

Modeling clusters from the ground up: A web data approach

Christoph Stich

University of Birmingham, Birmingham, UK

Emmanouil Tranos 

University of Bristol, Bristol, UK; Alan Turing Institute, London, UK

Max Nathan 

University College London, London, UK; Centre for Economic Performance, London, UK

EPB: Urban Analytics and City Science
2022, Vol. 0(0) 1–24

© The Author(s) 2022



Article reuse guidelines:

sagepub.com/journals-permissions

DOI: 10.1177/23998083221108185

journals.sagepub.com/home/epb



Abstract

This paper proposes a new methodological framework to identify economic clusters over space and time. We employ a unique open source dataset of geolocated and archived business webpages and interrogate them using Natural Language Processing to build bottom-up classifications of economic activities. We validate our method on an iconic UK tech cluster – Shoreditch, East London. We benchmark our results against existing case studies and administrative data, replicating the main features of the cluster and providing fresh insights. As well as overcoming limitations in conventional industrial classification, our method addresses some of the spatial and temporal limitations of the clustering literature.

Keywords

clusters, cities, technology industry, machine learning

JEL codes

C55, L86, O31, R12

Introduction

Modelling economic activities in space is a key theme of geographical research. Clusters are most simply understood as physically co-located, interacting groups of firms (Marshall, 1890), but now there is a vast literature on cluster formation, characteristics and dynamics (Duranton, 2011; Uyarra and Ramlogan, 2013).¹

Despite this wealth of activity, key questions about clusters remain unresolved. First, we are still unclear about the relative salience of different cluster microfoundations, especially the balance

Corresponding author:

Emmanouil Tranos, School of Geographical Sciences, University of Bristol, Bristol BS8 1RL, UK.

Email: e.tranos@bristol.ac.uk

between industrial specialization and diversity (Cariagiu et al., 2016; Frenken et al., 2015; Kerr and Kominers, 2015). Frameworks for cluster evolution are still in debate, particularly the desirable level of analytical generalizability (Martin and Sunley, 2011; Neffke et al., 2011). As a result, the feasibility of cluster policy and the appropriate policy mix also remain unclear (Duranton and Kerr, 2015).

In part, these questions are hard to answer because of some hard-to-fix empirical challenges. Identifying and describing clusters remains extremely challenging. For example, clustering does not always take place at the scale of available data, and working at inappropriate scales can distort results (Modifiable Areal Unit Problem, or MAUP). Researchers have turned to geocoded plant-level data to tackle this (Baldwin et al., 2010; Neffke et al., 2011). However, the industrial classifications used in such ‘administrative big data’ are backward-looking and tend to lag real-world industrial evolution (OECD, 2013; Papagiannidis et al., 2018). Defining clusters based on industries constrains our understanding of emergent sectors such as fintech or cleantech, which sit across multiple industry bins (Li et al., 2018). Using web data allows to capture company self-descriptions (Nathan and Rosso, 2015). Third, there are tradeoffs between data richness and reach. Firm censuses ask limited questions, while online sources often require extensive validation. Conversely, the case studies and small-*n* surveys used in some evolutionary studies, while rich, have limited reach (Gök et al., 2015).

This paper makes two contributions to tackling these difficulties. First, we propose a novel methodological approach to analyze clusters over time, based on geolocated web data and data science methods. Our approach tackles several of the analytical challenges facing empirical cluster research, including MAUP, the industrial classification problem and the richness/reach tradeoff. It enables us to explore key concepts in the literature at scale, notably cluster evolution and emergent structures of economic activity. Second, we provide new empirical insights for a well-known UK tech cluster in London, only hitherto explored through a handful of case studies (Foord, 2013; Martins, 2015b; Nathan and Vandore, 2014; Nathan et al., 2019). The Shoreditch cluster also gives us an established ground truth (Pickles, 1995) and clear empirical priors on which to benchmark our approach. We also compare our results against administrative microdata from Companies House, the UK companies register, showing that our approach delivers insight over and above what is possible with more conventional data.

Our approach is motivated by recent developments in qualitative GIScience (Martin and Schuurman, 2020). We exploit a cache of archived and geolocated website data 2000–2012, the JISC UK Web Domain Dataset (JISC and the Internet Archive, 2013; Jackson, 2017). While in the public domain, this dataset, like other web archives, has been rarely used by geographers (Tranos et al., 2020). We work first at the level of activities. We allow a single firm to be active in multiple activities, as described in website metadata. We extensively clean and validate these raw data, focusing on websites which meaningfully represent economic activity on the ground. We use topic modelling to bundle activities in economic space, working both across the cluster and within modelled ‘verticals’. We apply this approach to Shoreditch and expose its industrial micro-geography by observing co-location of related activities at the postcode level; we explore cluster-level topics, their granular content, and their evolution over time; and we provide a detailed breakdown of ‘creative digital’ industry space. We reproduce several stylized facts, for example picking out the growth of creative digital activities and the uptick of activity after the introduction of the ‘Tech City’ cluster program. We capture the evolution of the different economic activities and processes of branching out of new and technologically related activities. The use of recent historical web data allows us to validate our approach against the ground truth. Our proposed methodological framework is transferable to different geographical contexts and timeframes given the growth of web archives, which can provide current web data (Summers, 2020).

