upstream inputs (internal R&D, external knowledge), intermediate outputs (inventions) and firm 'performance' (productivity, stock market value and so on). Performance is partly driven by inventions successfully deployed internally, and/or commercialised (innovations). In practice, knowledge may emerge from interactions

[10] All results are significant at 1%. Rank-sum tests are preferred, as we do not know the underlying distribution of events. T-tests give virtually identical results. Median differences are much smaller.
[11] OECD definition. High growth firms have at least 10 workers in a given year, then exhibit annual growth of 20% or more in employment, revenue or revenue productivity over the following three years. Gazelles are high growth firms five years old or less.

with customers, suppliers and peers (Chesborough, 2003; Von Hippel 2005) as well as a firm's asset base, and is shaped by firms' absorptive capacity and evolution paths (Cohen and Levinthal, 1990; Blundell et al, 1995; Teece at al 1997).

Intermediates are product/process innovations protected either formally (via patents, trademarking or designs) or informally (via secrecy, confidentiality agreements or lead times) (Hall et al, 2014). Surveys suggest that firms typically use a range of IP protection tools, both formal and informal, if they do so at all (*ibid*). In particular, trademarks are an important formal complement to patents, both for IP protection (via legal protection for brands and marketing assets), but also to aid product differentiation and as a way to signal innovativeness to potential investors (Block et al 2015, Helmers and Rogers 2010, Fosfuri et al 2008). Among informal tools, lead time seems the most widely used (Hall et al, 2014).

We argue that product/service launch events – like those in Figure 1 – are a measure of firms' 'downstream' innovative activity: specifically, they represent inventions that have been commercialised into new-to-the-firm products and services. Crucially, while launches do not capture innovations protected by secrecy, they can (in theory) pick up *any* other public innovation however protected. Lags between formal IP filings and launches may also give some insight into lead time decisions.

Table 3 shows why launches are a useful complement to conventional innovation metrics, showing the coverage of patents, trademarks and reported launches at aggregated industry level in our data. As expected, patenting is most concentrated in manufacturing, but is also present in parts of the services sector, notably business services (including software and other 'knowledge-intensive' activities (Castellacci et al, 2008)). Given their broader functionality,

trademarks are more evenly distributed, with most activity in manufacturing, wholesale /retail/repair and social/personal services. Launch activity is more even and higher-frequency than either metric: we observe 10,349 launches in the financial years 2014 and 2015. In the same period, by contrast, we observe 2,908 patents and 3,902 trademarks.

Table 3 about here

If launches truly represent downstream innovative activity by firms, we should expect a positive significant link from past IP activity (reflecting upstream invention and indirectly, R&D) to launch activity (but not vice versa). Given their respective functions, we might also expect larger/stronger links from patenting than from trademarking. Raw industry-level correlations seem to bear this out (Figure 2).

Figure 2 about here

More formally, we can represent these relationships as a modified knowledge production function, in which launch activity L for firm *i* in year *t* is a function of past observable inventive activity in period *t-n*, firm characteristics, and wider local (*a*) and sectoral (*s*) conditions:

$$L_{itas} = a + b\text{PATS}_{it-n} + c\text{TM}_{it-n} + \boldsymbol{d}\mathbf{PAST\_P}_i + \boldsymbol{e}\mathbf{X}_{it} + T_t + A_a + S_s + e_{itas} \quad (1)$$

We define L as either a product launch dummy taking the value 0 or 1, or the count of launches *l*, where *l* = 0, ... *l*. PATS are patent stocks with a standard 15% depreciation rate (Hall and Harhoff 2012), which we vary in sensitivity tests. TM stocks are constructed the

same way. We are interested in the links from 'recent' (PATS) and 'historic' (Past_P) IP to launches, where historic IP stocks act as proxies for firms' absorptive capacity, including R&D activity (Blundell et al., 1995).[12] We define 'recent' patenting as occurring in any given five year period, so that *n* takes the value 0, 1 ... 5 for patents, for EPO/US/PCT filings in any given year back to 2009.[13] We define 'historic' patenting as taking place pre-2009. Following Blundell et al, **PAST_P** includes a dummy taking the value 1 if the firm patents before this date, and an average of pre-2009 patenting activity which takes values $p = 0, \ldots p$. For trademarks, *n* takes the value 0, 1 or 2 given available data. **X** includes predictors of firm growth such as lagged log turnover, age, startup dummy and company legal status dummies (public company or sole trader, with 'other' the reference category). We include area and 4-digit industry fixed effects as well as a time dummy. Areas are Travel to Work Areas, approximating local spatial economies.

We estimate in OLS using standard errors clustered on two-digit SICs. This is because nonlinear estimates typically converge to OLS results once converted to marginal effects (Angrist and Pischke, 2009); we later test this assumption in robustness checks. OLS is also more efficient given the very large number of fixed effects in our data. Note also that measurement error on both sides of (1) will affect our estimates. First, the majority of UK innovations are not protected with formal IP, even for R&D-intensive companies (Hall et al., 2013). Many new products / services involve multiple patents (e.g. the iPhone reportedly has over 100).[14] Second, reported launches are likely to be a lower bound on true levels of launch

[12] KPF approaches normally include R&D. Our data makes this challenging. Government surveys such as CIS and BERD are too small to match to our sample; commercial sources such as Orbis have limited direct coverage (7,600 'industrial companies' in the UK with R&D expenditure in annual accounts); UK SMEs only need to file minimal returns to Companies House, so that standard proxies are hard to reconstruct. For this reason we rely on past patenting and trademarks to provide a (lower bound) approximation of underlying R&D.

[13] We use filings to these offices as a proxy for invention quality: inventions filed in international domains rather than a single country will be 'worth' more for applicants (Helmers and Rogers 2010). Alternatives would be triadic patent family constructs as an ex-ante measure of quality, or patent citations as an ex-post measure.

[14] E.g: https://www.quora.com/How-many-patents-does-the-iPhone-use?, accessed 23 February 2017.

activity. We also know (from Section 2) that GI's modelling has ascription error that tends to allocate events away from single plant SMEs, the firms in our sample. Third, we are testing aggregate links for each firm using many years of patents and TMs, but only two financial years' worth of reported launches. While the measurement error may downward bias the estimates, we can consider the error in the product launch as good as random, conditional on the set of variables we have in the regression.

We start with a simple stepwise approach. Here *n* is set at 1, so all right hand side variables are lagged one period. Results are given in Table 4. Columns progressively add controls (column 2), past patents and trademarks (column 3), year (c4), area (c5) and sector (c6) effects.[15]

Table 4 about here

For the linear probability model (top panel) we can see that coefficients of past IP remain significant, although as expected, adding controls and fixed effects reduces coefficient size. Note that historic (pre-2009) patenting activity is a significant predictor of current launch activity, with coefficients substantially larger than the 1-year lag of patenting. We see similar patterns for the launch counts model (bottom panel), although in this case historic IP is not statistically significant.

In Table A2 in the appendix we cross-check the specification using firm fixed effects rather than PAST_IP, to better account for heterogeneity in firm-level innovative activity. Again, *n* is set at 1, and columns progressively add controls, firm fixed effects, then year. As expected,

[15] We also run regressions with IPC1 technology field*year effects, with nine count variables for each instance of patenting activity in each IPC1 field in each year. Results are identical to the fullest specification here.

this estimation strategy is too restrictive given the structure of our dataset: because we are using a short panel of firms, most of the variation used in the estimation is between firms. For these reasons, we prefer to control for firm heterogeneity via past IP activity.

Table 5 gives results for varying *t-n* lags. In the top panel, which shows the dummy model each 10 additions to patent stocks in a given year of a launch raises the probability of a launch in that year by 6% points (column 1); for lagged patenting, the size of the link varies between 6 and 9% points. Historic patenting is always significant, with a 3.3-3.6% point link from *any* historic patenting to launches. For counts (bottom panel), 10 additions to patent stocks in a given year is linked to just over 0.2 extra launch events. For lagged patenting, the link varies between 0.2 and 0.46 extra launches. Historic patenting and trademarking links are weaker, with the latter not always significant.

Table 5 about here

We extend the analysis in two robustness checks. First, in Appendix Tables A3 and A4 we change the specification of PATS to be cumulative (top panel) or with a 40% depreciation rate (bottom panel), following Li and Hall (2016). Neither re-specification changes the results substantively, although patent coefficients are slightly bigger with a 40% depreciation rate than with the standard 15% specification. Second, Table 3 suggests that patenting and trademarking have uneven coverage across manufacturing and services industries. Appendix tables A5 and A6 therefore split out the sample into firms in 'manufacturing' (SIC sections A-D) and 'services' (SIC sections G-O). We find that recent patent stocks are more likely to lead to a launch for manufacturing firms than for those in services, and produces more launches. Services firms are more likely to launch and conduct more launches for patent

stocks 3-5 years previously. As shown by Table 3, these firms are most likely to be in IP-intensive activities such as software or other business services.[16]

Finally, to crosscheck the direction of the IP-launch relationship, we run a cross-sectional placebo test where we regress current (2015) launches on past (2014) patents, then vice versa. If products are the 'downstream' result of 'upstream' inventive activity, the coefficient of past patents on present launches will be positive significant, and the coefficient of present launches on past IP will be zero, insignificant or both. Table 6 gives the results for naive OLS, and for models with area and industry dummies. The launch-to-patent relationship is close to zero, and orders of magnitude smaller than the patent-to-launch relationship.

Table 6 about here

4/ Digital tech firms and launches: identification

We use modelled launches to explore the innovative performance of ICT / digital firms, specifically the single plant SMEs that are of great interest to policymakers. Our identification strategy uses firm fixed effects on panel data: results are descriptive, not causal.

We define digital technology status using two alternative definitions of sector space: first, using the set of 'digital technologies' SIC codes defined by the UK Office of National Statistics (Harris, 2015), and second, using alternative sector-product definitions based on Gi modelled variables developed via machine learning (Nathan and Rosso, 2015). These

[16] In our data, manufacturing firms patent more than services firms across the whole sample, and for the sub-sample with events exposure. Of those firms who patent, patenting is higher for services than for manufacturing.

definitions give a UK digital tech sector of 12% and 20% of all firms, respectively, with the big data-powered measures picking up a substantive amount of activity outside 'core' ICT sectors and missed by conventional industry codes.

Our basic identification strategy uses the knowledge production function (2), which links launch activity to past patenting and trademarking (and thus to R&D), other firm characteristics, firm fixed effects and area, time and industry dummies:

$$L_{isat} = a + \mathbf{bIP}_{ist-n} + \mathbf{cX}_{ist} + I_i + S_s + A_a + T_t + e_{isat} \quad (2)$$

L indicates either a launch dummy or the count of launches, as in equation (1). Patenting and trademarking (**IP**) are defined as single-year activity lagged *t-n* years. Firm controls **X** now include age, startup status, log turnover, corporate group structure and recent high-growth activity, all lagged one period. Area, time and industry dummies are as in (1).

As noted in Section 2, just 2.88% of firms switch SIC codes during the panel period, and available GI classifications are non-time-varying. We thus follow Combes et al (2008): we save firm fixed effects estimates from (2) and regress these on digital tech status and other cross-sectional characteristics, using bootstrapped standard errors, 400 reps. This allows us to recover our coefficient of interest, β:

$$I\mathit{hat}_i = \alpha + \beta \mathrm{DIGITECH}_i + z_i \quad (3)$$

Estimated fixed effects I*hat* represent time-invariant company characteristics net of observables and wider conditions specified in (2). Coefficients of DIGITECH in (3) give us

the underlying propensity of the average digital tech firm to have product / service launches, when L is a dummy, and when L is a count, the level of activity that can be ascribed to digital tech status. We estimate in OLS as before, with robustness checks for functional form.

