Economic complexity and the emergence of new ideas

Alex Bishop*† Ana-Maria Dobre‡ Juan Mateos-Garcia*

September 2, 2018

Abstract

Recent research has shown a strong link between the economic complexity of a country and important economic indicators such as GDP per capita and lower inequality. However, the mechanisms underpinning this relationship remain poorly understood. While it is often argued that complex ecosystems provide a fertile ground for novel combinations of ideas leading to the emergence of impactful new technologies and industries, this process is hard to measure with aggregate data structured around industrial codes offering a lagging view of the economy. We seek to overcome this challenge by combining official data about industrial activity in UK Local Authority Districts (LADs) with a novel dataset containing the content and meta-data from almost a million UK business websites. We measure economic complexity using ECI and a Fitness based measure, which respectively capture a location's specialisation in unique, knowledge intensive sectors, and a weighted measure of economic diversity. We use a complex networks approach to topic modelling to detect 'topics' in the websites of businesses and analyse their relation to economic complexity and industrial sectors before going on to measure emergence by looking for novel words in the websites of businesses in different locations and sectors. We show that LADs with high ECI scores have a stronger share of companies active in emerging technologies, even after we control for their industrial composition. Further, and contrary to our initial expectations, emergent companies are more industrially diversified in locations with high ECI scores than locations with high Fitness scores. One potential reason is that the innovative sectors that high ECI locations specialise on are lead adopters of emerging technologies that are subsequently diffused into other parts of the local economy.

*Nesta, 58 Victoria Embankment, EC4Y 0DS
†Corresponding author: alex.bishop@nesta.org.uk
‡Glass

1 Introduction

1.1 A complex turn in the analysis of economic development

In the last ten years, researchers have found a robust link between the economic complexity of nations and their wealth in terms of GDP per capita and other important economic outcomes such as lower inequality [1, 2, 3]. The notion of economic complexity has been subsequently extended to the analysis of technologies (based on patent data) and industries (based on industrial clustering), and used to study regions as well as nations [4, 5, 6].

Economic complexity refers to the diversity and sophistication of productive capabilities present in a country or region: if a location displays a comparative advantage in the supply of complex products, this suggests the presence of various hard to imitate capabilities enabling it to capture more market share [5][1]. Economic complexity also shapes a location's future specialisation trajectory through the *Principle of Relatedness*. According to this principle, nations tend to enter activities for which they already have relevant knowledge and capabilities [7, 8]. Economically complex locations have more 'options' to diversify their economy in response to new opportunities and shocks.

The seminal approach for estimating economic complexity, the *Economic Complexity Index* (ECI) is based on the *method of reflections*, a recursive algorithm that produces a sum of the number of products where a country has a revealed comparative advantage (or of industries where it specialises) "weighted" by their ubiquity (their propensity to appear in many other countries) [1].

[1]This distinguishes economic complexity metrics from other measures of diversity that consider the number of activities present in a location without taking into account their broader distribution.

1.2 Complexity and emergence

Outside of economics, the concept of complexity is often associated with the idea of emergence: complex systems give rise to new phenomena which are qualitatively different from their constituent parts (the whole is more than the sum of the parts), like a whirlpool emerging from flowing water, or an ant-hive emerging from the decentralised behaviour of ants [9]. In economics, these emergent phenomena take the form of firms, industries, markets and innovations (novel combinations of ideas) that can give rise to new technologies and industries [10].

Economic geographers have long studied how spatial proximity facilitates these emergent phenomena by lowering transaction costs, helping to build up trust, facilitating interactive learning and accelerating the flow of ideas inside and between sectors through localised knowledge spillovers [11, 12]. Although all these processes could act as mediators between economic complexity and economic development, this link remains understudied[2]. This is in part due to limitations in the data: it is not possible to identify novel phenomena with structured data sources based on pre-existing product, industry and technological categories offering a lagging view of the economy. For example, new sectors such as the Internet of Things or Virtual Reality are not present in the Standard Industrial Classification (SIC) taxonomy used to structure economic data in the UK, dating back to 2007. But how can we determine if more economically complex areas have a stronger propensity to nurture emergent activities?

Here, we attempt to do this by applying Natural Language Processing (NLP) techniques to a novel dataset with text from hundreds of thousands of UK business websites collected by Glass, an economic intelligence start-up. Our prior is that the description of

[2]In a way, the Principle of Relatedness captures technological emergence because it describes how a location becomes specialised in an industry that was not present there before. However, that kind of novelty is different from the generation of truly novel activities - including innovative variation in the content of economic activities in a pre-existing sector.

business activities in those websites will include information about how those businesses are developing and adopting new ideas, helping us to create metrics of emergent activity that we can correlate with the economic complexity of different locations [13].

1.3 Measures of complexity

As mentioned in section 1.1, ECI and the method of reflections permit us to measure the complexity of economies and products. More precisely, by letting $X \in \mathbb{R}^{N_c \times N_p}$ denote the extensive country-product export matrix of N_c countries and N_p products where X_{cp} denotes the exports of country c in product p, and $M \in \mathbb{R}^{N_c \times N_p}$ the intensive country-product export matrix where M_{cp} is unity if country c has a revealed comparative advantage in product p and zero otherwise (1). The method of reflections proceeds by calculating the ubiquity (how easy it is to make a product), $k_p^{(0)} = \sum_c M_{cp}$, of a product p, and the diversity (how many products a country makes), $k_c^{(0)} = \sum_p M_{cp}$, of a country c and then recursively corrects each of these measures using the information contained within the other (2,3). A numerically stable way of calculating ECI was established in [14] by noting its equivalence (up to a sign) to calculating the eigenvector corresponding to the second largest eigenvalue of $\tilde{M}$ (4). ECI has been interpreted as both a corrected diversity measure and an eigenvector centrality algorithm [15]; however ECI is orthogonal to diversity (though correlated) [16]. Mealy et. al [17] demonstrate the correct interpretation of ECI is a spectral clustering algorithm that partitions the country-export similarity graph into two clusters such that countries with similar ECI have similar exports. This interpretation puts on solid footing previous (little discussed) observations[18] that the picture ECI (and its derivatives) gives are impressionistic in the sense that they are globally correct, but do not hold locally (e.g. 3rd vs 4th ranking may not be meaningful).

$$M_{cp} = \mathbb{1}\left(\frac{X_{cp}/\sum_p X_{cp}}{\sum_c X_{cp}/\sum_{c,p} X_{cp}} > 1\right) \tag{1}$$

$$k_c^{(N)} = \frac{1}{k_c^{(0)}} \sum_p M_{cp} k_p^{(N-1)} \tag{2}$$

$$k_p^{(N)} = \frac{1}{k_p^{(0)}} \sum_c M_{cp} k_c^{(N-1)} \tag{3}$$

$$\tilde{M} = \mathrm{diag}\left(k_c^{(0)}\right) X \mathrm{diag}\left(k_p^{(0)}\right) X^T \tag{4}$$

A non-linear modification to ECI called Fitness[19, 14] is also commonly used and addresses two shortcomings of ECI - though its non-linear nature may induce instabilities [20, 15]. Firstly, ECI is an average of 'corrected' product complexity which can lead to the counter-intuitive situation where a country producing a range of products from low to high complexity may be ranked lower than a country producing one moderately complex product. Furthermore, the linear nature of ECI (the relationship between the complexity of countries and products is linear) means that high complexity of countries producing unsophisticated products (e.g. oil) push up the complexity of the product and therefore the complexity of less complex countries also producing it.

The country complexity of the fitness measure, F_c^N (5), avoids the averaging problem and considers only the sum over all products weighted by complexity of products, Q_p^N (6).

$$F_c^{(N)} = \sum_p X_{cp} Q_p^{(N-1)} \tag{5}$$

$$Q_p^{(N)} = \sum_p \frac{X_{cp}}{F_c^{(N-1)}} \tag{6}$$

$$Fitness_c+ = \log(F_c^{(\infty)}) - \log\left(\sum_p \frac{X_{cp}}{\sum_c X_{cp}}\right) \tag{7}$$

The nested structure of M_{cp} suggests that the more countries make a product, the less complex it is, therefore the denominator of Q_p^N contains a sum over all countries making the product; however not all countries are alike. High complexity countries make many products so should not decrease a product's complexity too much therefore each term in the denominator is weighted with the inverse of country complexity.