Our framework illustrates the utility of qualitative spatial data derived from web archives and NLP to answer questions rooted within the core of geographical research. We contribute to an evolving literature which aims to expose the mechanisms of cluster formation, by moving beyond a pre-determined understanding of economic clusters in spatial, temporal and technological terms (Balland et al., 2015; Catini et al., 2015; Delgado et al., 2015; Ter Wal and Boschma, 2011). We also join a growing literature employing web data for answering economic geography research questions (Musso and Merletti, 2016; Papagiannidis et al., 2018).

Using web data to uncover business practices discusses how web data have helped to uncover business practices. *Data and methods* presents the data and methods. *Results* gives our results and *Conclusions* concludes.

Using web data to uncover business practices

Just like most economic activities, businesses leave digital traces that can be used to learn more about their behavior (Arribas-Bel, 2014; Rabari and Storper, 2014). One example is website data, which are readily available, cheap to obtain and extensive in terms of coverage. Most businesses maintain websites, which act as self-reporting platforms and include valuable information. Over 81% of firms with 10 or more employees had a website across OECD countries in 2018.² Coverage for smaller firms is only slightly less: in 2014 75% of all UK companies with at least one employee maintained a website (Gök et al., 2015). Business website text typically contains qualitative information on a variety of themes: from the types of economic activity and the firm outputs (products and services), to export orientation, research and development and innovation activities (Blazquez and Domenech, 2018a). Businesses may not necessarily expose all of their strategies on their websites, but neither do they do this for other conventional data collection methods (Arora et al., 2013). Importantly, the literature has identified a typology of business functions that such websites perform: they are designed to spread information and establish a public image for businesses, support online transactions and communicate with customers (Blazquez and Domenech, 2018a; Blazquez and Domenech, 2018b; Hernández et al., 2009). The quality of the web text is essential to achieve these objectives: “the firm must include on its website all the information it wants its real and potential clients to know, presenting it in the most adequate manner” (Hernández et al., 2009: 364). Among other things, the richness of web text also allows for potentially more flexible methods of industrial classification than conventional industry typologies (Papagiannidis et al., 2018). Crucially for our purposes, around 70% of all websites contain some place reference (Hill, 2009).

A handful of recent studies use web data and data science tools for industry and/or cluster analysis. Blazquez and Domenech (2018b) use data from corporate websites to test the export orientation of a small sample of 350 Spanish companies. They ‘nowcast’ and track important cluster features. Arora et al. (2013) and Shapira et al. (2016) study the early commercialization strategies of novel graphene technologies focusing on a sample of 65 small and medium-sized enterprises (SMEs) in the US, UK, and China. Gök et al. (2015) explore the R&D activities of 296 UK green goods SMEs and Li et al. (2018) focus on a similar sample of US-based SMEs to build a Triple Helix framework. Papagiannidis et al. (2015) use longitudinal archived web data to analyze the diffusion of different web technologies within and between specific sectors in the UK as well as across different mega-regions. Musso and Merletti (2016) and Hale et al. (2014) use these data to illustrate UK business’ web adoption in the late 1990s, and the linking practices of British university websites. Kinne and Axenbeck (2020) and Kinne and Resch (2018) in a large-scale study, scraped business websites to model firm innovative behaviour. The closest contribution to this paper is Papagiannidis et al. (2018), who retrieve the text and the metadata from the live websites of circa 8500 firms in the UK North-East, sampled from a market research database. They benchmark

classifications based on Standard Industrial Classification (SIC) codes against new classifications from web text, identifying clusters not shown by conventional typologies.

All these studies have important empirical limitations. Typically, only a few hundred subjects or less are covered, the temporal dimension is ignored, and the geolocation process is coarse at best. Conversely, we work with 12 years of data for thousands of business websites to explore cluster dynamics. We use postcode level information from self-reported trading addresses, rather than the registration addresses usually included in UK administrative data. Importantly, commercial or freely available firm data are not bias-free. Companies House, the UK's registrar of companies, does not include any information about business websites and only 24% of the records that [Papagiannidis et al. \(2018\)](#) used included business URLs.

Data and methods

We employ a unique source of archived web data, which have never been used before in such a context and extent: the JISC UK Web Domain Dataset ([JISC and the Internet Archive, 2013](#); [Tranos and Stich, 2020](#)). This is a bespoke subset of the Internet Archive (IA) and includes all the archived webpages under the .uk country code Top Level Domain (ccTLD),³ which is one of the oldest ccTLD created in 1985 ([Hope, 2017](#)) and was the second most popular in 1999 ([Zook, 2001](#)). Established in 1996, the IA is a non-profit organization that archives web content via a web crawler and a seed list of URLs. During the archival of the HTML documents from these URLs, it also discovers the hyperlinks included in these documents and uses them to discover more URLs following a snowball-like sampling technique ([Hale, Blank, and Alexander, 2017](#)). In 2016 the IA contained 273 billion webpages from 361 million websites ([Internet Archive, 2016](#)). While the IA continues its operation today, the preprocessing of their data by the British Library and, therefore, the time frame of this dataset ends in Q1 2013. Nevertheless, this dataset offers some rare advantages. As it is readily available it is more accessible to researchers outside digital humanities, which tend to monopolize the use of web archives ([Schroeder and Brügger, 2017](#)). We make use of this constraint to validate our approach against the ground truth of Shoreditch's recent history. Our results illustrate the potential of our approach for contemporary analysis and nowcasting applications using more recent web archives.⁴