4.1 / Identification challenges

Our basic approach faces a number of challenges. The main issue is the fact that launches are reported, not directly observed. As discussed previously, reported launches can be considered as a lower bound on true launches. However, we know that there is selection on firms with media-reported events of any kind, and these firms differ on observables from firms without events exposure.

Events exposure is likely to be driven by a) actual launch activity, and upstream inventive activity before it, plus b) firm capacity/decision to report and c) media awareness and behaviour. For example, reported launches may partly reflect what firms want to and/or are able to report. Young firms may need to promote themselves more than larger, more established firms; the latter may have greater capacity to report. Further, failing and low-growth firms are more likely to hide bad news (Enikopolov and Petrova, 2015). Second, and relatedly, both reported *and actual* launches reflect both innovative activity and strategic decisions with the firm. Specifically, launch activity and timing at least partly reflects individual managers' approach and capacity – plus wider market factors, changes in country-level policy regimes and wider trade frictions, which can be partially captured in industry*time interactions (Cockburn et al., 2016). Third, reported launches also reflect what media actors choose to and/or are able to report. Overall, media attention is likely more focused on established players, so that company age, size and legal status may be predictors

of event exposure for this reason too. However, in industries with low entry barriers (such as ICT), media reporting covers both dominant businesses and 'disruptive' new entrants (Lafrance, 2016). Industry fixed effects will help handle these sector differences.

For simplicity, assume that event exposure EVENTS takes the value 0 or 1 for a given firm *i*. EVENTS is then some function of firm-level characteristics / behaviour; media activity in that firm's sector; plus wider area, industry and time conditions:

$$EVENTS_{itas} = F(c\mathbf{EC}_{it-n,}\, d\mathbf{MEDIA}_{ist}, I'_i , Y'_t , S'_s , A'_a , u'_{itas}) \quad (4)$$

Then the reported launches we actually observe, RL, are some function of true launch activity, firm characteristics and behaviour, wider conditions, plus media exposure on a firm. In linear form, this can be written out as:

$$RL_{itas} = a + \mathbf{bIP}_{ist-n} + \mathbf{cX}_{ist-n} + I_i + Y_t + S_s + A_a + gEVENTS_{itas} + u_{itas} \quad (5)$$

In which case our second stage equation becomes:

$$I'hat_i = \alpha' + \beta' DIGITECH_i + z_i \quad (6)$$

The pointers above suggest that many of the drivers of event exposure in (4) can be captured in firms' observable characteristics and suitable fixed effect specifications. However, given the novelty of our data we have few *specific* empirical priors on how we should populate the **EC** and **MEDIA** vectors, or whether all salient factors are observable. A further issue is that reported launches are only observed when there is media exposure, so as currently set up there is a partial mechanical correlation between left and right hand sides of the model. That

is, in (5) observing launches depends on observing at least some event activity (whether or not this is a launch). Both issues bias up coefficients of DIGITECH in (6) if not dealt with.

To tackle these challenges, we proceed as follows. We first estimate (5) and (6) for firms with at least one reported event of any kind. Variation then comes from firms with events exposure including launches, versus firms with events exposure of other kinds: conditional on observable characteristics, DIGITECH is independent from the propensity to have events. (By removing EVENTS from the estimating equation, we also work around the simultaneity issue.) However, it is possible that unobserved firm-specific, time-varying determinants of event exposure drive our results. We therefore interpret results as associations holding *only for firms with at least some event exposure*. Next, we expand our analysis to the full sample. Here, we use instruments and sample reweighting approaches to directly control for – and quantify – firms' underlying propensity to have events coverage, and check whether this kills our main result.

4.2 / Main results: firms with events

Table 7 summarises the main results for regressions on the events sub-sample. The first and second panels cover tech firms as defined by ONS digital tech SICs (ONS 2015) crosswalked to SIC2003 codes. Our preferred specification is columns 5 and 6, which correspond to equations (5) and (6) respectively. We find that firms in digital tech SICs are 15.3 percentage points more likely to have at least one product launch (first panel); such firms also have 0.58 more launches than non-tech firms (second panel).

Table 7 about here

The average firm with media exposure has a 37 percentage point probability of a product launch, and conducts 1.5 launches (Table 1), so these are substantive innovation effects for this group of companies. For the probability model, the average probability increases by 45% (from 36% in Table 1 to 52%) for the counts model, the shift is from the 25th to the 90th percentile.

Panels C and D repeat the analysis for an alternative definition of tech sector space, built using GI sector and product codes (Nathan and Rosso, 2015). Here the magnitude of the results differs substantially when a broader definition of industry space is considered, based on industry codes and website content. Specifically, on this definition tech firms are 8.5 percentage points more likely than non-tech firms to have launches (Panel C), but produce 1.2 more launch events than non-tech firms (bottom panel). GI sector space defines a bigger set of activities than the ONS definition, including digitised activity outside core ICT. This broader set of 'tech' firms are less likely to have launches than the core set, but the launch-active produce substantially more launches than core ICT firms.

4.3 / Robustness checks: firms with events

We run a series of robustness checks, with results shown in Appendix tables A7 and A8. Both dummy and count models are robust to changes in the controls vector. Removing firms that change SIC code (and thus potentially into/out of digital tech status) also leaves main results unchanged, as does removing firms who change location. Results are also robust to further tightening of the identification strategy, including fitting industry*year effects; two-way clustering on area*industry; fitting technology field*year fixed effects; and dropping

singletons (Correira, 2015). We also run functional form tests on a reduced form specification, comparing LPM to logit and OLS to Poisson estimators. Marginal effects are very similar in linear and non-linear cases.

4.4 / Heterogeneity analysis

We extend our main results by exploring the role of urban location; and of firms at different stages of scaling. Specifically, in (7) we interact DIGITECH with location dummies for 'urban' TTWAs (that contain a city of at least 125,000 people), or with dummies for small firms (10-50 employees) and medium-size firms (51-250 employees). Micro firms, with under 10 employees, are the reference category in the employment regressions. Results are given in Appendix Tables A9 and A10. For probability models, we find no significant role of urban location, but positive links for medium size digital tech firms (significant for SIC definitions, and marginally so for GI). For count models, results vary by technology space. Medium size GI digital tech firms and small SIC digital tech firms produce more launches. We find no evidence that SIC digital tech firms in urban areas have more launches, but a strong positive link for urban GI tech firms. This implies that urban areas may better support launch activity for a broader set of digitised activities than for core ICT activities.

4.5 / Full sample analysis

In this section, we use the full sample of firms and use a range of methods to directly control for firms' underlying propensity to have events exposure. Since the majority of firms have no events exposure (and thus no launches), this will mechanically shift our estimates of

DIGITECH downwards. Further controlling for event exposure propensity should reduce coefficients yet further. If our main results survive these checks, then, this is a useful finding.

Tables 8 and 9 about here

In both cases, we start with a re-estimation of our baseline model (5) for the full sample, without controlling for event exposure. Results are given in columns 1 and 2 of Tables 8 and 9, for the probability and counts models respectively. As expected, DIGITECH coefficients are much smaller, with zero-events firms driving the result. In columns 3 and 4, we check how much of the firm fixed effect can be explained by (time-invariant) event exposure, on top of that explained by digital tech status. To do this we re-run (6) including an EVENTS dummy. We can see that events exposure explains just under a third of the fixed effect variation; DIGITECH coefficients drop somewhat but stay significant. Even if coefficients are much smaller, the effect, compared to the baseline (0.003 in Table 1) is large: digital companies have twice the probability to launch a product compared to non-digital companies.

Observing launch activity is conditional on event exposure. To tackle this, we instrument for EVENTS in (6) using a two-period lagged dummy for PLC legal status (a UK public company, compared to a partnership, LLC or sole trader). The intuition is that PLC status is a proxy for a firm's public profile and its capacity to attract media attention, while not directly influencing launch activity.[17] IV results are given in columns 5-7. The instrument is a strong predictor of events exposure, and passes the usual tests for weak identification and under-identification. Second stage IV coefficients of DIGITECH are very similar to the baseline estimates.

[17] Reduced form regressions of launch activity on plc status (unlagged, and lagged 1 and 2 periods back) show negative significant coefficients of plc status. We also run regressions with an unlagged instrument, with identical results to those reported here.

As an alternative, we reweight the sample of firms with events exposure, so that they more closely resemble the rest of the sample on observable characteristics. This is in the spirit of case-control strategies for patent analysis (Jaffe et al 1993, Thompson and Fox-Kean 2005), in which individual patenting firms are matched with control firms identical on observables. Given the large difference in observables between firms with and without events exposure – see Section 2 – this firm-level approach is not feasible. We follow a nonparametric reweighting procedure where the probability of launching a product for the companies with events is reweighted so that the frequencies of the firms' characteristics resemble those of the firms with no events[18]. Further, given the novelty of our data we are unclear on the precise set of relevant events exposure predictors. We therefore develop weights using Lasso regression to test choices in a way that penalises overfitting (given by minimising Mallow's Cp statistic). Specifically, we generate a longlist of predictors and test these against 'kitchen sink' specifications that use all 31 variables in our summary statistics.[19] Tables A11 and A12 give the results for longlist and kitchen sink respectively; lasso removes one of the longlist predictors (age) to give a Cp of 11; by contrast the kitchen sink specification gives a Cp score of 20 using 19 predictors, with the rest rejected. To make the weights, we estimate a probit regression of EVENTS on the cleaned vector of predictors, plus time, area and SIC2 industry dummies. For firms with events, weights *W* are given by a Bayes' rule:

$$W = (p/(1-p)) * ((1-pbar) / pbar) \quad (7)$$

[18] The re-weighted probability of an event is the one we would have observed if the information on events was available for the whole sample of firms. See DiNardo, Fortin, and Lemieux (1996).

[19] Results given for a kitchen sink predictor set cleaned for collinearity. We also estimate with a raw predictor set, which gives a Cp of 22.7 and drops a number of variables in the lasso estimation.

Where *p* is the probability that EVENT = 1 and *pbar* is the conditional probability of having an event (pr(EVENT=1|X)). Table A14 compares the means and standard deviations of firms without events (top panel) to unweighted and reweighted firms with events exposure (middle and bottom panels respectively). We can see that on a range of event predictors, reweighting successfully brings firms with events rather closer to those without (Table A13 in Appendix).

Table 10 about here

Results from reweighted sample regressions are given in Table 10, with *W* used as probability weights. For both the linear probability model (top panel) and counts model (bottom panel), reweighted DIGITECH estimates are slightly smaller than the baseline, as expected.

6/ Conclusions

This paper introduces a new and experimental measure of firm-level innovation: modelled product/service launches developed using machine learning, which we argue complements existing innovation metrics and provides a potentially important tool for decision-makers. We provide evidence that past patenting and trademarking predict launch activity at the firm level. We use panel fixed effects settings to look at launch performance in UK single plant SMEs; for companies with events coverage, digital tech SMEs are substantially more launch-active than non-tech firms. We use instruments and reweighting on the full sample of companies to handle firms' underlying propensity for events exposure: effect size is dampened, but results remain significant.

References

Aghion P. (2016) Entrepreneurship and growth: lessons from an intellectual journey. *Small Business Economics*: 1-16.

Angrist J and Pischke J-S. (2009) *Mostly Harmless Econometrics,* Princeton: Princeton University Press.