The underlying form of the fitness algorithm was rediscovered from an alternative perspective several years later and named ECI+[3] [21] before the similarities between the two algorithms were noted, sparking debate on the originality and utility of further research into alternative formulations of economic complexity. Provoking further discussion around this topic is not the purpose of this paper as we wish to explore the use of novel datasets to measure emergence and how these relate to economic complexity. To this end we use both ECI, and the Fitness on the extensive 'export' matrix normalised with geometric mean each iteration and incorporating the country specific offset of [21] which we shall henceforth call Fitness+. Choosing the extensive matrix has the advantage of measuring activity continuously, instead of as a binary variable (whether a country has a relative specialisation in a product or not).

[3]ECI+ adds a country dependent off-set in an attempt to correct for country size, and uses a geometric rather than harmonic mean normalisation at each iteration to account for the extensive nature of the data

1.4 Hypotheses

Previous research suggests that ECI and Fitness+ capture different dimensions of economic complexity. ECI measures specialisation in unique (complex, knowledge intensive) sectors, while Fitness+ measures the weighted diversity of an economy. We therefore expect these two metrics to be associated with different types of emergence:

1. In locations with high ECI scores, emergence should be concentrated in unique sectors (*concentrated emergence*)

2. In locations with high Fitness+ scores, emergence should be more widely dispersed across the local economy (*diversified emergence*)

Evidence of these links would support the idea that one of the channels through which economic complexity contributes to economic development is by facilitating the emergence of new ideas. It would also help us understand the specifics of this relationship. Thinking about the various benefits from co-location identified in economic geography, concentrated emergence would be conceptually closer to the idea of agglomeration economies between specialised or related industries, while diversified emergence would echo the idea of urbanisation economies based on less predictable flows of knowledge between co-located, unrelated sectors. We explore these links in section 3.

2 Data

2.1 Official data

We estimate our economic complexity indexes using employment data from BRES (The Business Register and Employment Survey) and IDBR (Interdepartmental Business Regis-

ter) accessed from NOMIS, an online portal with labour market data for the UK [4]. BRES is an annual survey of 80,000 businesses in Great Britain providing official employee and employment estimates by detailed geography and industry. The Interdepartmental Business Register is a register of 2.6 million VAT and PAYE registered businesses in the UK derived from administrative sources. It is used as the sampling frame for official business surveys in the UK.

Previous studies have shown a robust relationship between economic complexity and other important economic variables. Here, we use salary data from ASHE (Annual Survey of Hours and Earnings) and local economy estimates of GVA per capita generated by the Office for National Statistics (ONS) as economic benchmarks for our economic complexity metrics.[5].

2.2 Glass data

Glass is a UK startup that has developed Artificial Intelligence (AI) technology that can understand text at scale. The core technology is in the field of machine language 'understanding', which aims to give machines the power to understand not just words but entire sentences and eventually paragraphs. The company has built an intelligent web crawler that reads and interprets public websites automatically. To identify UK businesses, the crawler is set to read websites representing public and private organisations that target a UK audience, or have adopted the .uk domain address. Websites are considered if they are:

(i) Written in English,

(ii) Mention a UK address in their pages,

[4] https://www.nomisweb.co.uk/

[5] The ASHE data were also obtained from NOMIS. The GVA per capita data are available from https://bit.ly/2krQBCs.

(iii) Have some depth of representation for the organisation

Starting with over 204 million web pages, 894,277 business websites with a qualified depth of data are found. Then, the content of each website is read and relevant text entities (e.g. business descriptions) are detected with state of the art precision ($> 95\%$). Business descriptions are identified with a machine learning model that considers multiple features such as location on the web-page, use of specific keywords and phrases, sentence structure etc.

Based on descriptions and other attributes, each business is classified into one or more sectors and assigned a weight showing its proximity to the sector. Specialised businesses tend to have a single sector with high weight, while those with a diversified activity have multiple sector predictions with lower weight values. The sector classification has been trained using a sample of company classifications based on an industrial taxonomy created by LinkedIn.

2.3 Geographical unit of analysis

We study 380 Local Administrative Districts capturing local government boundaries in Great Britain. Following [17], we use this administrative geography instead of functional economic areas based on commuting patterns (Travel to Work Areas - TTWAs) because our benchmarking datasets (median salaries and GVA per capita) are not available at the TTWA level.

We geocode the business websites in the Glass data with their main postcode via NSPL (National Statistics Postcode Lookup), a lookup between postcodes and official geographies in the UK [6]. At the end of this process, we end with $\sim 400,000$ unique geocoded business websites.

[6] We focus on business websites with less than 5 postcodes, where the main postcode tends to provide a valid match for geocoding

2.4 Sector segmentation

BRES and IDBR data are classified into industries using 4-digit SIC codes. We have clustered these codes into *industrial segments* of related industries with the algorithm developed in [22], which measures this relatedness based on employment and business co-location, occupational composition of the workforce and business to business trade based on input-output tables. This gives us a list of 72 unique industries [7]. This segmentation should reduce misclassification in very granular SIC categories, and helps us focus subsequent complexity and clustering analyses on distinct industries, reducing the risk that we over-estimate the diversity of locations that host many different sectors that are very similar to each other.

As mentioned, the Glass data are labelled with multiple sector tags through a predictive analysis of their website content. The LinkedIn taxonomy used for this contains 108 industries. We classify each website into its top sector based on the weight generated by Glass, and then into its industrial segment through a provisional lookup table [8].

3 Results

3.1 The economic complexity of British LADs

We start by considering our complexity indexes (ECI and Fitness+), their geography and sectoral distribution and their relationship with other variables of interest, including measures of economic diversity and economic benchmarks [9]. One of our goals here is to assess

[7] See https://www.nesta.org.uk/blog/complex-places-for-complex-times-an-analysis-of-the-complexity-of-uk-local-economies-and-their-future-evolution/ for further detail.

[8] Longer term, we plan to match the Glass data with the UK business register (Companies House), and use the SIC codes available there to classify companies into segments, as well as obtain their registered and trading addresses.

[9] We have estimated diversity counting the number of sectors with a RCA above 1 in a LAD according to the BRES and IDBR data, and with Shannon entropy.

the differences between ECI and Fitness+, since this will have important implications for the analysis of their relationship with emergence.

3.1.1 Economic complexity indexes and their interpretation

We have estimated ECI and Fitness+ with the BRES and IDBR data following the steps described in [2] and [21]. We refer to them as `bres_eci`, `idbr_eci`, `bres_fit_p` and `idbr_fit_p`. We present the results in figure 1, and correlation between variables in the first set of rows of the correlation matrix in figure 2.

Variable	Top 3 LAD	Bottom 3 LAD
bres_eci	City of London, Islington, Camden	Fenland, Neath Port Talbot, South Holland
idbr_eci	Lambeth, St Albans, Chiltern	Blaenau Gwent, Scarborough, Isles of Scilly
bres_fit_p	Aberdeen City, Westminster, Birmingham	Rutland, Clackmannanshire, Isles of Scilly
idbr_fit_p	Birmingham, Westminster, Aberdeen City	Merthyr Tydfil, Clackmannanshire, Isles of Scilly

Table 1: Top 3 LAD and bottom 3 LAD by complexity index

The heatmap in figure 3 shows that locations with high scores in `bres_eci` and `idbr_eci`(top rows) tend to have strong specialisation in knowledge intensive business sectors such as `services_computing` and `services_kibs` and lower specialisation in manufacturing sectors. Fitness+ rankings are less visibly associated with distinct specialisation patterns.

We have also considered the distribution of pairwise similarities between LADs that score highly in different complexity measures (based on the cosine similarity between their specialisation vectors). Figure 4 shows that average similarities between LADs with high scores in `bres_eci` and `idbr_eci` tend to be higher than the average similarities in LADs with high scores in `bres_fit_p` and `idbr_fit_p` .