We rely on archived web data instead of live ones because this is the only way to obtain longitudinal web data. Moreover, the publicly available business registration data in the UK (Companies House) do not include business website URLs and, importantly, the process of matching business names with websites is not trivial. This might be possible for other countries, the business registration data of which contain URL information – see for instance the work of [Kinne and Axenbeck \(2020\)](#) for German businesses.

Our raw data consists of billions of timestamped URLs of .uk webpages, which have been archived in 2000–2012. We access their text programmatically through the IA API.⁵ We use a subset of all the archived .uk webpages, which include a string in the format of a UK postcode (e.g. EC1A 1AA) in the web text. Created by the British Library, this dataset includes 2.5 billion URLs ([Jackson, 2017](#)). The postcode-based geolocation method does not suffer by the widely discussed IP geolocation limitations ([Zook, 2000](#)) and by the 'here and now' problem often occur with data derived from social media ([Crampton et al., 2013](#)). Both ideas refer to the mismatch between the location an activity takes place and its reflection in the different layers of the internet: while the former refers to the difference between the physical address included in a website registration, which is used for the geolocation of IP addresses, and the actual location of the underpinning activity, the latter points out the difference between the location and the time social media content refers to and how this propagates over space and time through different social media channels.

Such data are not without limitations as some websites might escape web crawlers. [Ainsworth et al. \(2011\)](#) find that 35–90% of webpages have been archived globally by public archives. The IA, just like any other archive, only captures publicly available webpages and is constrained by robot exclusions.⁶ Webpages that attract more traffic also have higher probability of being archived. Nevertheless, the consensus is that the IA is the most extensive and complete archive in the world ([Ainsworth et al., 2011](#); [Holzmann et al., 2016](#)). Focusing on a subset of websites similar to the one used here, [Thelwall and Vaughan \(2004\)](#) indicate that the IA captures at least one webpage for 92% of all the US commercial websites.

Data cleaning

We start with all the archived .uk webpages with a string in the UK postcode format in the web text. UK postcodes are alphanumeric strings with a hierarchical structure which refer to very small areas. For densely populated areas, they might refer to a single building. Hence, we treat them as point data. We trim data to 2000–2012, as the archived web data before 2000 is sparse and for 2013 we only have data for the first quarter. We drop false positives postcodes and keep webpages under the .co.uk or .ltd.uk second level domains, which represent commercial activities ([Thelwall, 2000](#)). A potential caveat here is that a UK company might decide to use a ccTLD different than the .uk one (e.g. .com). However, the established popularity of the .uk provides confidence for using these data to capture economic activities anchored in the UK and, more specifically, within Shoreditch: during the first year of our study period three .co.uk websites were registered every minute ([OECD, 2001](#)); and [Hope \(2017\)](#) illustrated the strong preference of UK consumers towards .uk websites when they are looking for services or products.

We then use these webpages to rebuild archived websites: for example, [www.website1.co.uk/webpage1](#) and [www.website1.co.uk/webpage2](#) are part of the [www.website1.co.uk](#). We further subset these data and only keep webpages with at least one postcode within the Shoreditch area. Following [Nathan et al. \(2019\)](#), we define the Shoreditch as a 1 km zone around Old Street Roundabout.

Websites do not necessarily correspond to underlying firms. Matching to company-level administrative data is both challenging and provides limited added value in this case, so instead we run diagnostics to understand website-firm relationships. Using the above example, if each archived webpage includes the same postcode, then we link [www.website1.co.uk](#) to a unique postcode. Otherwise, we sum all the unique postcodes included in the archived webpages of a specific website and this is the total number of different postcodes included in this website. We repeat this exercise yearly for the period 2000–2012. [Figure 1](#) presents this distribution.

Websites located at the right end of the long tail include many postcodes, at least one of which falls within Shoreditch. These are typically online directories, which were popular in the beginning of the study period ([Figure S3](#)). We drop such websites as they are artefacts of the internet's past and they do not represent economic activities anchored to the study area. Instead, we focus on commercial websites with a clear location within Shoreditch. To begin with, we only include in the analysis websites with one unique postcode, which fall within Shoreditch (18% of all the websites with at least one postcode in Shoreditch for 2000–2012). We argue that these websites represent economic activities that take place within our study area. As discussed in *Using web data to uncover business practices*, businesses are motivated to include accurate information in their websites in order to establish a public image and communicate with their customers, among other things ([Blazquez and Domenech, 2018a](#); [Hernández et al., 2009](#)). [Figure 2](#) illustrates examples of such websites. It presents the homepage of commercial websites with a unique postcode within Shoreditch, where usually the economic activity is presented, and the 'contact' page, where usually the Shoreditch postcode can be found. At a second stage we run a sensitivity check by running the