Arora A, Belenzon S and Sheer L. (2017) Back to Basics: Why do Firms Invest in Research? *NBER Working Paper 23187*. Cambridge, MA: NBER.

Autor D, Dorn D, Hanson GH, et al. (2016) Foreign Competition and Domestic Innovation: Evidence from U.S. Patents. *NBER Working Paper 22879*. Cambridge, MA: NBER.

Block JH, Fisch CO, Hahn A, et al. (2015) Why do SMEs file trademarks? Insights from firms in innovative industries. *Research Policy* 44: 1915-1930.

Blundell R, Griffith R and Van Reenen J. (1995) Dynamic Count Data Models of Technological Innovation. *The Economic Journal* 105: 333-344.

Castellacci F. (2008) Technological paradigms, regimes and trajectories: Manufacturing and service industries in a new taxonomy of sectoral patterns of innovation. *Research Policy* 37: 978-994.

Chatterji A, Glaeser E and Kerr W. (2014) Clusters of Entrepreneurship and Innovation. *Innovation Policy and the Economy* 14: 129-166.

Chesborough H. (2003) *Open Innovation: The New Imperative for Creating and Profiting from Technology*, Cambridge, MA: Harvard Business School Press.

Cockburn IM, Lanjouw JO and Schankerman M. (2016) Patents and the Global Diffusion of New Drugs. *The American Economic Review* 106: 136-164.

Cohen WM and Levinthal DA. (1990) Absorptive Capacity: A New Perspective on Learning and Innovation. *Administrative Science Quarterly* 35: 128-152.

Combes P-P, Duranton G and Gobillon L. (2008) Spatial wage disparities: Sorting matters! *Journal of Urban Economics* 63: 723-742.

Correira S. (2015) Singletons, Cluster-Robust Standard Errors and Fixed Effects. Mimeo.

Coutu S. (2014) The Scale-Up Report on UK Economic Growth. London: Scale-Up Institute.

Cowen T. (2011) *The Great Stagnation*, New York: Dutton Adult.

Davies N. (2009) *Flat Earth News*, London: Vintage.

Decker R, Haltiwanger J, Jarmin RS, et al. (2016) Where Has All the Skewness Gone? The Decline in High-Growth (Young) Firms in the U.S. *European Economic Review* 86: 4-23.

DiNardo, John, Nicole M. Fortin, and Thomas Lemieux. 1996. "Labor Market Institutions and the Distribution of Wages, 1973-1992: A Semiparametric Approach." *Econometrica* 64 (5):1001–1044.

Einav, L., Levin, J., 2014. The Data Revolution and Economic Analysis. *Innovation Policy and the Economy* 14, 1–24.

Enikopolov R and Petrova M. (2015) Media Capture: Empirical Evidence. In: Anderson S, Waldfogel J and Strömberg D (eds) *Handbook of Media Economics.* North Holland: Elsevier, 687-700.

Fosfuri A, Giarratana MS and Luzzi A. (2008) The Penguin Has Entered the Building: The Commercialization of Open Source Software Products. *Organization Science* 19: 292-305.

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

Gentzkow M, Kelly BT and Taddy M. (2017) Text as Data. *NBER Working Paper 23276.* Cambridge, MA: NBER.

Gibbons S, Overman HG and Resende G. (2011) Real Earnings Disparities in Britain. SERC Discussion Paper DP0065. London: SERC.

Gordon RJ. (2016) *The Rise and Fall of American Growth*, Princeton: Princeton University Press.
Griliches Z. (1979) Issues in assessing the contribution of R&D to productivity growth. *Bell Journal of Economics* 10.
Guzman J and Stern S. (2015) Where is Silicon Valley? *Science* 347: 606-609.
Hall BH and Harhoff D. (2012) Recent Research on the Economics of Patents. *NBER Working Paper No. 17773*. Cambridge, MA: NBER
Hall BH, Helmers C, Rogers M, et al. (2013) The Importance (or not) of Patents to UK Firms. *National Bureau of Economic Research Working Paper 19089*. Cambridge, MA: NBER.
Hall BH, Helmers C, Rogers M, et al. (2014) The Choice Between Formal and Informal Intellectual Property: A review. *Journal of Economic Literature* 52: 375-423.
Haltiwanger J, Jarmin RS and Miranda J. (2013) Who Creates Jobs? Small versus Large versus Young. The Review of Economics and Statistics 95: 347-361.
Harris J. (2015) Identifying Science and Technology Businesses in Official Statistics. London: ONS.
Helmers C and Rogers M. (2010) Innovation and the Survival of New Firms in the UK. *Review of Industrial Organization* 36: 227-248.
Jaffe AB, Trajtenberg M and Henderson R. (1993) Geographic Localization of Knowledge Spillovers as Evidenced by Patent Citations. *The Quarterly Journal of Economics* 108: 577-598.
Jiang R, Banchs RE & Li H. (2016). Evaluating and Combining Named Entity Recognition Systems. ACL 2016, 21.
Katila R and Ahuja G. (2002) Something Old, Something New: A Longitudinal Study of Search Behavior and New Product Introduction. *The Academy of Management Journal* 45: 1183-1194.
Lafrance A. (2016) 'Access, Accountability Reporting and Silicon Valley. *Nieman Reports*, 17 August.
Li W and Hall BH. (2016) Depreciation of Business R&D Capital. *NBER Working Paper No. 22473*. Cambridge, MA: NBER.
Mairesse J and Mohnen P. (2010) Using Innovation Surveys for Econometric Analysis. In: Bronwyn HH and Nathan R (eds) *Handbook of the Economics of Innovation*. North-Holland, 1129-1155.
Mazzucato M. (2011) *The Entrepreneurial State*. London: Demos.
Mokyr J. (2013) 'Is technological progress a thing of the past?' Vox, 8 September. http://voxeu.org/article/technological-progress-thing-past, accessed 2 March 2017.
Motoyama Y and Bell-Masterson J. (2014) Beyond Metropolitan Start-Up Rates. Kansas City: Kauffman Foundation.
Nathan M and Rosso A. (2015) Mapping digital businesses with Big Data: some early findings from the UK *Research Policy* 44: 1714-1733.
OECD. (2013) Measuring the Internet Economy: A contribution to the research agenda. *OECD Digital Economy Papers 226*. OECD Publishing.
Office of National Statistics. (2016) Business Structure Database, 1997-2015 SN: 6697. *Secure Data Service Access [computer file]*. Colchester: UK Data Archive.
Roberts ME, Stewart BM and Tingley D. (2016) stm: R Package for Structural Topic Models. http://www.structuraltopicmodel.com.
Teece DJ, Pisano G and Shuen A. (1997) Dynamic Capabilities and Strategic Management. *Strategic Management Journal* 18: 509-533.
Thompson P and Fox-Kean M. (2005) Patent Citations and the Geography of Knowledge Spillovers: A Reassessment. *American Economic Review* 95: 450-460.

Varian HR. (2014) Big Data: New Tricks for Econometrics. *Journal of Economic Perspectives* 28: 3-28.
Viner K. (2016) 'How Technology Disrupted the Truth'. *Guardian.* 12 July.
Von Hippel E. (2005) *Democratising Innovation*, Cambridge, MA: MIT Press.

Figures and tables.

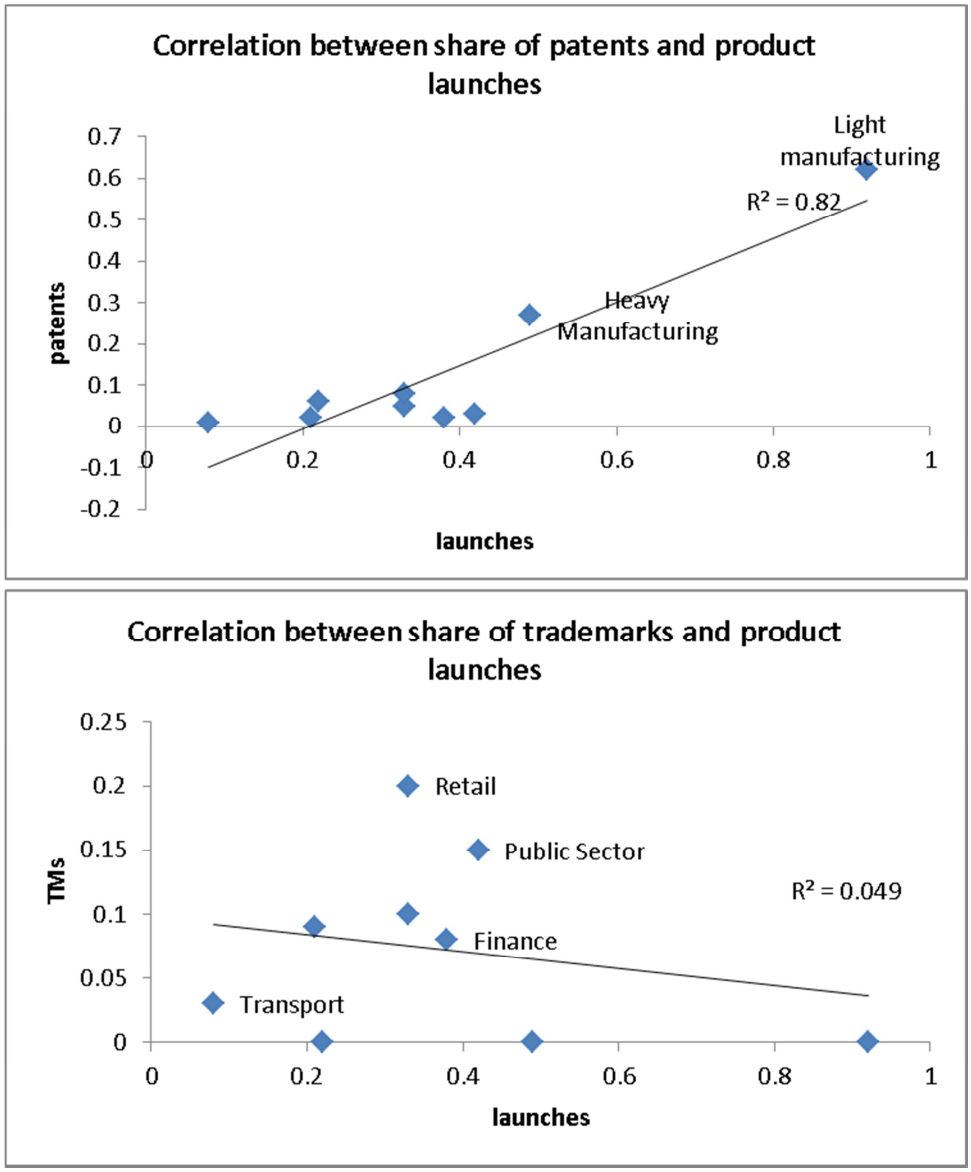
Figure 1. Example 'events', showing raw text and classification.

company_id	13724
event_type_id	product_launch
event_date	2014-03
Sample fragment	**Masterwork** goes large with new die cutter. Postpress equipment manufacturer Masterwork Graphic Equipment has expanded its range of products with the addition of the MK1450ER large-format die cutter with stripping and blanking facilities….
source_name	xxx
doc_title	Masterwork goes large with new die cutter
url	http://www.XXX/NewsStory.aspx?i=2296

company_id	1542955
event_type_id	product_launch
event_date	2013-12
Sample fragment	**Hammond Electronics** has launched a range of designspecific moulded enclosures to support the new types of credit card sized, low cost bare board computers, which, typically running Linux, provide basic functionality across a wide range of applications…
source_name	xxx
doc_title	Enclosures for credit-card sized computers
url	http://www.XXXX/content/enclosures-credit-card-sized-computers

Source: GI.
Note: As agreed with data provider we cannot report the source name. 'fragment' gives the raw text, with company subject in **bold**. 'doc_title', 'url' and 'event_date' provide further raw info. 'event_type_id' is the Gi classification.