This lends further support to the idea that LADs with high ECI scores are sectorally homogeneous, with a strong presence of unique, complex creative, digital and knowledge intensive sectors. Meanwhile, locations with high Fitness+ are more heterogeneous and

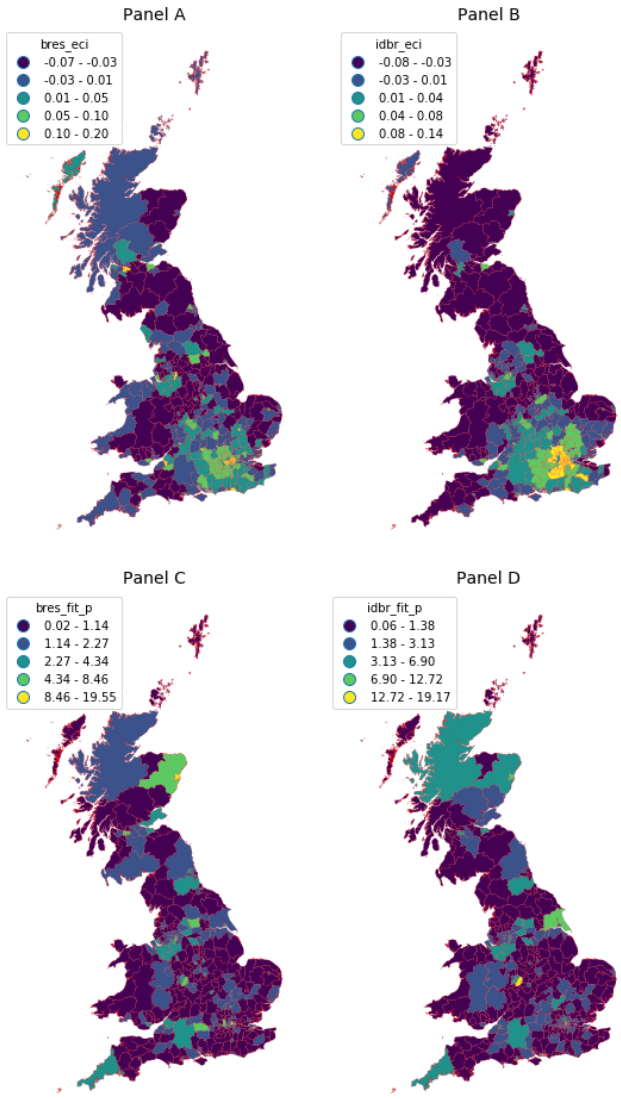

Figure 1: Economic complexity of British Local Authorities based on ECI (panels A and B) and Fitness+ (Panels C and D)

.

somewhat more diverse (although it is worth pointing out that their correlations with diversity metrics in figure 2 are not always strong - this is not surprising, given that Fitness+ up-weights the presence of unique sectors in a location, based on the difficulty of

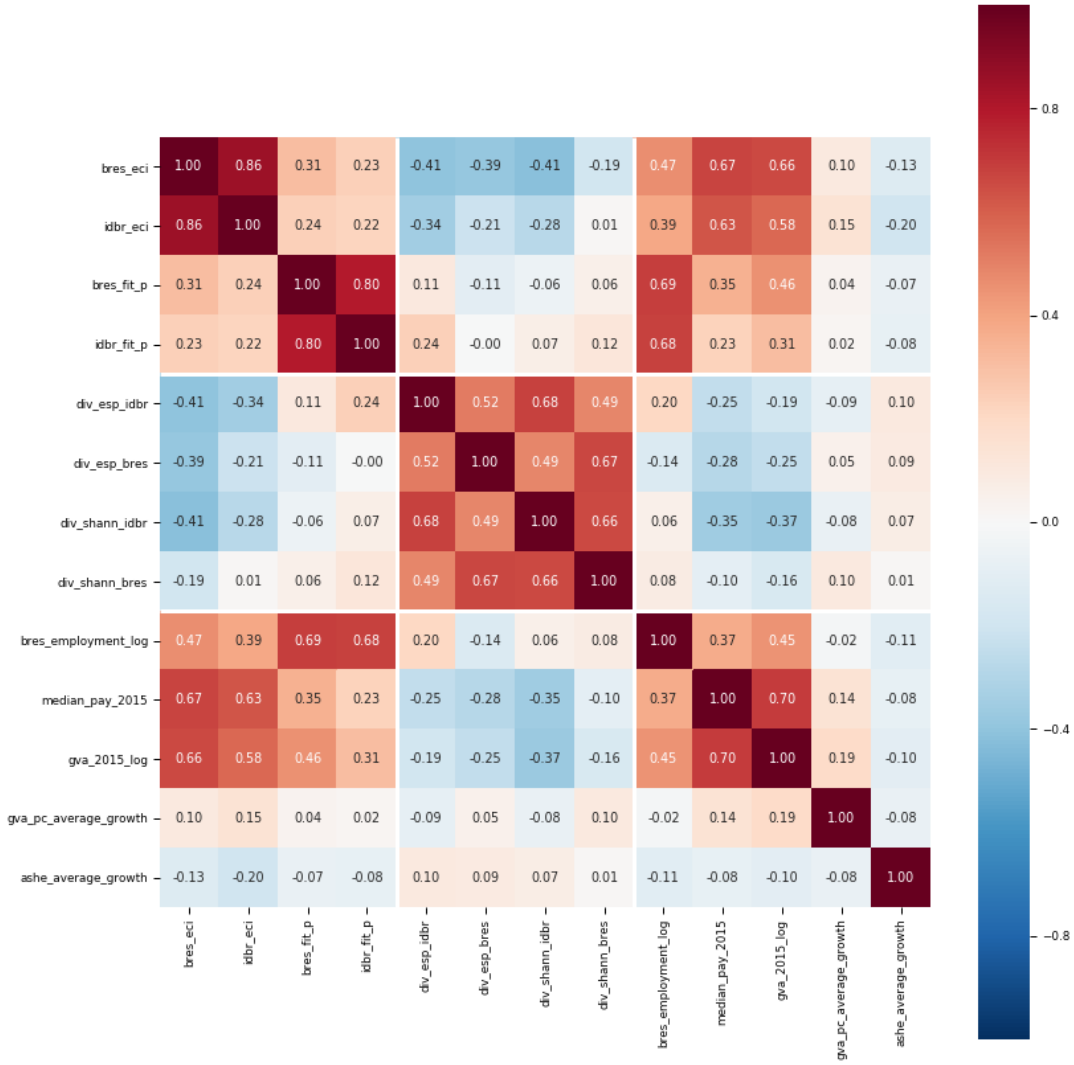

Figure 2: Correlation matrix for key variable sets (complexity, diversity and economic benchmarks)

developing a specialisation in them).

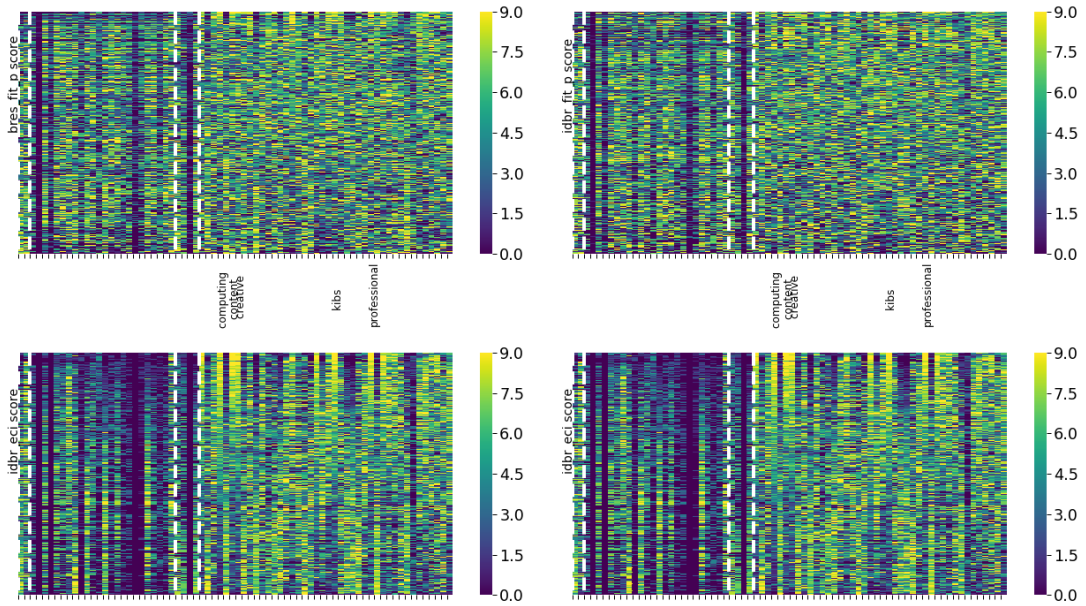

Figure 3: RCA in sector ranked by ECI index based on ECI algorithm (first row) and Fitness+ algorithm (second row). RCAs have been quantised into deciles to reduce the impact of extreme values. Vertical dashed lines indicate broad industrial categories (Construction, Manufacturing, Primary, and Services)

3.1.2 Economic complexity and economic performance

We conclude by considering the link between economic complexity and our economic benchmarks (median salary and GVA per capita (logged) in 2015, and average growth in those two variables between 2010 and 2016). In general, we find that the association between ECI metrics (`bres_eci` and `idbr_eci`) and the 2015 benchmarks is positive, and stronger for ECI than Fitness+. By contrast, the association between simpler metrics of diversity and economic performance is negative. This could reflect previous findings showing that the mere co-location of unrelated sectors is unlikely to be beneficial because agglomeration economies and spillovers require a shared knowledge base [8].