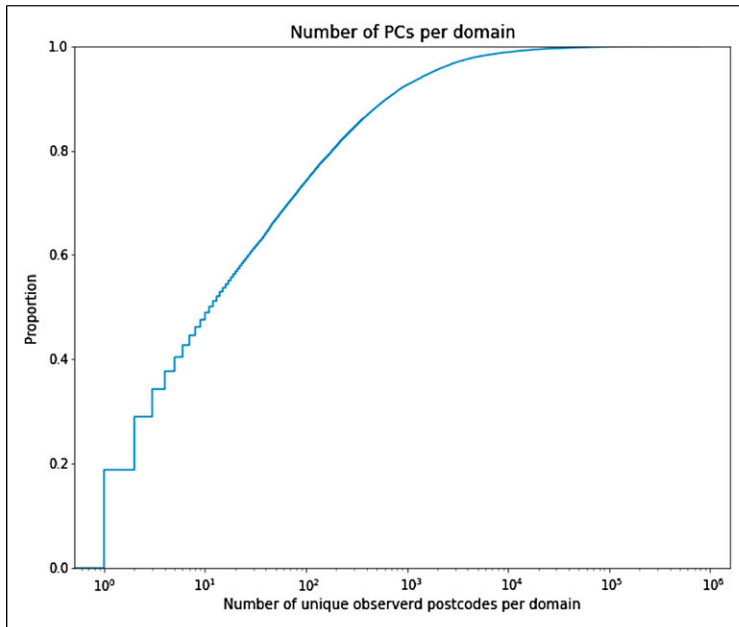


Figure 1. N. of postcodes per website distribution 2000–2012.

analysis to a larger sample that includes websites with up to 11 postcodes, at least one of which is in Shoreditch (50% of all the websites with at least one postcode in Shoreditch in 2000–2012). These sites plausibly represent economic activity in multiple locations, but may also represent generic economic activity less connected to the cluster.

We deal briefly here with two other concerns. Firms use websites in numerous ways, including defensive purposes akin to trademarking future products (Blazquez and Domenech, 2018a). Defunct firms’ websites may also live on after the underlying business has closed. However, the likelihood of having such websites in our data is small because the IA crawler finds, and archives websites based on hyperlinks from other websites leading to that website. We expect ‘placeholder’ or defunct websites to contain zero or very few live hyperlinks from other sites. Moreover, we would not expect defunct firms to continue paying domain names fees. Also, once a website is archived by the IA, chances are that this website will keep on being archived. Previous research has indicated that only 7.5% of the websites which contain at least one postcode appear in two or more years without these 2 years being consequent (Tranos and Stich, 2020). Lastly, the first year that a website appears in our data does not necessarily reflect the firm or the website creation year, but instead the first year the website was archived.

Topic modelling

To analyze the cleaned website text we use Latent Dirichlet allocation (LDA) and, specifically, an extension by Blei and Lafferty (2006), which accounts for the temporal evolution of the dataset. LDA is a widely used tool in natural language processing. Several studies have utilized LDA in spatial settings, such as the spatial distribution of topics on Twitter (Lansley and Longley, 2016; Martin and Schuurman, 2017), improving geographic information retrieval (Li et al., 2007), understanding residents’ views of their neighborhoods (Hu et al., 2019) or identifying classes of economic activities in a region (Papagiannidis et al., 2018).