Figure 2. Correlation between share of companies patenting/trademarking and launching a product. 1-digit industry level.



Source: GI / CH / Orbis / UKIPO.

Table 1. Event coverage and type, 2014-2015.

Variable	Obs	Mean	Std. Dev.
Firm has event	2,406,346	0.009	0.093
Event count	2,406,346	0.034	1.602
Firm has product launch	2,406,346	0.003	0.057
Product launch count	2,406,346	0.013	1.204
Firms with events in either 2014, 2015 or both			
Variable	**Obs**	**Mean**	**Std. Dev.**
Firm has event	32,852	0.642	0.479
Event count	32,852	2.475	13.492
Firm has product launch	21,098	0.368	0.482
Product launch count	21,098	1.494	12.770

Year	Product Launches (%)
2014	11,137 (52.8)
2015	9,961 (47.2)
Total	*21,098 100*

Source: GI. Minima and maxima suppressed to avoid disclosure.

Table 2. Comparing firms with events vs. those without, 2014-2015.

Variable	Mean for firm with	
	no events	*events*
Age entered BSD	11.389	15.973
Age entered BSD / incorporated	12.432	18.221
Year incorporated	2004	1998
Startup	0.144	0.04
Immediate foreign ownership	0.166	0
Firm is in a group of enterprises	0	0.058
Number of companies in the group	0.008	0.208
Firm is a public limited company (plc)	0.982	0.923
Firm is a sole proprietor	0.001	0
Patent stocks	0.008	0.195
Weighted patent stocks	0.008	0.194
EPO/US/PCT patent stocks	0.004	0.107
Weighted EPO/US/PCT patents stocks	0.004	0.107
TM stocks	0.004	0.035
Jobs two-year average	5.422	22.935
Jobs growth two-year ave (%)	1.6	2.9
High jobs growth firm	0.015	0.062
High jobs growth firm <= 5yo (gazelle)	0	0.009
Revenue two year average (,000)	867	14668
Revenue growth two-year ave (%)	1	4.6
High revenue growth firm	0.15	0.213
High rev growth firm <= 5yrs old (gazelle)	0.056	0.046
Revenue/worker two year average (,000)	147.993	736.509
Productivity growth two-year ave (%)	-0.7	1.7
High rev/worker growth firm	0.129	0.148
High rev/worker growth firm <= 5yrs old	0.038	0.023

Source: BSD / CH / GI.
Sample as in Table 1. All differences are significant at 1% using rank sum tests and t-tests.

Table 3. Coverage of firms by SIC1 sectors for product launch, patents and trademarks.

sic03	sic03 section name	no launch	launch	no patent	patent	no tm	tm
A	Agriculture, hunting and forestry	99.9	0.1	99.99	.	99.94	0.06
B	Fishing	99.89	.	100	.	100	.
C	Mining and quarrying	98.77	1.23	99.51	.	99.88	.
D	Manufacturing	99.41	0.59	99.65	0.35	99.74	0.26
E	Electricity, Gas and Water Supply	99.84	.	99.94	.	99.94	.
F	Construction	99.92	0.08	99.99	0.01	99.97	0.03
G	Wholesale and retail trade, etc	99.6	0.4	99.93	0.07	99.76	0.24
H	Hotels and restaurants	99.91	0.09	100	.	99.95	0.05
I	Transport, storage and communications	99.67	0.33	99.97	0.03	99.92	0.08
J	Financial intermediation	99.53	0.47	100	.	99.9	0.1
K	Real estate, renting and business activities	99.67	0.33	99.92	0.08	99.9	0.1
L	Public administration and defence, etc	100	.	100	.	100	.
M	Education	99.73	0.27	99.99	.	99.87	0.13
N	Health and social work	99.82	0.18	99.97	0.03	99.93	0.07
O	Other community, social, personal services	99.58	0.42	99.97	0.03	99.85	0.15
P	Household domestic employment	100	.	100	.	100	.
Q	Extra-terrestrial organisations, bodies	100	.	100	.	100	.
	All	*99.68*	*0.32*	*99.92*	*0.08*	*99.88*	*0.12*

Source: Gi / CH / BSD / Orbis / UKIPO. Sample = 2.4m observations. Cells with under 10 observations suppressed to avoid disclosure.

Table 4. Linking past IP activity to product launches. Stepwise regressions.

A. pr(Launch)	**(1)**	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**
L1.PCT / EPO / US patent stock	0.008***	0.008***	0.006***	0.006***	0.006***	0.006***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
L1.TM stocks	0.010***	0.008***	0.008***	0.008***	0.008***	0.007***
	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Ave pre-2009 patenting			-0.005*	-0.005*	-0.004*	-0.004*
			(0.002)	(0.002)	(0.002)	(0.002)
Firm patents pre-2009			0.039***	0.039***	0.038***	0.033***
			(0.007)	(0.007)	(0.007)	(0.006)
Observations	2406346	2246386	2246386	2246386	2225176	2225176
R^2	0.0014	0.0069	0.0074	0.0074	0.0081	0.0126
B. Launch counts	**(1)**	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**
L1.PCT / EPO / US patent stocks	0.029***	0.024***	0.022***	0.022***	0.022***	0.021***
	(0.004)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)
L1.TM stocks	0.029***	0.019**	0.018**	0.019**	0.019**	0.016**
	(0.008)	(0.008)	(0.008)	(0.007)	(0.007)	(0.007)
Ave pre-2009 patenting			-0.031**	-0.031**	-0.030**	-0.030**
			(0.012)	(0.012)	(0.012)	(0.012)
Firm patents pre-2009			0.119**	0.119**	0.118**	0.107**
			(0.049)	(0.049)	(0.049)	(0.048)
Observations	2406346	2246386	2246386	2246386	2225176	2225176
R^2	0.0000	0.0004	0.0004	0.0004	0.0005	0.0009

Source: BSD / CH /Orbis / IPO / GI. All models fit controls, area (TTWA), time (year) and 4-digit SIC industry dummies. Controls fitted include log 2-year mean turnover, age since BSD entry OR incorporation, number of companies per entref, enterprise has >1 associated company dummy, legal status dummies (public company, sole proprietor, reference = other). Controls lagged one year except age. Standard errors clustered on 2-digit SIC. *** significant at 1%, ** significant at 5%, * significant at 10%. Constant not shown.

Table 5. Linking past IP activity to product launches. Variable lags.

A. Pr(product launch)	**(1)** **L0**	**(2)** **L1**	**(3)** **L2**	**(4)** **L3**	**(5)** **L4**	**(6)** **L5**
PCT / EPO / US patent stock	0.006*** (0.001)	0.006*** (0.001)	0.007*** (0.001)	0.008*** (0.002)	0.009*** (0.002)	0.007* (0.004)
TM stock	0.010*** (0.002)	0.007*** (0.001)	0.010** (0.005)	0.010** (0.005)	0.010* (0.005)	0.010** (0.005)
Ave pre-2009 patenting	-0.004* (0.002)	-0.004* (0.002)	-0.005** (0.002)	-0.005*** (0.002)	-0.004** (0.002)	-0.001 (0.002)
Firm patents pre-2009	0.033*** (0.006)	0.033*** (0.006)	0.035*** (0.006)	0.035*** (0.006)	0.036*** (0.006)	0.036*** (0.007)
Observations R^2	2225176 0.0129	2225176 0.0126	2225176 0.0125	1938369 0.0130	1691279 0.0138	1517738 0.0143
B. Launch counts	**(1)** **L0**	**(2)** **L1**	**(3)** **L2**	**(4)** **L3**	**(5)** **L4**	**(6)** **L5**
PCT / EPO / US patent stocks	0.022*** (0.006)	0.021*** (0.005)	0.027*** (0.006)	0.030*** (0.008)	0.041*** (0.015)	0.046* (0.023)
TM stock	0.023** (0.011)	0.016** (0.007)	0.029 (0.028)	0.030 (0.029)	0.030 (0.030)	0.032 (0.032)
Ave pre-2009 patenting	-0.029** (0.011)	-0.030** (0.012)	-0.034*** (0.011)	-0.033*** (0.009)	-0.035*** (0.009)	-0.029*** (0.009)
Firm patents pre-2009	0.104** (0.048)	0.107** (0.048)	0.111** (0.049)	0.113** (0.049)	0.113** (0.049)	0.113** (0.050)
Observations R^2	2225176 0.0009	2225176 0.0009	2225176 0.0009	1938369 0.0010	1691279 0.0010	1517738 0.0010

Source: BSD / CH /Orbis / IPO / GI. All models fit controls, area (TTWA), time (year) and 4-digit SIC industry dummies. Patent stocks are lagged back up to 5 periods, TM stocks up to 2 periods. Controls fitted as per Table 4. Standard errors clustered on 2-digit SIC. *** significant at 1%, ** significant at 5%, * significant at 10%.

Table 6. Innovation regressions: placebo test.

	OLS	FE	OLS	FE
	2015 launch counts		2014 patent stocks	
2014 patent stocks	0.0204***	0.0194***		
	(0.003)	(0.003)		
2015 product launch counts			0.0009***	0.0008***
			(0.000)	(0.000)
Constant	0.0120***		0.0047***	
	(0.002)		(0.001)	
Observations	2370158	2347297	2370158	2347297
R^2	0.0000	0.0006	0.0000	0.0110

Source: BSD/CH/GI/Orbis.

OLS models fit bivariate model. FE models add Travel to Work Area and industry dummies.

Table 7. Linking tech firm status to launch activity, for firms with event exposure.

A. pr(launch)	**(1)**	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**
Digital tech firm SIC	0.163***		0.166***	0.185***		0.153***
	(0.025)		(0.021)	(0.013)		(0.007)
Observations	29616	29616	29616	29616	29616	29616
R^2	0.0164	0.0026	0.0194	0.0857	0.6463	0.0213
B. Launch counts	**(1)**	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**
Digital tech firm SIC	0.532***		0.578***	1.056***		0.576***
	(0.147)		(0.163)	(0.265)		(0.120)
Observations	29616	29616	29616	29616	29616	29616
R^2	0.0003	0.0018	0.0021	0.0205	0.9813	0.0003
C. pr(launch)	**(1)**	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**
Digital tech firm GI	0.098***		0.101***	0.045***		0.085***
	(0.021)		(0.021)	(0.008)		(0.005)
Observations	29616	29616	29616	29616	29616	29616
R^2	0.0103	0.0026	0.0134	0.0872	0.6463	0.0116
D. Launch counts	**(1)**	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**
Digital tech firm GI	1.167***		1.235***	1.084***		1.228***
	(0.259)		(0.246)	(0.296)		(0.215)
Observations	29616	29616	29616	29616	29616	29616
R^2	0.0024	0.0018	0.0044	0.0218	0.9813	0.0027
Controls	N	Y	Y	Y	Y	
Area/time/sector dummies	N	N	N	Y	Y	
Firm fixed effects	N	N	N	N	Y	

Source: BSD/CH/Gi/Orbis/IPO.
Notes: Standard errors in parentheses, clustered on 2-digit SIC/NACE, except for column (6) which fits bootstrapped standard errors, 400 reps. Controls fitted include firm age, startup dummy, 1-year lags of log turnover, firm stock of EPO/PCT/US patents, firm stock of trademarks, multiple companies dummy, number of linked companies, legal status dummies. Time unit is year. Area is Travel to Work Area, which approximates local labour market. Sector is 4-digit SIC/NACE. *** significant at 1%, ** 5%, * 10%.