When looking at average growth in economic benchmarks, we note with interest that

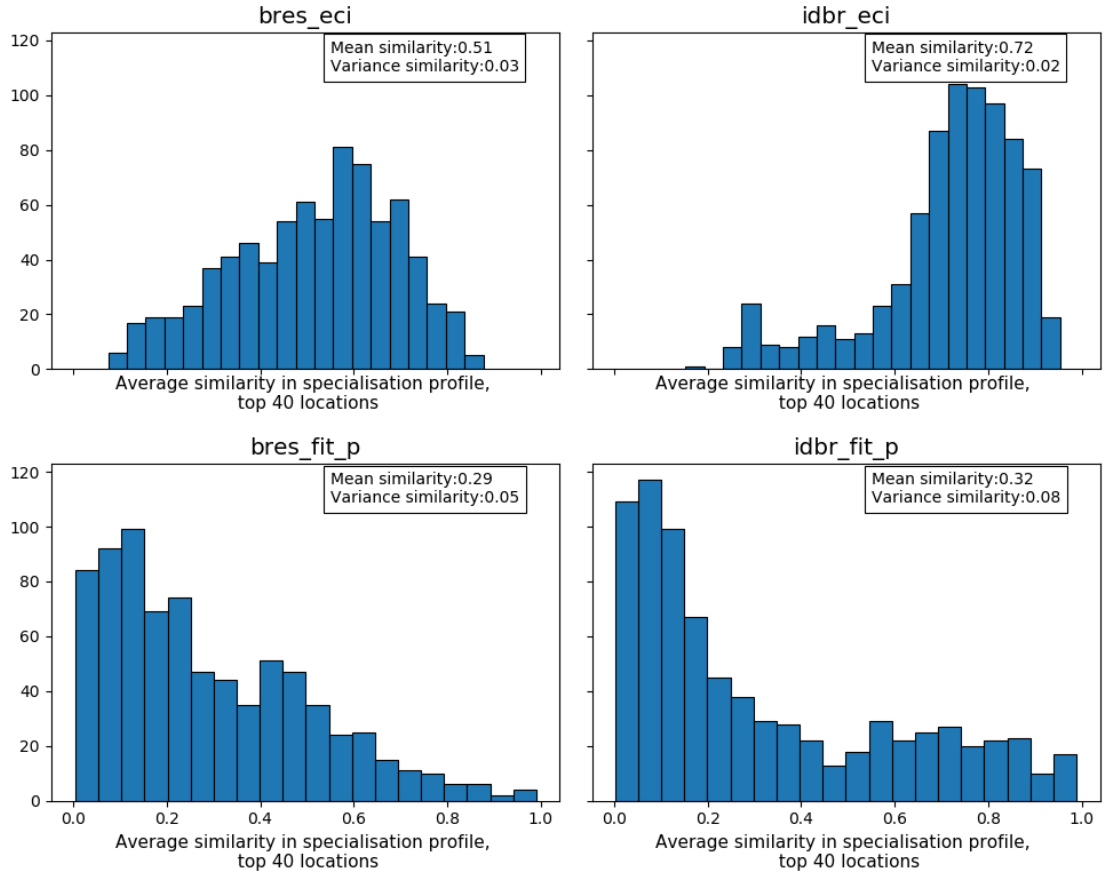

Figure 4: Histogram of pairwise sector similarities for top 40 locations in each complexity measure. Similarity is measured using cosine similarity.)

while the association between ECI measures and growth in GVA per capita is positive, their association with growth in median salary is negative. This result, which could reflect increasing income inequality in 'creative' cities in the UK [23] contrasts with the findings in [4] for China, and deserve further exploration.

3.1.3 Observations

LADs with high ECI scores tend to specialise in complex sectors but do not necessarily have a highly diversified local economy. High Fitness+ LADs also specialise in unique sectors but with some differences with ECI (for example, we detect a strong correlation between `primary_oil_gas` and `manufacture_light` RCAs and Fitness+ indexes.

One potential interpretation of these differences is that LADs that are part of dense urban conurbations (e.g. London boroughs in Greater London) can develop a strong specialisation in KIBS sectors because they can obtain other inputs and services from other LADs around them (hence their high ECI). Meanwhile, larger and more self-contained local economies like Birmingham and Leeds have a higher Fitness+. The strong correlation between total employment in a LAD and its Fitness+ supports this interpretation [10].

3.2 Complexity of language

Before proceeding to detect emergence within the descriptions of business websites in the Glass dataset, we investigate the link between language and economic complexity using topic modelling. Essentially by considering our documents (business descriptions) as being a mixture of topics with each topic itself being a mixture of words we can learn topics in an unsupervised manner from the likelihood of word co-occurrences within our documents. This approach enables us to analyse both the extent to which sectors and topics co-relate, and identify topics (and thus words) which correspond to high economic complexity. The most widely used and well known topic model is Latent Dirichlet Allocation (LDA) [24] which uses sparse Dirichlet priors on the number of topics per document and number of words per topic (assumes few topics per document and few words per topic). Given a topic

[10] This also creates the risk that some of our findings might be biased by systematic differences in the overlap between LAD boundaries and functional economic areas across the UK. Going forward, we will address this issue by reproducing our analysis at the TTWA level, and considering spatial autocorrelation in economic complexity metrics.

distribution per document and a word distribution per topic, the generative process of the LDA model generates the n^{th} word in a document by drawing a topic, z, from the multinomial distribution of topics, and then draws the word from the multinomial distribution of words for topic z. Recent work [25] has exploited a mathematical connection between topic models and finding community structure in networks, namely the mathematical equivalence between the Stochastic Block Model (SBM) and probabilistic Latent Semantic Indexing (pLSI), to develop an approach to topic modelling by deriving a non-parametric Bayesian parametrisation (topSBM) of pLSI adapted from a hierarchical SBM (hSBM) [26]. We use this new approach to topic modelling as it confers multiple advantages over LDA such as automatically selecting the number of topics and allowing for a more heterogeneous topic mixture than is permitted by a Dirichlet prior.

We process and tokenise a random sample of 100,000 Glass business descriptions and fit a topSBM topic model. The bipartite document-word network and corresponding topic hierarchy is visualised in figure 5 and illustrates the hierarchical nature of the method, we focus on the second coarsest level of the hierarchy which identifies 125 topics[11].

By grouping documents by their primary sector and averaging the contributions to each topic over the sector we can see the relationship between topics and sectors (Figure 6) and inspect popular terms by sector such as in table 2 which gives the top 5 words from the main topic for a handful of sectors.

Furthermore, one can link the topic distributions to various measures of complexity by matching a companies topic distribution to the complexity of the LAD to which it belongs, averaging for each LAD, and finding the topics that are over-represented in the top-n complexity LADs compared to the bottom-n complexity LADs. Table 3 (Table 4) gives the top 5 terms in each of the top 10 (4) topics that are over-represented in the most (least)

[11]The coarsest level identifies 15 topics which does not provide enough resolution for this analysis.