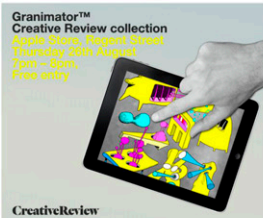
Home and contact us webpage snapshots	Source URL									
<p>1</p> <p>Welcome to ASMG</p> <p>ASMG specialise in developing tax efficient solutions with the aim of maximizing our clients wealth. We are a team of experienced tax professionals, our handcount being drawn from the Inland Revenue, "Big 4" and select tax oriented legal practices. We have implemented structures for successful private entrepreneurs and high net worth individuals.</p> <p>We pride ourselves on being more than just tax specialists. We exist to protect and enhance the wealth of our clients, and our services are specifically designed for that purpose. For regular, recurring support and on reaching business and personal crossroads, our people are there to help and to guide.</p> <p>ASMG people are committed, proactive and passionate about delivering the very best possible service and value to our clients. Our approach is based on building strong relationships to ensure a personal service. We provide a breadth and depth of skill to solve the most complex needs of our clients in a manner that is consistently second in commercial reality.</p> <ul style="list-style-type: none"> • Home • About us • Services • Careers • Testimonials • Enquiries • Contact us • Accessibility • Terms of use • Privacy policy <p>Visit this page Telephone: 020 79 430 1000 E-mail: info@asmg.co.uk</p> <p>Copyright ©2005, ASMG Ltd Site design by Blue Eye E-commerce Site uses XHTML and CSS and strives to be as accessible as possible.</p>	<p>http://web.archive.org/web/20060621095920/http://www.asmg.co.uk/80/</p>									
<p>2</p> <p>ustwo™ is a digital user interface company that develops pioneering user experiences and apps for some of the world's leading brands.</p> <p>ustwo™ and Creative Review to host a special event at the Apple Store</p>  <p>Granimator™ Creative Review collection Apple Store, Regent Street Thursday 26th August 7pm – 8pm, Free entry</p> <p>CreativeReview</p>	<p>http://web.archive.org/web/20100813113036/http://www.ustwo.co.uk/</p>									
<p>Get in touch</p> <p>If you've got some juicy work for us, have a great idea you want to share or just want to get to know us then we'd be happy to hear from you. If you are a recruitment agent then we'd love to hear from you too!</p> <table border="0"> <tr> <td data-bbox="447 1134 546 1285"> <p>London, UK</p> <p>+44 20 7611 9999 info@ustwo.co.uk</p> </td> <td data-bbox="559 1134 658 1285"> <p>Malmö, Sweden</p> <p>+46 40 333 9999 info@ustwo.se</p> </td> <td data-bbox="671 1134 769 1285"> <p>Business</p> <p>+44 20 7611 9999 info@ustwo.co.uk</p> </td> </tr> <tr> <td colspan="3"> <p>General</p> <p>+44 20 7611 9999 info@ustwo.co.uk</p> </td> </tr> <tr> <td colspan="3"> <p>Press</p> <p>+44 20 7611 9999 info@ustwo.co.uk</p> </td> </tr> </table> <p>Get in touch with us</p> <p>Get in touch with us</p>	<p>London, UK</p> <p>+44 20 7611 9999 info@ustwo.co.uk</p>	<p>Malmö, Sweden</p> <p>+46 40 333 9999 info@ustwo.se</p>	<p>Business</p> <p>+44 20 7611 9999 info@ustwo.co.uk</p>	<p>General</p> <p>+44 20 7611 9999 info@ustwo.co.uk</p>			<p>Press</p> <p>+44 20 7611 9999 info@ustwo.co.uk</p>			<p>http://web.archive.org/web/20100813113036/http://www.ustwo.co.uk/contact/</p>
<p>London, UK</p> <p>+44 20 7611 9999 info@ustwo.co.uk</p>	<p>Malmö, Sweden</p> <p>+46 40 333 9999 info@ustwo.se</p>	<p>Business</p> <p>+44 20 7611 9999 info@ustwo.co.uk</p>								
<p>General</p> <p>+44 20 7611 9999 info@ustwo.co.uk</p>										
<p>Press</p> <p>+44 20 7611 9999 info@ustwo.co.uk</p>										
<p>3</p> <p>DELICIOUS PHOTO FASHION, FOOD, LIFE, STYLE, TRAVEL WHAT, WHO, WHERE CLIENTS</p> <p>Delicious is a photo agency representing photographers who understand the recipe for compelling brands and create advertising images designed to stir emotions, quicken the pulse and make your mouth water. Delicious.</p> <p>Delicious Photo is a subsidiary of Delishious Group, London EC2A 3BE. © 2010 Delishious Group Ltd. All rights reserved.</p>	<p>http://web.archive.org/web/20100722203026/http://www.delicious-photo.co.uk/80/</p>									

Figure 2. Snapshots of examples of websites with a unique postcode in Shoreditch.

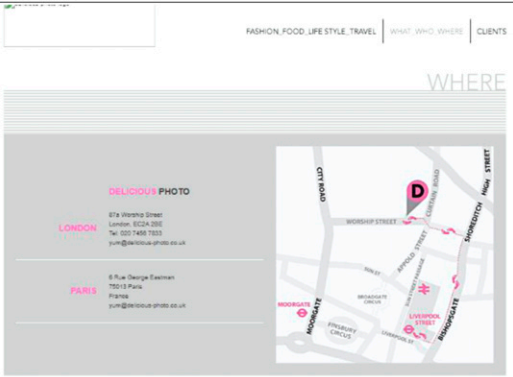
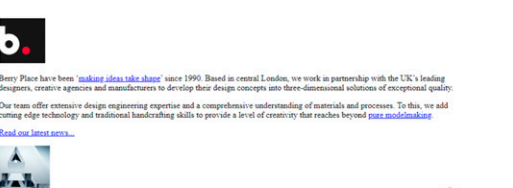
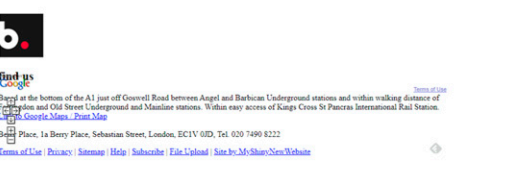
		http://web.archive.org/web/201100722025731/http://www.delicious-photo.co.uk/whow.html
4	 <p>Berry Place have been 'making ideas take shape' since 1990. Based in central London, we work in partnership with the UK's leading designers, creative agencies and manufacturers to develop their design concepts into three-dimensional solutions of exceptional quality.</p> <p>Our team offer extensive design engineering expertise and a comprehensive understanding of materials and processes. To this, we add cutting edge technology and traditional handcrafting skills to provide a level of creativity that reaches beyond mere modelmaking.</p> <p>Read our latest news...</p>	http://web.archive.org/web/201100706150800/http://www.berry-
	 <p>Berry Place is at the bottom of the A1 just off Goswell Road between Angel and Barbican Underground stations and within walking distance of Finsbury and Old Street Underground and Mainline stations. Within easy access of Kings Cross St Pancras International Rail Station.</p> <p>Berry Place Google Maps Print Map</p> <p>Berry Place, 1a Berry Place, Sebastian Street, London, EC1V 0JD, Tel: 020 7490 8222</p> <p>Terms of Use Privacy Sitemap Help Subscribe File Upload Sign Up My Sign Up New Website</p>	http://web.archive.org/web/201100706150749/http://www.berry-