Table 8. Tech firms and product launch activity. Full sample analysis with IV.

Depvar = pr(launch)	**(1)**	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**	**(7)**
					pr(event)	**iv stage 2**	
Digital tech firm SIC	0.0043***		0.0031***			0.0032***	
	(0.000)		(0.000)			(0.000)	
Digital tech firm GI		0.0026***		0.0017***			0.0018***
		(0.000)		(0.000)			(0.000)
firm has event			0.2952***	0.2952***		0.2626***	0.2618***
			(0.003)	(0.003)		(0.003)	(0.013)
L2. Firm is PLC					*-0.0244****		
					(0.001)		
Observations	1938273	1938273	1938273	1938273	*1938273*	1938273	1938273
R^2	0.001	0.000	0.320	0.320	*0.002*	0.316	0.316
F					*617.096*		
Kleibergen-Paap under-identification test chi^2						885.786	888.991
Kleibergen-Paap chi^2 p-value						0.000	0.000
Kleibergen-Paap weak identification test F						901.235	904.612
Hansen over-identification test j						0.000	0.000

Source: BSD / CH /Orbis / IPO / GI. Second stage coefficients only except where stated. Dependent variable in iv first / column (5) is pr(event). All models use controls as in Table 7, plus firm/area/SIC4/year FE. Standard errors are bootstrapped (400 reps) so IV regressions use Kleibergen-Paap test statistics. Weak identification test uses Stock-Yogo thresholds: 10% threshold = 16.38. Under-identification test: null = under-identified. *** significant at 1%, ** significant at 5%, * significant at 10%.

Table 9. Tech firms and product launch activity. Full sample analysis with IV. Counts model.

Depvar = launch counts	**(1)**	**(2)**	**(3)**	**(4)**	**(5)** *pr(event)*	**(6)** iv stage 2	**(7)**
Digital tech firm SIC	0.0156***		0.0098***			0.0140***	
	(0.002)		(0.002)			(0.002)	
Digital tech firm GI		0.0247***		0.0203***			0.0233***
		(0.004)		(0.004)			(0.004)
firm has event			1.4077***	1.4069***		0.3743***	0.4522***
			(0.094)	(0.094)		(0.095)	(0.088)
L2. Firm is PLC					*-0.0244****		
					(0.001)		
Observations	1938273	1938273	1938273	1938273	*1938273*	1938273	1938273
R^2	0.000	0.000	0.011	0.011	*0.002*	0.005	0.006
F					*617.096*		
Kleibergen-Paap under-identification test chi^2						885.786	888.991
Kleibergen-Paap chi^2 p-value						0.000	0.000
Kleibergen-Paap weak identification test F						901.235	904.612
Hansen over-identification test j						0.000	0.000

Source: BSD / CH /Orbis / IPO / GI. Second stage coefficients only except where stated. Dependent variable in iv first / column (5) is pr(event). All models use controls as in Table 7, plus firm/area/SIC4/year FE. Standard errors are bootstrapped (400 reps), so IV regressions use Kleibergen-Paap test statistics. Weak identification test uses Stock-Yogo thresholds: 10% threshold = 16.38. Under-identification test: null = under-identified. *** significant at 1%, ** significant at 5%, * significant at 10%.

Table 10. Tech firms and product launches. Full sample analysis with reweighting.

A. Pr(launch)	**baseline**			**reweighted sample**		
	(1)	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**
Digital tech firm SIC		0.0046***			0.0038***	
		(0.000)			(0.000)	
Digital tech firm GI			0.0026***			0.0011***
			(0.000)			(0.000)
Observations	1763970	1763970	1763970	1763970	1763970	1763970
R^2	0.710	0.001	0.000	0.898	0.000	0.000
B. Launch counts	**baseline**			**reweighted sample**		
	(1)	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**
Digital tech firm SIC		0.0167***			0.0161***	
		(0.002)			(0.002)	
Digital tech firm GI			0.0242***			0.0230***
			(0.004)			(0.004)
Observations	1763970	1763970	1763970	1763970	1763970	1763970
R^2	0.981	0.000	0.000	0.975	0.000	0.000

Source: BSD / CH /Orbis / IPO / GI. Second stage coefficients only except where stated. All models use controls as in Table 7, plus events dummy and firm/area/SIC4/year FE. Standard errors bootstrapped, 400 reps. *** significant at 1%, ** significant at 5%, * significant at 10%.

INNOVATIVE EVENTS: APPENDIX

A1 / Events data

This paper uses variables that model events in a company's lifecycle (hence 'events'), developed by the data science firm Growth Intelligence (Gi). Each 'event' is based on content taken from company websites or from 3,740 online news sources (including major sources such as Reuters or Yahoo news, as well as industry sources such as IT Briefing and PRWeb. Our raw data consists of 318,899 observations covering financial years 2014 and 2015 (in calendar time, August 2013 to November 2014 inclusive). The fundamental challenge in using the events dataset for inference is dealing with its unstructured nature. We develop a number of substantive checks and improvements on the raw data

Feature extraction and company name recognition

We first clean the data to remove all-fields duplicates and the small number of events projected for dates in the future. We remove 'farmed' content by not allowing identical text fragments to appear more than once a day anywhere in the data. Third, we conduct checks for the quality of GI's feature extraction and syntax parsing.

We begin with a simple manual check for 'negative events' – that is, reports describing something that has *not* occurred. On a 1% sample of product/service launch events, we find a negative event error rate of 0.6% (5/823). Next, we conduct more systematic checks on a sample of 'hard cases'. We define 'hard cases' as observations where there are *a priori* reasons to believe GI's ascription of news article text to a given company may be incorrect: specifically, because the text includes either a large tech company (e.g. Google, Facebook) or a large press agency (e.g. Reuters, Bloomberg). These company names often appear in

everyday contexts outside activities by that company. For example, 'Google' is now commonly used as a noun or verb; many company websites and online news articles will include social media-related text along the lines of 'follow us on Facebook'; many news reports about other companies are filed by large press agencies. In this way, the set of hard cases provides a natural upper bound on the error rate in GI's analysis. By restricting the sample to single plant SMEs, we further guard against error by removing these hard cases from the data.

Specifically, ascription error could arise from failure to extract text from credible online sources ('content farming error'), or, once text has been extracted, from failures of name entity recognition or selection ('algorithm error'). We define large tech and media companies through Wikipedia reports of global market cap or market share. We draw 5,000 event observations (news articles) ascribed to one of these companies (hence 'big digital', 12.5% of the 40,000 observations ascribed to such firms).

Analysis using title and text fragment fields suggests around 16% content farming error in the 'big digital' sample, especially what we term 'copyright clutter' (where ascription has been done on article source/copyright text) and what we term 'social media clutter' (ascription based on 'follow us on facebook' type text). Note that GI's ascription is based on the full text from each event text, not just the fields provided to us, so true error rates due to clutter may be lower than this.

We also conduct further, experimental tests on a sub-sample of the 'big digital' companies. We use the URL field to re-extract the original text, then to reverse-engineer GI's feature extraction and syntax parsing routines. We are only able to perform this exercise on websites

that are a) scrapable b) active (return a 200 to standard HTTP requests). This reduces our sample size to 1,746. We then build a web crawler to retrieve the original webpage text, and train a Name Entity Recognition (NER) model to identify company names from the re-extracted data. The model is built from Stanford NER Conditional Random Field Classifiers, which is the current gold standard (with over 80% accuracy) (Jiang et al 2016). We use the CoNLL, MU6, MU7 abd ACE 2002 training datasets, which are substantively based on news corpora. For each observation, we proceed as follows. We extract all company names C_{ner} (we already know the GI company name C_{gi}). Let $C_{candidates}$ be a subset of C_{ner} occurring in the title / headline of each article, and identified as potential subjects in the text. We assume that the correct subject(s) of the event described will be a) identified as subject at least once in the text (and probably multiple times), and b) be mentioned in the article title. For precision, we therefore drop 413 cases where no company is mentioned in the article title.

This leaves us with three scenarios. If GI's ascription is correct, C_{gi} is in $C_{candidates}$, and $C_{candidates} = 1$. If GI's ascription is probably correct, C_{gi} is in $C_{candidates}$, and $C_{candidates} \geq 1$. If GI's ascription is incorrect, C_{gi} is not in $C_{candidates.}$ For the 'digital' sample, we find 95.1% incorrect ascription on the 977 remaining observations. Note that this is a lower bound on the true error rate: if we assume all 'probably correct' cases are incorrect, the error rate rises to 99% of cases.[1] Note that our focus on single plant SMEs removes hard cases from the data, minimising the ascription error rate on the rest of the sample. However, to the extent that mis-ascription 'gives' events to large tech and media firms which actually belong to SMEs, we have a lower bound on the true level of event activity for our firms of interest.

[1] We check for out-of-sample error rates by running these routines for the full set of 40,000 observations, with very similar results.

Event observations and actual events

A further substantive issue is that in its raw form, an event observation may not perfectly correspond to some underlying (real world) event. For example, a major merger is likely to be reported hundreds of times; each of these is currently reported as a distinct event occurrence. We use structural topic modelling (STM) to deal with this. Topic modelling algorithms handle this issue by clustering text fragments that talk about the same topic in different ways, using different text but similar content words (Roberts et al., 2016). In STM, each text fragment is modelled as a document. A topic is defined as a mixture over words where each word is associated to a probability of belonging to a topic. A document is a mixture over topics; therefore each document can be associated to multiple topics. For each text fragment we have a *topical prevalence* and a *topical content*. The prevalence refers to how much a document is associated with a topic, and it is computed using the shared words in the document, while the content refers to the words used within the topic. We use the topical prevalence to group event fragments within the same topic. We use the 90% threshold, so we assume that events belong to the same cluster if they share at least 90% of the content.[2]

Before modelling the data, we stem the fragments (reducing the words to their roots) and remove stopwords (definite and indefinite articles, pronouns, etc.).[3] We then group individual event observations according to three variables – type of event, company and event date[4] –

[2] This threshold can be modified.
[3] For more precise information on the model and on the implementation in R see Roberts et al (2016).
[4] We use the day, but future analysis will be extended using a longer time frame (days or weeks) as the same event may be reported for more than a day. Variations on this might include allowing a weekly bound. However, a bound is hard to identify as we do not know when the actual event took place. Also bounds may differ across event types. Is it better to use the first day that event appeared or is it better to use the day with the highest frequency, or is it better to use the last day the event is reported as we can be more confident that on that day the event has already happened.

and run the model within each group. If an event is reported by several sources in different formats on the same day, the STM algorithm identifies the repetitions and keeps one of them. STM processing substantially reduces the number of events observations, to 202,912.

A2 / Panel build

To build the panel, we match Companies House companies to enterprises in the Business Structure Database (BSD). Growth Intelligence (GI) data is pre-matched to Companies House identifiers. We then match in patents and trademarks, again using Companies House identifiers.

BSD-Companies House-GI matching

Company level data (Companies House and GI) is based on companies active as of August 2012 (financial year 2013). The UK Data Service team therefore matches companies to enterprises using the 2013 BSD cross-section, which comprises 1,818,263 unique entrefs (which denote individual enterprises in the BSD). The initial matching rate is 61.1% (1,877,600 / 3,074,845 observations matched). Note that due to data protection legislation, we are unable to do this matching ourselves.