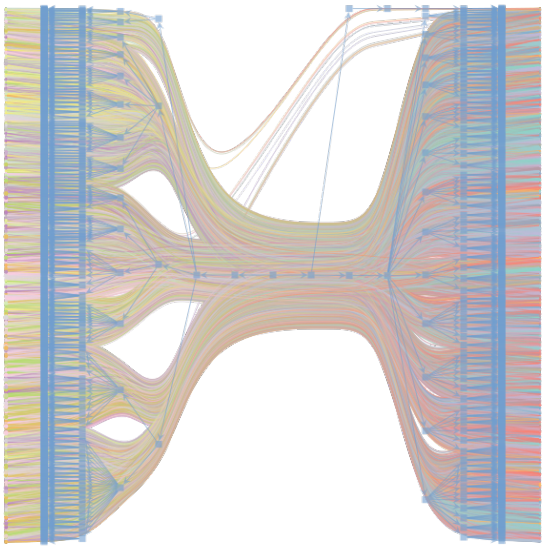

Figure 5: Figure showing the hierarchical community structure inferred by the SBM in the word-document network. Document (Word) nodes are on the left (right). On the uppermost level (middle), each node belongs to the same group. At the next-lower level the network is split into two groups corresponding to the bipartite nature of the network. On each next lower level nodes are further divided into word-groups (topics) or document-groups.

complex areas. The low complexity terms are generally uninteresting and seem to mostly correspond to community/family related terms. With the high complexity terms we observe that our four main measures of economic complexity all pick up similar terms (mostly relating to the information economy) as among the most complex but with the ECI terms tending to favour global and creative terms (possibly due to the fact that high ECI LADs are less isolated) whereas the Fitness+ place slightly more weight on business/management terms. Interestingly row 8 of the `bres_eci` measure contains the terms 'john', 'mark', and 'david'. It is unclear whether this illustrates the gender inequality present at the top echelons of industry (there are more CEO's named John or David than females in the S&P 1500) or whether the distribution of female names is less concentrated than that of Males.

	sector	top words
0	services_computing	software, data, platform, communications, infrastructure, analysis, enterprise, implement, powerful, implementation
1	services_r_&_d	research, change, campaign, science, target, profile, influence, targeted, scientific, transformation
2	services_kibs	aim, helping, future, partnership, committed, improve, enable, supported, potential, supporting
3	services_publishing	website, series, post, subject, news, read, book, story, voice, reading
4	services_agricultural	garden, accommodation, gardens, pet, animals, animal, dog, pets, yard, veterinary
5	manufacture_materials	furniture, kitchen, lighting, door, showroom, glass, fitting, doors, windows, stone
6	primary_fishing	food, natural, fresh, farm, coffee, agricultural, organic, foods, skin, farming

Table 2: Top 5 words from the topic with the heaviest weight in a given sector based on figure 6

	bres_eci	bres_fit_p	idbr_eci	idbr_fit_p
0	creative, marketing, agency, media, content	investment, growth, portfolio, finance, capital	investment, growth, portfolio, finance, capital	investment, growth, portfolio, finance, capital
1	software, data, platform, communications, infrastructure	software, data, platform, communications, infrastructure	creative, marketing, agency, media, content	software, data, platform, communications, infrastructure
2	investment, growth, portfolio, finance, capital	creative, marketing, agency, media, content	software, data, platform, communications, infrastructure	creative, marketing, agency, media, content
3	approach, understand, process, understanding, relationships	approach, understand, process, understanding, relationships	approach, understand, process, understanding, relationships	recruitment, talent, candidates, career, roles
4	unique, create, bring, creating, passionate	businesses, network, strategy, strategic, strategies	unique, create, bring, creating, passionate	approach, understand, process, understanding, relationships
5	opportunity, live, share, real, exciting	development, develop, partners, key, achieve	international, global, worldwide, countries, globe	financial, firm, firms, financial_services, authority
6	development, develop, partners, key, achieve	value, professionals, long_term, provider, successfully	recruitment, talent, candidates, career, roles	businesses, network, strategy, strategic, strategies
7	international, global, worldwide, countries, globe	expertise, enables, specific, ability, complex	development, develop, partners, key, achieve	value, professionals, long_term, provider, successfully
8	london, joined, john, mark, david	unique, create, bring, creating, passionate	focus, success, successful, existing, manage	international, global, worldwide, countries, globe

Table 3: Top 5 terms in each of the top 10 topics that are over-represented in the most complex areas for differing complexity measures.

	bres_eci	bres_fit_p	idbr_eci	idbr_fit_p
0	home, area, family, located, england	home, area, family, located, england	road, room, park, village, town	road, room, park, village, town
1	requirements, ltd, customer, customer_service, cost_effective	road, room, park, village, town	home, area, family, located, england	home, area, family, located, england
2	product, manufacturing, manufacturers, manufacture, manufacturer	enjoy, hope, welcome, families, fun	local, year, known, main, country	local, year, known, main, country
3	road, room, park, village, town	day, happy, visit, feel, location	requirements, ltd, customer, customer_service, cost_effective	enjoy, hope, welcome, families, fun
4	high_quality, pride, standard, family_run, highest_quality	local, year, known, main, country	enjoy, hope, welcome, families, fun	day, happy, visit, feel, location
5	complete, ensuring, required, allows, fully	good, come, free, regular, little	day, happy, visit, feel, location	requirements, ltd, customer, customer_service, cost_effective
6	standards, safety, contract, health_safety, approved	church, christian, god, nursery, churches	electrical, cleaning, air, waste, roof	high_quality, pride, standard, family_run, highest_quality
7	stock, items, accessories, packaging, goods	school, children, child, students, parents	standards, safety, contract, health_safety, approved	good, come, free, regular, little
8	components, automotive, oil_gas, electronic, electronics	garden, accommodation, gardens, pet, animals	national, awards, award, operates, expanded	available, great, like, look, sure
9	engineering, construction, engineers, industrial, project_management	learning, courses, teaching, pupils, primary_school	product, manufacturing, manufacturers, manufacture, manufacturer	product, manufacturing, manufacturers, manufacture, manufacturer

Table 4: Top 5 terms in each of the top 4 topics that are over-represented in the least complex areas for differing complexity measures.

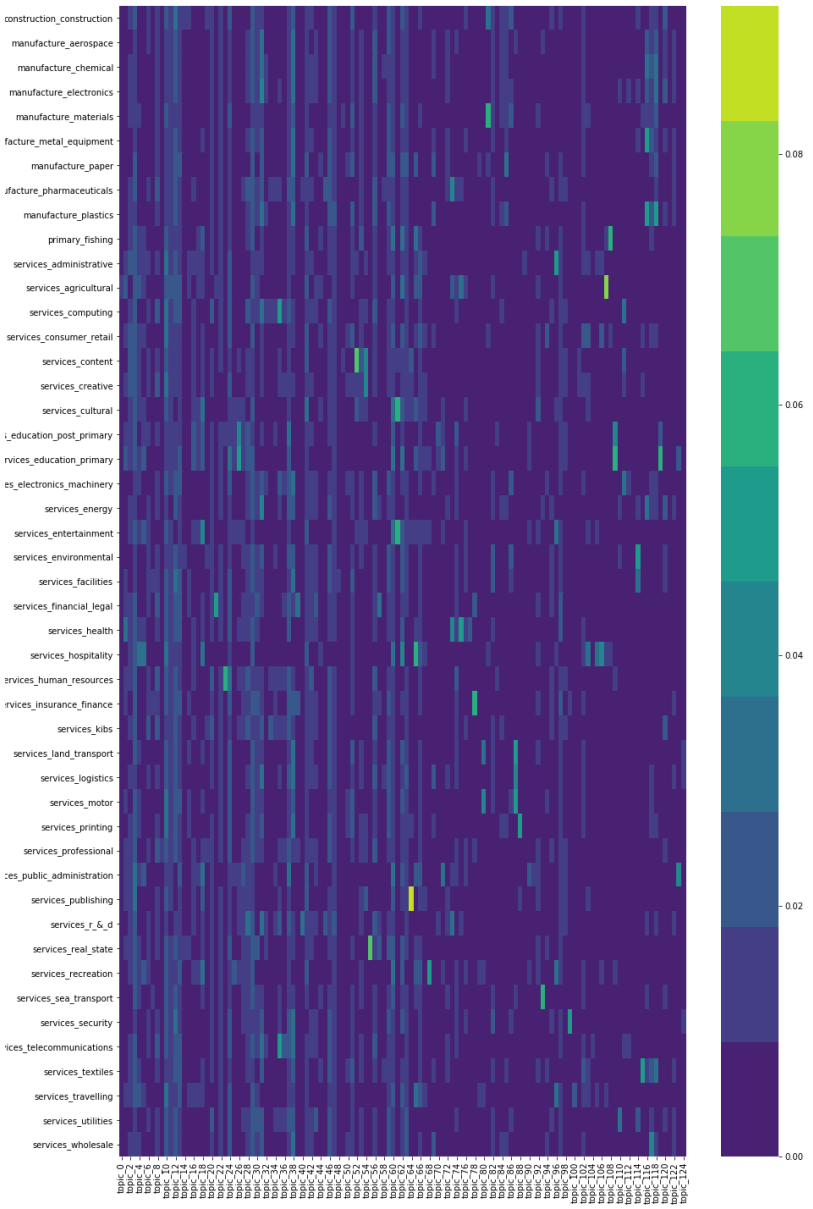

Figure 6: Heatmap illustrating the relationship between topics and sectors by aggregating the topic mixture of documents by sector.

3.3 Measuring emergence in web data

We use a set of keywords about emergent technologies extracted from CrunchBase [12] - a technology company directory) to identify relevant businesses in a cross-sectional snapshot of the Glass data.