Figure 2. Continued.

This approach has advantages over administrative datasets, which classify firms into industries using standardized typologies such as NAICS (in the US) or SIC/NACE (in the EU). Typically, firms are given only one code, where the underlying classification system may be several years old (in the case of current SIC/NACE, over a decade old). Here, we use website metadata to describe firms' economic activities ('terms') in the year of extraction and use LDA to bundle this into larger 'topics' which represent parts of activity space. This strategy means that each company can be part of several 'topics' at the same time, reflecting the fact that businesses can be active in several industries simultaneously. We combine topic and term-level information to identify specialized and cross-topic activities, such as the use of general-purpose technologies. Classification is also based on contemporaneous description by the firm itself. In the spirit of evolutionary economic geography, we then look at the growth and change of topics over time.

The intuition of LDA is that each website – or *document*, per NLP terminology – is composed by several different overlapping topics, which form the overall economic activity space. However, we cannot directly observe these topics, only the words that make up the documents. Formally, we assume that there is a generative process with hidden variables that defines a joint probability distribution for both the hidden and observed variables (Blei, 2012). LDA can then be described as finding a mixture of topics for each document:

$$P(t_i|d) = \sum_{j=1}^Z P(t_i|z_i = j)P(z_i = j|d) \quad (1)$$

where t are the terms of a document d , z_i is a latent topic and Z is the total number of latent topics (Krestel et al., 2009). To estimate the joint probability distribution, Blei et al. (2003) propose to use variational Bayes approximation of the posterior distribution. However, traditional LDA does not take the evolution of topics over time into account and topics are fixed over the whole study period. To overcome this problem, we adopt the approach of Blei and Lafferty (2006) to use probabilistic time series models to study the temporal dynamics of topics. This approach is widely used in the literature to study a variety of topics (Blei and Lafferty, 2006; Lee et al., 2016; Shalit et al., 2013) and allows topics to change between time slices, analogous to the branching process in cluster evolution (Boschma and Frenken, 2011).

We run the dynamic LDA on the human assigned keywords that describe the purpose of each website, to exclude extraneous vocabulary from our corpus. These keywords are part of HTML documents and are used from search engines to classify webpages.⁷ We follow standard NLP procedures to clean the keyword-based corpus. We exclude all English stop words and use the Snowball Stemmer (Porter, 2006) to only consider the word stems.

We use *genism* (Rehurek and Sojka, 2010) for modelling the dynamic LDA and *pyLDavis* (Sievert and Shirley, 2014) for visualizing the generated models. We set the maximum number of iterations to 300. To find an appropriate random seed for the topic modelling we create a population of 25 models with varying seeds. We then select the seed for our analysis that produces a model that is closest to the average of the log-likelihood of the population of models. Given that typical evaluation metrics such as coherence score are not available for the dynamic LDA, and that our research aim is to analyze the industrial structure of Shoreditch we opted for the highest number of topics up to the point that the derived topics could not have been interpreted by the authors. Hence, *Cluster-level analysis* presents the LDA outputs for $k = 15$ topics.

Solutions with less topics, which can be provided upon request, lead to similar conclusions when we look at the topic terms and more aggregated bundles of economic activities. As *Cluster-level analysis* illustrates, $k = 15$ led to fine-grained topics that were still interpretable and well delineated. This human judgement in selecting k is supported by the literature as previous work has shown that metrics based on the log-likelihood such as perplexity do often not agree with human judgement (Chang et al., 2009).

Robustness checks

We deploy different strategies to assess the robustness of our findings. Crucially, we implement our approach to model a well-known technology cluster in East London (Shoreditch). The theoretical and empirical stylized facts allow us to benchmark our results against established ground truth and previous literature. We also reproduce our cluster-level analysis using a larger set of websites containing up to 11 different postcodes, which represent larger multi-site firms, including chains. These may be economically important but less embedded in the cluster itself. Finally, we compare results derived from web-based methods with a more traditional approach based on administrative microdata from the UK company register (Companies House). This exercise illustrates how our research framework complements established analytical approaches in understanding clusters.