We then conduct a number of cleaning and matching sub-routines to optimise the match. Specifically, we drop all observations with no entref, neither in the 2013 BSD nor in the BSD-CH match; drop firms who left the BSD before 2012; drop public sector observations except public sector corporations (e.g. nationalised banks). At the end of these preliminary

cleaning steps we have 1,423,558 observations, for 1,416,218 unique enterprises. This is 75.8% of the original matched sample.

Some of the remaining enterprises are still matched to more than one legal entity (specifically, 78,379 observations, 1.6% of entrefs, 5.8% of observations). These firms are older, larger and richer than sample as a whole.[5] Because we do not have access to identifying information on the BSD side of the data, we are unable to observe the true corporate structures that match to each BSD enterprise. We therefore develop heuristics to give us a panel with 1:1 enterprise:company matching. The majority of corporate legal structures should reduce to this form, especially the single plant SMEs we focus on. We:

1) Keep companies in an enterprise:company group with non-missing year incorporated. Duplicates drop to 2.63% of observations from 5.8% of observations.
2) Keep companies in an enterprise:company group with non-missing CH revenue information and this reduces duplicates to 1.59% of observations. We prefer to have observations with revenues rather than none. Given the observable characteristics of these firms, they are more likely to have revenues to report.
3) Keep companies in an enterprise:company group with highest-reported CH revenue. This step reduces duplicates to 0.08% of observations, as these are likely to be reporting the revenues of the other companies in the group.
4) Shuffle the data and drop any remaining duplicates.[6]

[5] Specifically, the firms in these 1-to-many matches are older than average (mean incorporation 1990 vs. 2002); enter the BSD earlier (1984 vs. 2001); have more plants (94 vs. 6); have higher employment (3096 vs. 187) and employees (3095 vs. 187); have higher annual turnover (£1,200,313 vs. £70,983); are more likely to file revenue to Companies House; and report higher 2010-2013 revenue to Companies House (average £12.4bn vs £2.53bn).

[6] As a sensitivity check we compare characteristics of the retained observations against the modal values of group of linked companies. We find there's a 0.67*** in incorporation years; a 0.70*** correlation in modal founding years; a 0.86*** correlation in modal GI sector ; a 0.86* correlation in group-modal GI products; there's a 0.82*** correlation with the retained and group-modal SIC5 codes. Overall, we conclude that these cleaning rules do not systematically misrepresent underlying corporate structure.

At the end of these further cleaning steps we have 1,261,590 observations, for the same number of unique enterprises. This is 67.2% of the original matched sample.

Finally, we append BSD cross-sections for 2009-2015 and 'panellise' selected CH and GI variables to make a panel of enterprises / firms 2009-2015. This is used to build lags used in the regressions. The raw panel is unbalanced because firms enter the BSD at different times, and because firms drop in and out when they no longer fulfil the BSD criteria (their turnover drops below VAT threshold; they have no employees on PAYE or both of these criteria). In some other cases, especially in earlier years, they file zero against employment or turnover. We fill in gaps in years, while preserving firms' different entry points to the panel. We use a simple interpolation rule to fill in time-varying variables. 4.9% [check] of observations are interpolated. Having built lagged variables, we then reduce the panel to 2014 and 2015, and to single plant SMEs (companies with 250 employees or less).

Patents and trademarks data

We use fuzzy matching routines to match patents data and trademarks data to the panel. Raw patents data is taken from Orbis and covers 169,417 patents filed by 17,131 firms between 1950-2015. Patents are filed to UK, European (EPO), US, PCT and other offices. Patents are dated by priority year, that is, the first year an application enters any patent office in the world. Using application years places patenting activity as close as possible to the underlying invention. 10,360 patents are filed in 2014-2015 by 2973 firms, of which 6440 go to EPO/PCT/US. Orbis has pre-matched patent applicants to UK companies and provides Bureau van Dijk identifiers, which in the majority of cases are identical to, or slightly modified versions of, UK Companies House identifiers. In other cases we match patents to

firms using fuzzy matching on company/applicant names and full UK postcodes.[7] The overall match rate for patents to BSD/CH/GI data is 80.5% for 2014-2015 (2683/3332 observations). We match for 82.5% of companies (2452/2973 firms) in 2014/15.

Trademarks data covers calendar years 2012-2014, and comprises 8,493 UK trademarks filed by 5189 firms. 7129 trademarks are filed in 2014-2015 by 4395 firms. We use fuzzy matching based on company name and postcode to link trademark applicants and Companies House companies. The overall match rate is 89.1% for 2014-2015 (3918 / 4395 obs). We match for 89.1% of firms (3918 / 4395) in 2014/15.

Panel and variables

The final panel contains 2,406,346 observations for 1.22m enterprises in the financial years 2014-2015 (that is, 2013/14 and 2014/2015). Panel variables are defined as follows.

- Age – firms enter the BSD when they start paying UK sales tax (levied on companies with an annual revenue of £75,000 or more), have an employee, or both. Firms enter Companies House when they are incorporated – they may be pre-revenue and pre-employees. We set company age to be incorporation date. Where this is missing we use date of BSD entry.
- Employment and employment growth – following Haltiwanger et al (2013) we use a two-year moving average of employment to correct to regression to the mean. We then define employment growth as the change in E_t and E_{t-1}, weighted by the average of E_t and E_{t-1}. This bounds employment growth to ±200%, removing outliers.

[7] We also match a further 2,404 observations using variations on company name. We do not use these as we cannot be sure that applicants are based in the UK.

- Revenue and revenue growth – defined in the same way as employment, above.
- Revenue per worker / 'revenue productivity' – the BSD does not provide information on conventional labour productivity or TFP measures, but does allow us to directly observe revenue productivity. We define revenue productivity and its growth in the same way as revenue and employment.
- Patents – patents data is coded by application year, that is, the year in which a given patent submission was first submitted to any office in the world. We distinguish between patents filed at major patent offices (USPTO, EPO, PCT framework) and the entire pool of patents, which includes the above plus patents filed only with the UK Intellectual Property Office and with other single-country offices. We make unweighted counts and applicant-weighted counts, where raw patents are divided by the number of applicants. Our preferred measure for regressions is firm-level patent stock with a 15% annual rate depreciation (Hall and Harhoff 2012). Variants use a 40% depreciation rate and a simple cumulative measure.
- Trademarks – trademarks data are coded by application year to the UK IPO. We make simple counts and a TM stock measure specified with a 15% depreciation rate.
- High growth firms and gazelles – we follow the OECD definition of high-growth firms as those with a minimum of ten staff in a given period, where employment or revenue grows by at least 20% in the following three years inclusive. Gazelle firms are high-growth firms less than five years old. We also define high-growth and gazelle firms on the basis of revenue productivity.
- Number of plants – the BSD allows enterprises to exist with zero plants (for example, when all staff are laid off for a period). For ease of interpretation, we set the minimum plant size to be one.

- Legal status – dummies taking the value 1 if the company is a PLC, sole proprietor or partnership / other.
- Enterprise group – a dummy variable taking the value 1 if firms are part of a larger group of companies.
- Companies per enterprise group – for firms in enterprise groups, a count of the number of companies in the group.
- Industry – we use 4-digit SIC 2003 codes as our basic industry unit. 2.88% of companies in the BSD change SIC in our sample. In some cases this is due to change in company activity mix; in other cases ONS reclassifies to correct error, so that reported changes are a lower bound on actual changes. The GI classification of firms is time-invariant, and is based on reported and modelled characteristics for firms active in August 2012 (Nathan and Rosso, 2015).
- Area – we place enterprises in Travel to Work Areas (TTWAs), which are based on commuting patterns and are the best available proxy for local economies; there are 243 of these across the UK. We also use an urban/rural classification of TTWAs taken from Gibbons et al (2011), where 'urban' TTWAs contain at least one city of at least 125,000 people. 2.24% of enterprises change TTWA during the panel period.

References for appendices

Davies N. (2009) *Flat Earth News,* London: Vintage.

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

Gibbons S, Overman HG and Resende G. (2011) Real Earnings Disparities in Britain. *SERC Discussion Paper DP0065.* London: SERC.

Hall BH and Harhoff D. (2012) Recent Research on the Economics of Patents. *NBER Working Paper No. 17773.* Cambridge, MA: NBER.

Haltiwanger J, Jarmin RS and Miranda J. (2013) Who Creates Jobs? Small versus Large versus Young. *The Review of Economics and Statistics* 95: 347-361.

Jiang R, Banchs RE & Li H. (2016). Evaluating and Combining Named Entity Recognition Systems. ACL 2016, 21.

Lafrance A. (2016) Access, Accountability Reporting and Silicon Valley. *Nieman Reports*, 17 August.

Roberts ME, Stewart BM and Tingley D. (2016) stm: R Package for Structural Topic Models. http://www.structuraltopicmodel.com.

Viner K. (2016) How Technology Disrupted the Truth. *Guardian.* 12 July.

A3 / Additional results

Table A1. Summary statistics for single plant SMEs, 2014-2015

Variable	N	mean	sd
Firm has event	2,406,346	0.009	0.093
Total events per firm	2,406,346	0.034	1.602
New product/service launch	2,406,346	0.003	0.057
Total product/service launches per firm	2,406,346	0.013	1.204
Patent stocks	2,406,000	0.009	0.376
Weighted patent stocks	2,406,000	0.009	0.374
EPO / PCT / US patent stocks	2,406,000	0.005	0.250
Weighted EPO / PCT / US patent stocks	2,406,000	0.005	0.249
Weighted trademark stocks	2,406,346	0.002	0.071
Digital tech firm SIC	2,406,346	0.096	0.295
Digital tech firm GI	2,406,346	0.198	0.399
Age in years since BSD entry	2,406,346	11.429	9.533
Age since BSD entry OR incorporation	2,406,346	12.483	12.105
Incorporation year	2,403,004	2004	11.337
Firm is 3 years old or less	2,406,346	0.143	0.35
Foreign ownership	253,320	0.172	0.377
Enterprise has >1 associated company	2,403,975	0.004	0.061
Number of companies per entref	2,403,975	0.01	0.546
Public company	2,406,342	0.982	0.133
Sole proprietor	2,406,342	0.001	0.025
Total employment excluding owners	2,406,346	5.628	13.392
Employment two-year average	2,406,346	5.576	14.506
Annual % jobs growth	2,406,346	0.016	0.319
High jobs growth firm	2,266,150	0.015	0.122
High jobs growth firm ≤ 5 years old	2,266,150	0.004	0.059
Revenue two-year average (,000)	2,406,346	988	66687
Annual % revenue growth	2,406,346	0.01	0.48
High revenue growth firm	2,266,150	0.151	0.358
High revenue growth firm ≤ 5 years old	2,266,150	0.056	0.231
rev per worker two-year average (,000)	2,385,023	154	5209
Annual % rev per worker growth	2,406,346	-0.007	0.494
High rev per worker growth firm	2,266,150	0.129	0.335
High rev per worker growth firm ≤ 5 years old	2,266,150	0.038	0.19

Source: BSD / CH / GI. Minima and maxima suppressed by UK Data Service. Weighted patents and trademarks are weighted by number of applicants. Digital tech firms are defined using SIC codes issued by the UK Office of National Statistics ('SIC') or using bespoke sector/product information developed by Growth Intelligence ('GI'). High growth firms (jobs / revenue / revenue per worker) are defined using the OECD definition of high growth firms. See Appendix A2 for further variable details.

Table A2. Linking past IP activity to product launches. Cross-check with firm fixed effects.