3.4 Identifying emergence in CrunchBase data

One important dimension of emergence is technological: new technologies give rise to new industries, and to transformations in production processes and business models [10]. Based on our literature review, we would expect economically complex areas to generate higher levels of activity in emergent technologies.

Here we explore that question using a cross-sectional Glass dataset with company descriptions and LAD information for 353,307 unique business websites. Since these data lack a temporal dimension, we need an external source of information to identify novel technologies. To do this, we have used CrunchBase, a directory of technology companies increasingly popular in the analysis of start-up ecosystems [27, 28, 29]. More specifically, we have taken 238,629 companies in the CrunchBase directory, concatenated their pre-processed descriptions based on the year when they were founded, and identified salient terms that year based on their `TF-IDF` score, which normalises the occurrence of words in an observation (in this case a year) by their occurrence in the corpus.

As the results in table 5 show, this analysis captures different 'eras' in the digital technology landscape, starting with social networks and smartphones in 2008, moving into Artificial Intelligence since 2013, and then into Blockchain, crypto-currencies and virtual reality (VR) in the last three years.

We have decided to focus on salient terms from the last three years, and look for

[12] https://www.crunchbase.com/

Year	Top terms
2008	social network, online marketing, social networking, engine optimization, app development
2009	social network, online marketing, social networking, twitter, engine optimization
2010	social network, online marketing, platform allows, deal, social networks
2011	social network, smartphone, based platform, commerce platform, accelerator
2012	social network, application enables, smartphone, crowdfunding, accelerator
2013	accelerator, social network, crowdfunding, bitcoin, ai
2014	ai, peer, bitcoin, crowdfunding, accelerator
2015	ai, peer, vr, artificial intelligence, machine learning
2016	ai, blockchain, machine learning, vr, artificial intelligence
2017	blockchain, ai, cryptocurrency, artificial intelligence, decentralized
2018	blockchain, ai, crypto, cryptocurrency, decentralized

Table 5: Top 5 terms by TF-IDF score and year, CrunchBase

company descriptions with those terms in the Glass data. In order to avoid low recall due to small differences in spelling or variation in terminology to refer to the same industries or technologies, we have expanded the initial vocabulary of emergent technology trends extracted from CrunchBase with similar terms based on a `word2vec`[30] model trained on the Glass data [13].

This expands the initial set of terms to the following vocabulary:

accelerator, accelerators, ai, algorithmic, algorithms, animation, animations, ar, artificial_intelligence, augmented_reality, autonomous, bi, big_data, bioscience, biotech, biotechnology, bitcoin, blockchain, catapult, cleantech, cross platform, crowdfunding, crypto, cryptocurrencies, cryptocurrency, cyber, cybersecurity, 3d_animation, 3d_modelling, 3d_printing, 3d_scanning, data_analytics, data_science, data_visualization, decentralised, decentralized, digital_technologies, disruptive_technologies, drone, drones, early_adopters, emerging_markets, fintech, fuel_cell, gamification, highly_scalable, iaas, immersive, incubator, internet_of_things, iot, iptv, machine_learning, medtech, motion_graphics, ocu-

[13] `word2vec` projects words in a corpus into a multidimensional space based on their co-location. The vectors for words which are semantically similar are located closer to each other in this space.

lus_rift, paas, patent_pending, patented, patented_technology, predictive_analytics, remote_sensing, robotics, saas, saas_software, sensor_technology, sequencing, smart_cities, smart_grid, uav, unique_patented, virtual_reality, vr, wearables

In general, the terms in the list capture emerging technologies or related ideas and concepts [14].

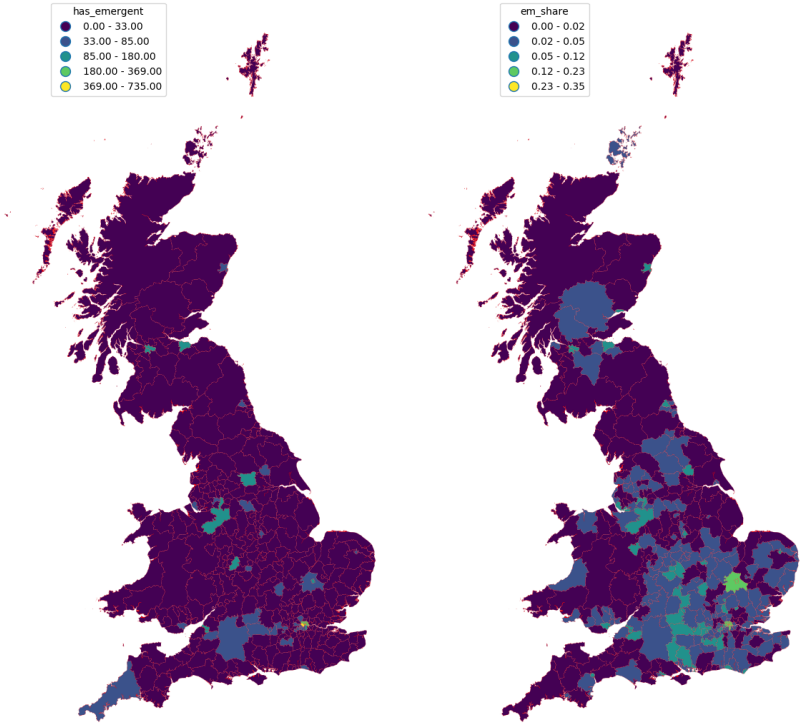

Figure 7: Emergent company activity by LAD: Company counts (left panel) and Share of all companies (right panel)

.

[14]Having said this, some of the terms in the expanded list are somewhat generic and could create false positives. We have randomly checked the results, removed those deemed likely to generate too many false-positives, and found that the majority of labels are valid. Having said this, refining this keyword search is an area that we plan to develop further.

Emergent companies in Glass' cross-sectional data

We have identified 5,652 companies mentioning at least one of the terms above. We map them in figure 7. Even after normalising by total business counts using IDBR data, most of the LADs with a strong presence of emergent companies are in London and the South East of the UK. Unsurprisingly, creative and digital technology clusters such as Cambridge, Oxford, Guildford or Bristol also rank highly in their 'emergence shares' (see table 6) [31].

Local Authority District	Emergent count	Emergent share (%)
City of London	534.0	0.35
Cambridge	110.0	0.23
Hackney	260.0	0.23
Islington	298.0	0.23
Tower Hamlets	222.0	0.19
Westminster	735.0	0.18
Camden	369.0	0.17
Southwark	180.0	0.15
South Cambridgeshire	80.0	0.13
Vale of White Horse	54.0	0.12
Oxford	53.0	0.11
Edinburgh, City of	174.0	0.11
Rushmoor	30.0	0.10
Guildford	64.0	0.10
Manchester	161.0	0.10
Bristol, City of	141.0	0.09
Milton Keynes	82.0	0.09
Hammersmith and Fulham	83.0	0.08
York	50.0	0.08
Reading	47.0	0.08

Table 6: Top 20 cities by share of emergent companies in Glass data

Figure 8 shows that knowledge intensive business sectors such as services_r&d, services_content, services_computing etc. have the highest share of emerging companies.

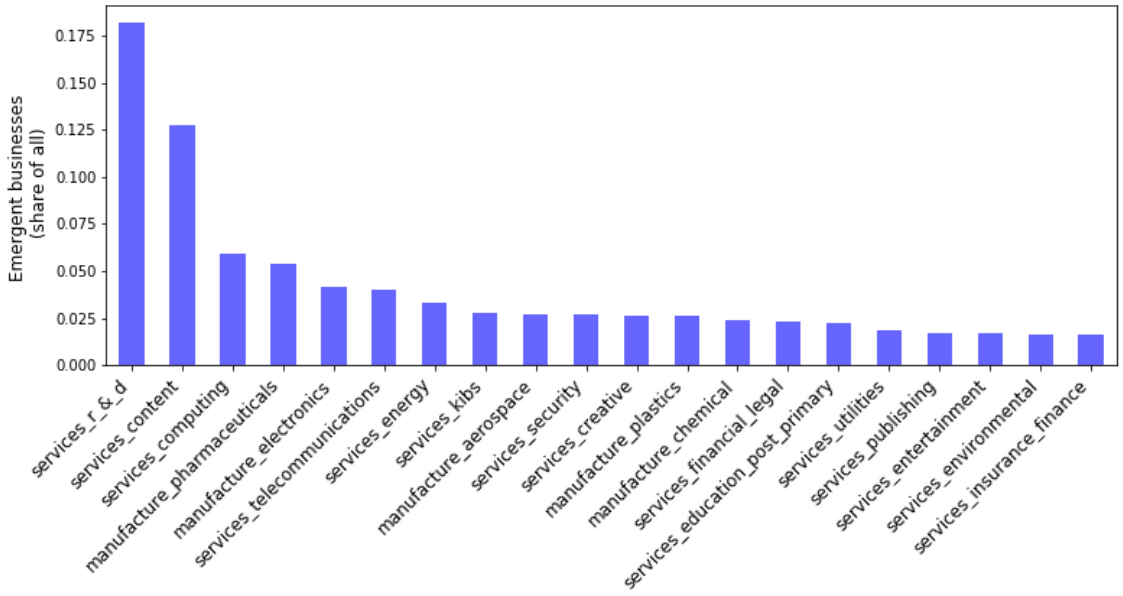

Figure 8: Share of emergent companies in total by sector (top 20 sectors)

.