Our case study: Shoreditch

Shoreditch (known as ‘Tech City’) is a good test case, having much in common with urban technology production districts in large cities around the world (e.g. in New York, San Francisco, Berlin, Stockholm and Tel Aviv), including its evolution from ‘depressed’ ex-industrial area to ‘vibrant’ post-industrial milieu (Hall, 1998; Hutton, 2008; Scott, 1997, 2014; Zukin, 1982). Here we set out some stylized facts, drawing on existing qualitative and quantitative case studies, which form the ground truth that we want our framework to reproduce: beyond this, we want to deliver additional insights not uncovered by previous work.

The cluster is located in a set of ex-industrial East London neighborhoods a few miles from the West End and close to the City of London and is tightly drawn around the Old St roundabout (‘Silicon Roundabout’). Historically a working-class district organized around warehousing and light/craft manufacturing (including printing), Shoreditch declined in the post-WWII period. By the 1980s the area had large amounts of empty warehouse and office space. By the mid-1990s, these were taken up by a mix of artists (Harris, 2012), loft-dwellers (Hamnett, 2003) and (in the early 1990s) advertising, media and ‘new media’ firms moving east from more expensive central areas, followed shortly by a wave of dotcoms (Hutton, 2008; Pratt, 2009). This mixture of creative industries and technology firms has gradually evolved into the current ‘creative digital’ cluster (Foord, 2013; Nathan et al., 2019). Proximity to London’s main financial district gives the area a body of financial and business services firms, with several new office developments in recent years. The area has become a desirable residential neighborhood, with extensive new luxury apartment developments and local amenities for well-off incomers. At the same time, leisure and night-time economy has emerged, with many cafes, bars and restaurants doubling as ‘soft infrastructure’ where creative professionals meet (Currid, 2007; Martins, 2015b). Like similar clusters, the creative technology community grew ‘organically’ for many years before coming to the attention of policymakers (Foord, 2013; Jones, 2017; Nathan and Vandore, 2014; Pratt, 2009). The flagship ‘Tech City’ cluster development program was launched in 2010, and the cluster has become substantially larger and costlier in the following years (Nathan et al., 2019).

Results

Exploratory spatial analysis

Figure 3 presents the distribution of the number of websites per postcode in Shoreditch in 2012. Readers are reminded that postcodes in the UK are very small areas and for dense urban areas, they can even consist of a single building. Therefore, it is difficult to justify the extreme outlier at the right end of the distribution in Figure 3, according to which more than 80 unique websites point to a specific postcode in Shoreditch (EC1V 2NX). The Supplementary Materials section illustrates the interesting story behind this outlier, which has been removed.

Cluster-level analysis

We present here the LDA results for the 8154 commercial websites with one unique postcode within Shoreditch. One of the LDA parameters that needs to be exogenously defined is the number of topics. As mentioned in *Topic modelling*, because we aim to explore the industrial structure of Shoreditch we opted for the highest number of topics up to the point that the derived topics could not have been interpreted by the authors. Hence, Table 1 presents the LDA outputs for $k = 15$ topics.

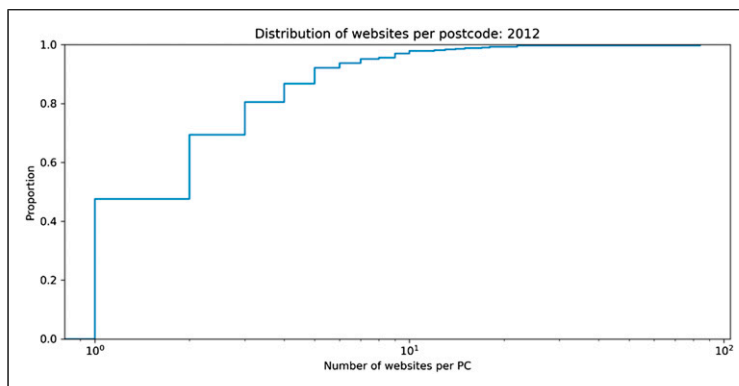


Figure 3. Number of websites per postcode in Shoreditch in 2012.

The last column of [Table 1](#) presents the 20 most frequent terms – that is stemmed website keywords – for each topic for the last year in the study period (2012). We use these terms to label each topic and their underlying term-level relationships ([Sievert and Shirley, 2014](#)). We rank these topics based on the overall frequency of their terms. Importantly, the topics correspond closely to the stylized facts about the cluster.

The digital and creative character of Shoreditch is clearly depicted in topics 1, 3, 8, 9, 12 and 14. Digital media is the most prevalent one (topic 1) and is a good representation of the area’s creative and media-orientated technology cluster, as illustrated in recent case studies ([Foord, 2013](#); [Jones, 2017](#); [Nathan and Vandore, 2014](#); [Nathan et al., 2019](#)). Its terms highlight economic activities related to online content creation and services, including roots in printing, graphics and ‘new media’: *design, web, websit, graphic, digit*. Other terms – *creativ, media, print, imag* – illustrate the area’s more recent creative core. A third group of terms covers the area’s digitized advertising and marketing activities, with terms such as *brand, advertis, and indet*.