A. pr(Launch)	**(1)**	**(2)**	**(3)**	**(4)**	**(5)**
L1.PCT / EPO / US patent stock	0.008***	0.008***	0.008***	-0.002	-0.002
	(0.001)	(0.001)	(0.001)	(0.003)	(0.010)
L1.TM stock	0.010***	0.008***	0.008***	-0.005	-0.006
	(0.002)	(0.001)	(0.001)	(0.003)	(0.004)
Observations	2406346	2246386	2246386	2406346	2246386
R^2	0.0014	0.0069	0.0069	0.7140	0.7158
B. Launch counts	**(1)**	**(2)**	**(3)**	**(4)**	
L1.PCT / EPO / US patent stock	0.029***	0.024***	0.024***	-0.008	-0.023
	(0.004)	(0.005)	(0.005)	(0.005)	(0.027)
L1.TM stock	0.029***	0.019**	0.019**	-0.011	-0.013
	(0.008)	(0.008)	(0.008)	(0.011)	(0.012)
Observations	2406346	2246386	2246386	2406346	2246386
R^2	0.0000	0.0004	0.0004	0.9805	0.9809
Controls	N	Y	Y	N	Y
Year fixed effect	N	N	Y	Y	Y
Firm fixed effects	N	N	N	Y	Y

Source: BSD / CH /Orbis / IPO / GI. Column 1 fits IP variables, column 2 adds controls, column 3 adds firm fixed effects, column 4 adds year dummy, column 5 adds TTWA area dummy, column 6 adds SIC4 industry dummy. Controls fitted include log 2-year mean turnover, age since BSD entry OR incorporation, number of companies per entref, enterprise has >1 associated company dummy, legal status dummies (public company, sole proprietor, reference = other). Controls lagged one year except age. Standard errors bootstrapped, 400 reps. *** significant at 1%, ** significant at 5%, * significant at 10%.

Table A3. Linking past IP activity to product launches. Dummy model. Cumulative patenting.

pr(product launch)	Cumulative					
	(1) L0	(2) L1	(3) L2	(4) L3	(5) L4	(6) L5
Cumulative PCT / EPO / US patent count	0.004***	0.004***	0.005***	0.007***	0.007***	0.006*
	(0.001)	(0.001)	(0.001)	(0.001)	(0.002)	(0.003)
TM stock	0.010***	0.007***	0.010**	0.010**	0.010*	0.010**
	(0.002)	(0.001)	(0.005)	(0.005)	(0.005)	(0.005)
Ave pre-2009 patenting	-0.002	-0.004*	-0.005**	-0.005***	-0.004*	-0.001
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
Firm patents pre-2009	0.036***	0.036***	0.037***	0.038***	0.037***	0.036***
	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)	(0.007)
Observations	2225176	2225176	2225176	1938369	1691279	1517738
R^2	0.0128	0.0126	0.0124	0.0130	0.0137	0.0143
pr(product launch)	**40% depreciation**					
	(1) L0	(2) L1	(3) L2	(4) L3	(5) L4	(6) L5
Patent stocks depreciated 40% pa	0.009***	0.008***	0.009***	0.011***	0.012***	0.008**
	(0.001)	(0.001)	(0.001)	(0.002)	(0.003)	(0.004)
TM stock	0.010***	0.007***	0.010**	0.010**	0.010*	0.010**
	(0.002)	(0.001)	(0.005)	(0.005)	(0.005)	(0.005)
Ave pre-2009 patenting	-0.002	-0.003	-0.004*	-0.005***	-0.004**	-0.001
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
Firm patents pre-2009	0.032***	0.033***	0.034***	0.035***	0.036***	0.036***
	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)	(0.007)
Observations	2225176	2225176	2225176	1938369	1691279	1517738
R^2	0.0129	0.0126	0.0125	0.0131	0.0138	0.0143

Source: BSD / CH / Orbis / IPO / GI. All models fit controls, area, time and industry dummies. Standard errors clustered on 2-digit SIC. Controls as in Table A3. *** significant at 1%, ** significant at 5%, * significant at 10%.

Table A4. Linking past IP activity to product launches. Counts model. Cumulative patenting.

Launch counts	Cumulative					
	(1) L0	(2) L1	(3) L2	(4) L3	(5) L4	(6) L5
Cumulative PCT / EPO / US patent count	0.014**	0.014***	0.021***	0.025***	0.035**	0.043**
	(0.006)	(0.005)	(0.005)	(0.008)	(0.013)	(0.021)
TM count	0.024**	0.017**	0.030	0.031	0.030	0.032
	(0.011)	(0.007)	(0.028)	(0.028)	(0.030)	(0.032)
Ave pre-2009 patenting	-0.022**	-0.027**	-0.033***	-0.032***	-0.034***	-0.028***
	(0.011)	(0.010)	(0.010)	(0.008)	(0.008)	(0.009)
Firm patents pre-2009	0.116**	0.117**	0.121**	0.121**	0.121**	0.116**
	(0.048)	(0.048)	(0.048)	(0.048)	(0.048)	(0.049)
Observations	2225176	2225176	2225176	1938369	1691279	1517738
R^2	0.0009	0.0009	0.0009	0.0009	0.0010	0.0010
Launch counts	**40% depreciation**					
	(1) L0	(2) L1	(3) L2	(4) L3	(5) L4	(6) L5
Patent stocks depreciated 40% pa	0.032***	0.028***	0.035***	0.036***	0.051***	0.055*
	(0.010)	(0.007)	(0.007)	(0.010)	(0.019)	(0.028)
TM count	0.023**	0.016**	0.029	0.030	0.030	0.032
	(0.011)	(0.007)	(0.028)	(0.029)	(0.030)	(0.032)
Ave pre-2009 patenting	-0.022*	-0.024**	-0.030***	-0.030***	-0.035***	-0.030***
	(0.011)	(0.012)	(0.011)	(0.009)	(0.009)	(0.009)
Firm patents pre-2009	0.103**	0.105**	0.109**	0.112**	0.113**	0.113**
	(0.048)	(0.048)	(0.048)	(0.049)	(0.049)	(0.050)
Observations	2225176	2225176	2225176	1938369	1691279	1517738
R^2	0.0009	0.0009	0.0009	0.0009	0.0010	0.0010

Source: BSD / CH / Orbis / IPO / GI. All models fit controls, area time and industry dummies. Standard errors clustered on 2-digit SIC. Controls as in Table A3. *** significant at 1%, ** significant at 5%, * significant at 10%.

Table A5. Linking the probability of a launch on past patents and trademarking. Manufacturing vs services.

Manufacturing (A-D)	**(1)** **L0**	**(2)** **L1**	**(3)** **L2**	**(4)** **L3**	**(5)** **L4**	**(6)** **L5**
PCT / EPO / US patent stock	0.009*** (0.003)	0.009*** (0.003)	0.007** (0.002)	0.008** (0.003)	0.009** (0.003)	0.006 (0.005)
TM stock	0.010*** (0.003)	0.004 (0.003)	0.002 (0.005)	0.002 (0.006)	0.002 (0.006)	0.002 (0.007)
Ave pre-2009 patenting	-0.011 (0.009)	-0.011 (0.009)	-0.008 (0.008)	-0.008 (0.008)	-0.008 (0.008)	-0.004 (0.008)
Firm patents pre-2009	0.030 (0.018)	0.031 (0.018)	0.030* (0.018)	0.031* (0.017)	0.031* (0.017)	0.030* (0.017)
Observations R^2	211440 0.0230	211440 0.0225	211440 0.0218	194045 0.0229	178708 0.0237	166858 0.0241
Services (G-O)	**(1)** **L0**	**(2)** **L1**	**(3)** **L2**	**(4)** **L3**	**(5)** **L4**	**(6)** **L5**
PCT / EPO / US patent stock	0.005*** (0.001)	0.005*** (0.001)	0.007*** (0.001)	0.008*** (0.002)	0.009** (0.003)	0.006 (0.005)
TM stock	0.009*** (0.002)	0.007*** (0.001)	0.012** (0.006)	0.012** (0.006)	0.012* (0.006)	0.013** (0.006)
Ave pre-2009 patenting	-0.003** (0.001)	-0.003** (0.001)	-0.005*** (0.001)	-0.005*** (0.001)	-0.004** (0.002)	-0.001 (0.002)
Firm patents pre-2009	0.040*** (0.006)	0.041*** (0.006)	0.042*** (0.006)	0.043*** (0.006)	0.043*** (0.006)	0.044*** (0.006)
Observations R^2	1757429 0.0120	1757429 0.0118	1757429 0.0117	1516727 0.0122	1310311 0.0130	1166719 0.0136

Source: BSD / CH /Orbis / IPO / GI. Specification as in Table A3. *L* denotes lag period. Trademarks lagged back 2 periods in columns 3-5. *** significant at 1%, ** significant at 5%, * significant at 10%. SIC sections A-D cover Agriculture, hunting and forestry; Fishing; Mining and quarrying; Manufacturing. SIC sections G-O cover Wholesale and retail trade, repair of vehicles and goods; Hotels and restaurants; Transport, storage and communications; Financial intermediation; Real estate, renting and business activities; Public administration and defence; Education; Health and social work; Other community, social and personal services.

Table A6. Linking launch counts on past patents and trademarking. Manufacturing vs services.

Manufacturing (A-D)	**(1)** **L0**	**(2)** **L1**	**(3)** **L2**	**(4)** **L3**	**(5)** **L4**	**(6)** **L5**
PCT / EPO / US patent stocks	0.023*** (0.006)	0.021*** (0.006)	0.018*** (0.005)	0.023*** (0.008)	0.033** (0.014)	0.042* (0.024)
TM count	0.027** (0.013)	0.015 (0.014)	0.003 (0.016)	0.003 (0.017)	0.002 (0.018)	0.001 (0.019)
Ave pre-2009 patenting	-0.051 (0.033)	-0.050 (0.033)	-0.043 (0.030)	-0.046 (0.031)	-0.049 (0.033)	-0.044 (0.034)
Firm patents pre-2009	0.104 (0.074)	0.106 (0.075)	0.105 (0.073)	0.106 (0.072)	0.108 (0.074)	0.105 (0.074)
Observations R^2	211440 0.0054	211440 0.0053	211440 0.0052	194045 0.0055	178708 0.0058	166858 0.0061
Services (G-O)	**(1)** **L0**	**(2)** **L1**	**(3)** **L2**	**(4)** **L3**	**(5)** **L4**	**(6)** **L5**
PCT / EPO / US patent stocks	0.021*** (0.007)	0.021*** (0.006)	0.028*** (0.006)	0.030*** (0.010)	0.042* (0.020)	0.045 (0.032)
TM count	0.020 (0.014)	0.015* (0.008)	0.036 (0.035)	0.036 (0.035)	0.038 (0.037)	0.040 (0.039)
Ave pre-2009 patenting	-0.029** (0.014)	-0.031** (0.014)	-0.036** (0.014)	-0.034*** (0.011)	-0.037*** (0.010)	-0.030*** (0.010)
Firm patents pre-2009	0.139* (0.079)	0.141* (0.080)	0.144* (0.081)	0.147* (0.084)	0.147* (0.085)	0.149* (0.085)
Observations R^2	1757429 0.0009	1757429 0.0009	1757429 0.0009	1516727 0.0010	1310311 0.0010	1166719 0.0011

Source: BSD / CH /Orbis / IPO / GI. Specification as in Table A3, main paper. *L* denotes lag period. Trademarks lagged back 2 periods in columns 3-5. *** significant at 1%, ** significant at 5%, * significant at 10%. SIC sections A-D cover Agriculture, hunting and forestry; Fishing; Mining and quarrying; Manufacturing. SIC sections G-O cover Wholesale and retail trade, repair of vehicles and goods; Hotels and restaurants; Transport, storage and communications; Financial intermediation; Real estate, renting and business activities; Public administration and defence; Education; Health and social work; Other community, social and personal services.