ECI and emergence

What is the link between economic complexity and these measures of emergence?

In figure 9 we plot the ECI and Fitness+ scores in a LAD against the share of emergent companies in the business population. We find a strong association between a LAD's ECI and its share of emergent companies. The relation is weaker in the case of Fitness+.

Although this link is consistent with the idea that economically complex areas have a stronger propensity to facilitate emergence, this could be simply driven by the fact that as figure 8 shows, most emergent companies tend to operate in creative, digital and knowledge intensive business sectors that are strongly present in high ECI locations.

To control for differences in industrial composition between LADs, we compare ECI and Fitness+ with the share of emergent companies in a selected group of sectors in a LAD, including the top five industries by emergent share in the total, that is: services_r_&_d, ser-

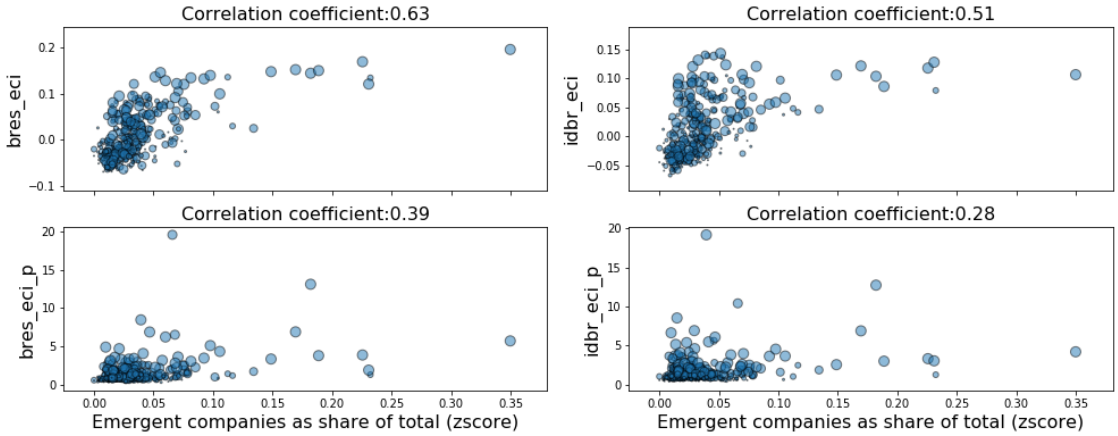

Figure 9: Correlation between share of emergent companies and ECI (first row) and Fitness+ (second row)

.

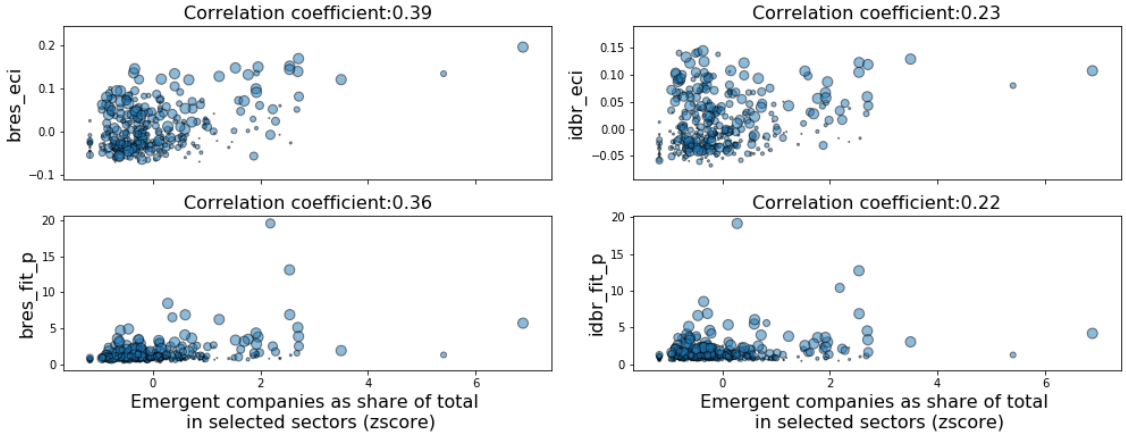

Figure 10: Correlation between share of emergent companies in top 5 sectors by emergence share (services_r_&_d, services_content, services_computing, manufacture_pharmaceuticals, manufacture_electronics) and ECI (first row) and Fitness+ (second row)

.

vices_content, services_computing, manufacture_pharmaceuticals, manufacture_electronics.

As the scatter-plots and correlation coefficients in figure 10 show, we still find a link between ECI and emergence.

ECI and diversified emergence

So far we have shown a link between ECI and emergence concentrated in selected knowledge intensive sectors. The association between Fitness+ and these emergence metrics is weaker. Now, we explore the second hypothesis set out in in section 1.4: that locations with higher Fitness+ scores will experience more *diversified emergence*: their emergent companies will be dispersed across a wider range of sectors.

In figure 11 we plot ECI against the Shannon entropy of the counts of emergent businesses by official sector in a LAD, a measure of the sectoral dispersion of its emergent companies. Although the association between this measure of emergence diversity and Fitness+ is somewhat higher than with the other metrics of emergence considered above, its relationship with ECI remains stronger: LADs with higher ECI tend to be more diversified in their emergent activities!

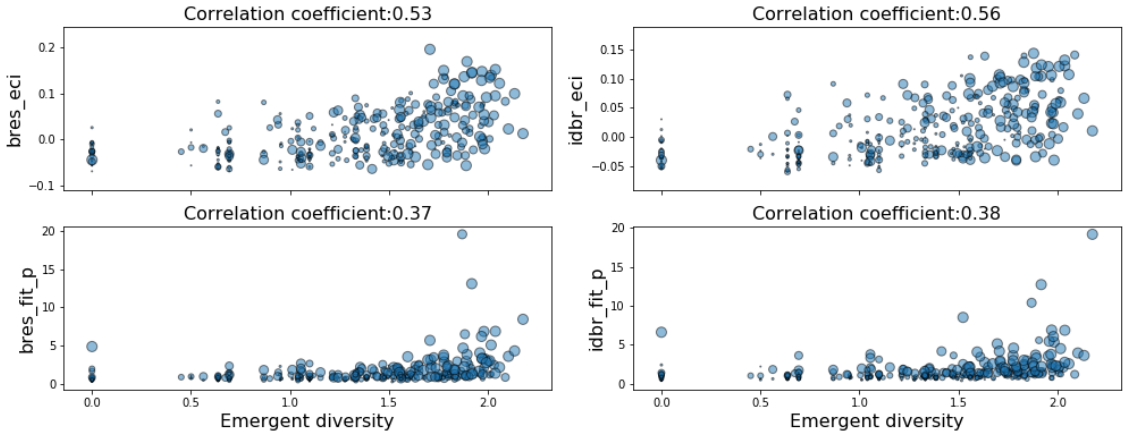

Figure 11: Correlation between emergent sector diversity and ECI (first row) and Fitness+ (second row)

.

4 Discussions

We have explored the economic complexity of UK local authorities using two indexes, ECI and Fitness+. Our results support the idea that high ECI scores capture specialisation in knowledge intensive, unique sectors such as computing, creative services, finance or research, while Fitness+ is closer to a weighted measure of diversity. LADs that are part of cities and conurbations in the South East of England (most prominently London) score higher in ECI, while large, more self-contained local economies in the Midlands, the North and Scotland score higher in Fitness+. ECI is more strongly associated with higher productivity measured in terms of GVA per capita and median salaries, although there is some evidence suggesting that it could also be linked to increasing inequality (measured through the evolution in median salaries). This idea is lent further credence by the topic modelling performed on the cross-sectional web data from Glass and linked to our indexes of economic complexity which suggests that large differences in ECI scores correspond to a marked change in the magnitude of certain topics.