Table A7. Robustness checks for firms with events. Launch dummy model.

Main checks	Digital tech SIC	Digital tech GI	Obs
Baseline	0.152***	0.085***	
	(0.007)	(0.004)	29616
	0.0213	0.0116	
Specification checks			
Log employees, 2-year average	0.153***	0.087***	
	(0.006)	(0.005)	29598
	0.0217	*0.0123*	
Patents 2-period lag	0.152***	0.085***	
	(0.007)	(0.005)	29616
	0.0213	*0.0115*	
Drop SIC switchers	0.149***	0.079***	
	(0.007)	(0.005)	28573
	0.0198	*0.0099*	
Drop TTWA switches	0.160***	0.087***	
	(0.007)	(0.005)	28754
	0.0225	*0.0118*	
SIC4*year dummies	0.160***	0.095***	
	(0.007)	(0.005)	29616
	0.0238	*0.0146*	
SIC4*TTWA clustered errors	0.152***	0.085***	29616
	(0.007)	(0.005)	
	0.0213	*0.0116*	
IPC1*year fixed effects	0.147***	0.078***	
	(0.007)	(0.005)	29616
	0.0195	0.0095	
Drop singletons	0.150***	0.083***	
	(0.007)	(0.005)	28220
	0.0214	0.0115	
Functional form checks			
OLS	0.165***	0.101***	
	(0.025)	(0.021)	29187
	0.0166	*0.0107*	
	-1.642e+04	*-1.651e+04*	
Logit, marginal effects	0.144***	0.094***	
	(0.018)	(0.018)	29187
	0.0138	*0.0093*	
	-1.593e+04	*-1.600e+04*	

Source: BSD / CH /Orbis / IPO / GI. Second stage coefficients except for functional form checks. R^2 and log-likelihood for functional form tests are in italics. Logit coefficients are marginal effects on the mean. All models use controls as in Table 12, main paper, plus firm/area/SIC4/year FE, except where specified. Standard errors are bootstrapped, 400 reps. *** significant at 1%, ** significant at 5%, * significant at 10%. 8

Table A8. Robustness checks for firms with events. Launch counts model.

Main checks	Digital tech SIC	Digital tech GI	Obs
Baseline	0.583***	1.249***	
	(0.120)	(0.220)	29187
	0.0003	*0.0027*	
Log employees, 2-year average	0.583***	1.265***	
	(0.120)	(0.221)	29171
	0.0003	*0.0028*	
Patents 2-period lag	0.581***	1.247***	
	(0.120)	(0.220)	29187
	0.0003	0.0027	
Drop SIC switchers	0.566***	1.246***	
	(0.118)	(0.213)	28173
	0.0003	0.0026	
Drop TTWA switches	0.570***	1.260***	
	(0.118)	(0.228)	28343
	0.0003	0.0027	
SIC4*year dummies	0.592***	1.261***	
	(0.120)	(0.220)	29187
	0.0003	*0.0028*	
SIC4*TTWA clustered errors	0.583***	1.249***	29187
	(0.120)	(0.220)	
	0.0003	*0.0027*	
IPC1*year fixed effects	0.585***	1.256***	
	(0.120)	(0.220)	29187
	0.0003	0.0027	
Drop singletons	0.578***	1.169***	
	(0.117)	(0.209)	28220
	0.0003	*0.0024*	
Functional form checks			
OLS	0.538***	1.189***	
	(0.149)	(0.264)	29187
	0.0003	*0.0024*	
	-1.11E+05	*-1.11E+05*	
Poisson, marginal effects	0.455***	1.004***	
	(0.101)	(0.213)	29187
	0.0046	*0.0385*	
	-9.24E+04	*-8.93E+04*	

Source: BSD / CH /Orbis / IPO / GI. Second stage coefficients except for functional form checks. R2 and log-likelihood for functional form tests are in italics. Poisson coefficients are marginal effects on the mean. All models use controls as in Table 12, main paper, plus firm/area/SIC4/year FE, except where specified. Standard errors are bootstrapped, 400 reps. *** significant at 1%, ** significant at 5%, * significant at 10%.

Table A9. Tech firms and product launches. Firms with events. Interactions.

Depvar = pr(launch)	**(1)**	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**
Digital tech firm SIC	0.152***	0.141***	0.150***			
	(0.007)	(0.009)	(0.018)			
Digital tech firm GI				0.085***	0.082***	0.107***
				(0.005)	(0.006)	(0.013)
Small business		0.035***			0.036***	
		(0.005)			(0.005)	
Medium business		0.076***			0.083***	
		(0.007)			(0.008)	
Digital Tech#small business		0.005			0.002	
		(0.014)			(0.011)	
Digital Tech#medium business		0.067***			0.028	
		(0.022)			(0.017)	
Urban TTWA			-0.005			-0.001
			(0.005)			(0.006)
Digital tech #urban TTWA			0.003			-0.026*
			(0.019)			(0.014)
Observations	29,616	28,988	29,616	29,616	28,988	29,616
R^2	0.021	0.029	0.021	0.012	0.019	0.012

Source: BSD / CH /Orbis / IPO / GI. Second stage coefficients only. All models use controls as above, plus firm/area/SIC4/year FE. Standard errors bootstrapped, 400 reps. Controls as in Table 12. Reference category for size is micro firm (<10 employees). Urban Travel To Work Areas contain an urban core of at least 125,000 people, as defined in Gibbons et al (2011). *** significant at 1%, ** significant at 5%, * significant at 10%.

Table A10. Tech firms and product launch counts. Firms with events. Interactions.

Depvar = launch counts	**(1)**	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**
Digital tech SIC	0.576***	0.300***	0.440***			
	(0.120)	(0.094)	(0.122)			
Digital tech GI				1.228***	0.671***	0.531***
				(0.215)	(0.204)	(0.183)
Small business		-0.216***			-0.238***	
		(0.077)			(0.050)	
Medium business		1.022**			0.180	
		(0.467)			(0.206)	
Digital Tech#small business		0.659***			0.397	
		(0.234)			(0.249)	
Digital Tech#medium		0.442			3.968**	
business		(0.640)			(1.663)	
Urban TTWA			0.300***			0.082
			(0.111)			(0.078)
Digital tech #urban TTWA			0.157			0.805**
			(0.193)			(0.317)
Observations	29,616	28,988	29,616	29,616	28,988	29,616
R^2	0.000	0.002	0.000	0.003	0.007	0.003

Source: BSD / CH /Orbis / IPO / GI. Second stage coefficients only. All models use controls as above, plus firm/area/SIC4/year FE. Standard errors bootstrapped, 400 reps. Controls as in Table 12. Reference category for size is micro firm (<10 employees). Urban Travel To Work Areas contain an urban core of at least 125,000 people, as defined in Gibbons et al (2011). *** significant at 1%, ** significant at 5%, * significant at 10%.

Table A11. Lasso diagnostics for event prediction regressions.

Step	Cp	R^2	Action
1	37864.26	0	
2	26356.49	0.0058	+ L1.ln(turnover)
3	11615.81	0.0132	+L1.firm has >1 associated company
4	9265.177	0.0143	+ L1.firm is PLC
5	2797.294	0.0176	+ L1.EPO / PCT / US patent stocks
6	1590.066	0.0182	+ L1.TM stocks
7	412.4431	0.0188	+ L1.Number of linked companies
8	173.5881	0.0189	+ L1.Firm is sole proprietor
9	137.9492	0.0189	+ Firm is a startup (<=3 years old)
10	47.0679	0.019	+ L1.High rev per worker growth firm
11	11	* 0.0190	+ L1.% rev per worker growth, two year average

Source: BSD / CH /Orbis / IPO / GI. Underlying sample = 2.04m observations.

Table A12. Lasso diagnostics for event prediction: cross-check using kitchen-sink regression.

Step	Cp	R^2	Action
1	35175.28	0	
2	31181.5	0.0143	+L1.firm has >1 associated company
3	24487.01	0.0383	+ L1.High employment growth firm
4	17526.35	0.0633	+age
5	8119.979	0.0971	+ Firm patents pre-2009
6	5914.698	0.105	+L1.Foreign ownership
7	4852.502	0.1088	+ L1.firm is PLC
8	4845.979	0.1088	+L1.revenue two-year average
9	4389.533	0.1105	+ L1.Number of linked companies
10	2640.637	0.1168	+ L1.% rev per worker growth, two year average
11	1851.11	0.1196	+ L1.TM stocks
12	1510.283	0.1208	+ L1.High rev per worker growth firm
13	1130.308	0.1222	+L1.Firm is in urban TTWA
14	459.3101	0.1246	+ L1.EPO / PCT / US patent stocks
15	427.4299	0.1247	+ L1.High employment growth firm
16	378.6706	0.1249	+ L1.% employment growth, two year average
17	271.978	0.1253	+Growth Intelligence estimated revenue
18	267.2369	0.1253	+ Firm average patents pre-2009
19	27.2957	0.1262	+ Firm is a startup (<=3 years old)
20	20	* 0.1262	+ L1.High rev per worker growth firm

Source: BSD / CH /Orbis / IPO / GI. Underlying sample = 2.04m observations.

Table A13. Comparing weighted and unweighted samples.

Variable	Obs	Mean	Std. Dev.
Firms without events			
L1. ln(turnover)	2,228,312	5.137	1.402
Firm is a startup	2,385,248	0.144	0.351
L1.EPO / PCT / US patent stocks	2,385,248	0.005	0.215
L1.TM stocks	2,385,248	0.002	0.077
L1.Number of linked companies	2,383,148	0.008	0.512
L1.firm has >1 associated company	2,383,148	0.003	0.057
L1.firm is PLC	2,385,247	0.983	0.131
L1.Firm is sole proprietor	2,385,247	0.000	0.019
L1.% rev per worker growth, two year average	2,385,248	-0.015	0.486
L1.High rev per worker growth firm	1,954,547	0.122	0.327
Firms with events, unweighted			
pr(launch)	*21,098*	*0.368*	*0.482*
launch count	*21,098*	*1.494*	*12.770*
L1. ln(turnover)	20,221	6.768	2.062
Firm is a startup	21,098	0.040	0.196
L1.EPO / PCT / US patent stocks	21,098	0.104	1.167
L1.TM stocks	21,098	0.020	0.230
L1.Number of linked companies	20,827	0.208	2.103
L1.firm has >1 associated company	20,827	0.058	0.234
L1.firm is PLC	21,098	0.922	0.268
L1.Firm is sole proprietor	21,098	0.000	0.010
L1.% rev per worker growth, two year average	21,098	0.011	0.537
L1.High rev per worker growth firm	19,978	0.139	0.346
Firms with events, reweighted			
pr(launch)	*18,931*	*0.469*	*0.499*
launch count	*18,931*	*2.340*	*17.819*
L1. ln(turnover)	18,931	9.088	2.107
Firm is a startup	18,931	0.000	0.020
L1.EPO / PCT / US patent stocks	18,931	11.459	26.952
L1.TM stocks	18,931	0.089	0.644
L1.Number of linked companies	18,931	1.358	6.089
L1.firm has >1 associated company	18,931	0.387	0.487
L1.firm is PLC	18,931	0.848	0.359
L1.Firm is sole proprietor	18,931	0.000	0.001
L1.% rev per worker growth, two year average	18,931	0.117	0.549
L1.High rev per worker growth firm	18,931	0.294	0.456

Source: BSD / CH /Orbis / IPO / GI. Minima and maxima suppressed to avoid disclosure.