Our exploration of cross-sectional web data from Glass highlights a potential mechanism linking ECI and higher productivity: locations with a high ECI score present a larger share of companies using new, emergent technologies underpinning new industries, business models and production processes. This relationship is not driven by differences in the sectoral composition of LADs: the share of emergent companies in digital, creative and knowledge intensive business sectors is also higher in high-ECI LADs. Contrary to what we expected initially, locations with higher Fitness+ scores do not tend to have higher levels of industrial diversification in their emergent activities. This could be partly explained by the focus on digital technologies in our definition of emergence. We would expect these technologies to be adopted first by the sectors that high ECI locations specialise on, and then to diffuse into other local industries. This underscores the value of knowledge intensive

business services for a local economy, not just as providers of innovative services, but also as lead adopters of new technologies that subsequently spread into the local economy.

5 Conclusion

Our work extends the analysis of economic complexity to local economies in the UK, showing the influence of their sectoral composition on growth prospects, and providing evidence about the mechanisms that link economic complexity and development through the emergence of new ideas. An avenue of research we do not pursue in this paper, though an important one, is the need for a more in depth analysis of how well the assumptions built into the development of complexity measures such as ECI and fitness hold on a sub-national level where boundaries are not as exclusive and less well defined. Our main policy implication is that local and national policymakers seeking to drive productivity growth in the UK will need to consider carefully specialisation profiles of its local economies, and also the broader regional contexts that enable specialisation in high growth, productivity-inducing sectors. As we have pointed out, boroughs in London and Local Authority Districts in dense, well connected parts of the UK such as the South East and the East seem able to specialise in the 'vanguard' sectors at the heart of economic complexity to a greater degree than larger, more self-contained and (perhaps) isolated cities in other parts of the country. At the same time, policymakers should recognise that a policy of specialisation in such industries might do little do address local economic inequality, and be detrimental for economic resilience against shocks concentrated on those sectors. The linking of novel data sources such as the Glass web data and official data make it possible to develop tools to aid policymakers in understanding the developing specialisation profiles of a local economy in a timely manner. Our analysis is not without limitations. This includes our reliance on LADs instead of functional economic areas as the unit of geographical analysis, the risk

of noise in our geocoding and sectoral classification of the Glass data, and our focus on descriptive, bivariate analyses of Glass' cross-sectional data. We will address these issues by reproducing our analysis with TTWAs as the unit of analysis, matching the Glass data with Companies House to obtain registered and trading addresses and sectors of activity for registered businesses, and modelling the link between economic complexity, emergence of new ideas and economic performance in a multivariate framework controlling for potential confounders such as population size, workforce education & skills, and considering the interdependencies between different metrics of economic complexity. Our analysis opens up many interesting questions for further research. They include a deeper consideration of spatial autocorrelation in economic complexity (and any interaction with the size effects observed with Fitness+), exploring the suggestive link between ECI and deepening economic inequality we highlighted in section 2.1, and using other data available from Glass as well as other open sources such as public funding for R&D to further explore the link between economic complexity and the emergence of new ideas in scientific and technological domains, as well as the extent and structure of collaboration networks that underpin these links.

References

[1] Cesar A. Hidalgo, Bailey Klinger, A.-L. Barabasi, and Ricardo Hausmann. The product space conditions the development of nations. *Science*, 317(5837):482–487, 2007.

[2] Cesar A. Hidalgo and Ricardo Hausmann. The building blocks of economic complexity. *Proceedings of the National Academy of Sciences*, 106(26):10570–10575, June 2009.

[3] Dominik Hartmann, Miguel R. Guevara, Cristian Jara-Figueroa, Manuel Aristaran, and Cesar A. Hidalgo. Linking economic complexity, institutions, and income inequality. *World Development*, 93:75–93, 2017.

[4] Jian Gao and Tao Zhou. Quantifying China's Regional Economic Complexity. *Physica A: Statistical Mechanics and its Applications*, 492:1591–1603, February 2018. arXiv: 1703.01292.

[5] Pierre-Alexandre Balland and David Rigby. The Geography of Complex Knowledge. *Economic Geography*, 93(1):1–23, January 2017.

[6] Inga Ivanova, Oivind Strand, Duncan Kushnir, and Loet Leydesdorff. Economic and Technological Complexity: A Model Study of Indicators of Knowledge-based Innovation Systems. *arXiv:1602.02348 [cs, q-fin]*, February 2016. arXiv: 1602.02348.

[7] César A Hidalgo, Pierre-Alexandre Balland, Ron Boschma, Mercedes Delgado, Maryann Feldman, Koen Frenken, and S Zhu. The principle of relatedness, 2017.

[8] Koen Frenken, Frank Van Oort, and Thijs Verburg. Related variety, unrelated variety and regional economic growth. *Regional studies*, 41(5):685–697, 2007.

[9] Cesar Hidalgo. Why information grows. *The evolution of Order, from Atoms to Economies.(Ebook) New York: Basic Books*, 2015.

[10] Daniele Rotolo, Diana Hicks, and Ben R Martin. What is an emerging technology? *Research Policy*, 44(10):1827–1843, 2015.

[11] David B Audretsch and Maryann P Feldman. R&d spillovers and the geography of innovation and production. *The American economic review*, 86(3):630–640, 1996.

[12] Jane Jacobs. *The economy of cities*. Vintage, 2016.

[13] Hasan Bakhshi and Juan Mateos-Garcia. New data for innovation policy. Working Paper, Nesta, London, 2016.

[14] Guido Caldarelli, Matthieu Cristelli, Andrea Gabrielli, Luciano Pietronero, Antonio Scala, and Andrea Tacchella. A network analysis of countries export flows: Firm grounds for the building blocks of the economy. *PLOS ONE*, 7(10):1–11, 10 2012.

[15] Greg Morrison, Sergey V Buldyrev, Michele Imbruno, Omar Alonso Doria Arrieta, Armando Rungi, Massimo Riccaboni, and Fabio Pammolli. On economic complexity and the fitness of nations. *Scientific Reports*, 7(1):15332, 2017.

[16] Eric Kemp-Benedict. An interpretation and critique of the method of reflections. 2014.

[17] Penny Mealy, J. Doyne Farmer, and Alexander Teytelboym. A New Interpretation of the Economic Complexity Index. *arXiv:1711.08245 [q-fin]*, November 2017. arXiv: 1711.08245.

[18] L. Pietronero, M. Cristelli, A. Gabrielli, D. Mazzilli, E. Pugliese, A. Tacchella, and A. Zaccaria. Economic Complexity: "Buttarla in caciara" vs a constructive approach. *ArXiv e-prints*, September 2017.

[19] Andrea Tacchella, Matthieu Cristelli, Guido Caldarelli, Andrea Gabrielli, and Luciano Pietronero. A new metrics for countries' fitness and products' complexity. *Scientific reports*, 2:723, 2012.

[20] Emanuele Pugliese, Andrea Zaccaria, and Luciano Pietronero. On the convergence of the fitness-complexity algorithm. *The European Physical Journal Special Topics*, 225(10):1893–1911, Oct 2016.

[21] Saleh Albeaik, Mary Kaltenberg, Mansour Alsaleh, and Cesar A. Hidalgo. Improving the Economic Complexity Index. *arXiv:1707.05826 [physics, q-fin]*, July 2017. arXiv: 1707.05826.

[22] Mercedes Delgado, Michael E Porter, and Scott Stern. Defining clusters of related industries. *Journal of Economic Geography*, 16(1):1–38, 2015.

[23] Centre for Cities. City monitor - Cities Outlook 2018 | Centre for Cities.

[24] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.
[25] Martin Gerlach, Tiago P. Peixoto, and Eduardo G. Altmann. A network approach to topic models. *Science Advances*, 4(7), 2018.
[26] Tiago P. Peixoto. Bayesian stochastic blockmodeling. abs/1705.10225, 2017.
[27] Jean-Michel Dalle, Matthijs den Besten, and Carlo Menon. Using crunchbase for economic and managerial research. 2017.
[28] Stefano Breschi, Julie Lassébie, and Carlo Menon. A portrait of innovative start-ups across countries. 2018.
[29] Max Nathan, Tom Kemeny, and Bader Almeer. Using crunchbase to explore innovative ecosystems in the us and uk. 2017.
[30] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.
[31] J Mateos-Garcia, J Klinger, and K Stathoulopoulos. creative nation: How the creative industries are powering the uks nations and regions. *London: Nesta*, 2018.