Topics 3, 8, 9 and 12 depict the art scene of Shoreditch. The pre-WW2 craft tradition is reflected in topic 3 (*shop, jewelleri, accessori, furniture, bespoke, bag, make*). Music and performance arts are grouped in topic 8 (*music, event, record, show, club, danc*), while visual arts can be found in topic 9 (*design, art, photograph, architecture, architect, interior*). Topic 12 represents fashion related economic activities (*fashion, design, cloth, watch*). These LDA findings are in accordance with previous research and reflect past urban economic developments programs, which aimed to support creative industries including fashion, jewelry and furniture makers ([Foord, 2013](#)). Linked to the above is topic 14, which corresponds to the hospitality industry. This topic maps closely the typology of ancillary spaces for creative workers in Shoreditch uncovered in interviews by [Martins \(2015a\)](#): bar/pubs, coffee shops, restaurants, hotels, members’ club, parks, squares and street markets.

The second batch of topics are linked to business and financial activities. Topic 2 represents business services and finance as it includes terms such as *account, job, manag, compani, recruit, invest, and finance*. Financial and investment services are also present in topic 5 (*insur, compani, provid, loan, mortgag, onlin, credit, secur, broker*) and 6 (*trade, share, price, market, stock, money, exchang, financi, analysi*).

Topics 4, 7, 10 and 11 represent a bundle of advanced producer services, a key feature of global cities such as London ([Taylor et al., 2014](#)). For instance, we can identify business technology services (*system, servic, call, support, softwar, mobil, solut, network, phone, comput, applic, data, server, technolog*), consultancy agents (*consult, public, train, market, relat, communic, manag*,

Table 1. LDA topics.

Topic	Label	Term frequency (%)	20 most frequent terms
1	Digital media	13.7	design, web, brand, market, graphic, digit, websit, creativ, agenc, media, develop, product, advertis, onlin, print, site, consult, ident, imag, compani
2	Business services and finance	10.5	account, job, manag, compani, recruit, invest, servic, busi, financi, fund, tax, financ, advic, corpor, consult, market, bank, trust, pension, career
3	Craft	9.0	shop, onlin, jewelleri, game, theatr, product, store, fit, love, new, children, made, con, box, accessori, furnitur, bespok, order, bag, make
4	Business technology services	8.9	system, servic, manag, consult, call, support, softwar, mo-bil, solut, busi, network, phone, comput, applic, data, server, technolog, develop, number, cost
5	Financial services	6.8	servic, offic, busi, insur, compani, provid, loan, mortgag, onlin, credit, secur, centr, broker, commerci, mail, financ, unit, clean, cours, profession
6	Investment services	6.3	trade, share, price, market, stock, money, exchang, fi-nanci, offer, equiti, time, invest, day, rate, deal, inform, book, free, cash, analysi
7	Consultancy agents	6.2	consult, public, train, market, relat, communic, manag, re-search, agenc, strategi, develop, social, sector, educ, learn, project, health, cours, communiti, media
8	Music and performance arts	5.9	music, event, film, news, record, show, club, studio, parti, confer, danc, venu, entertain, sport, art, video, pop, rock, band, wed
9	Visual arts	5.9	design, art, photograph, architectur, architect, interior, photographi, galleri, east, space, artist, white, contemporari, exhibit, keyword, street, ferri, colour, bike, black
10	Legal services	5.1	hire, car, law, solicitor, legal, lawyer, citi, hotel, firm, ser- vic, room, discount, investig, clinic, commerci, litig, em-ploy, station, airport, great
11	Business support	5.0	servic, busi, print, name, onlin, sell, design, card, domain, recoveri, digit, work, colour, internet, build, host, net, printer, deliveri, document
12	Fashion and trade	4.8	fashion, design, cloth, beauti, gift, card, street, wholesal, best, women, place, watch, award, seal, univers, top, east, shop, old, organ
13	Real estate	4.5	home, properti, agent, sale, hous, holiday, let, buy, estat, manag, rent, real, develop, residenti, opportun, travel, hotel, flat, work, build
14	Hospitality industry	4.1	food, restaur, bar, book, cater, street, cours, citi, parti, translat, privat, servic, drink, wine, dentist, lunch, dine, corpor, take, languag
15	Wellbeing	3.4	therapi, citi, massag, injuri, treatment, back, sport, thera-pist, west, street, pain, central, ship, stress, care, south, get, well, cargo, hill

Note: terms are stemmed.

research, agenc, strategi), legal services (*law, solicitor, legal, lawyer, citi*) and broader business support (*servic, busi, print, name, onlin, sell, design, card, domain, recoveri, digit*).

Finally, the LDA revealed two topics linked to the urban nature of Shoreditch. Topic 13 reflects real estate (*home, properti, agent, sale, hous, holiday, let, buy*) and topic 15 wellbeing activities (*therapi, citi, massag, injuri, treatment, back, sport, therapist*).

Cluster evolution

Evolutionary frameworks highlight the way economic systems such as clusters ‘branch’ over time, with new industries emerging out of technologically related prior layers (Martin and Sunley, 2006; Neffke et al., 2011). Our framework can explore these temporal dynamics by looking at the topic prevalence (Figure 4) and within topics term frequency (Figure 5).

Again, our framework cleanly reproduces existing stylized facts (Cushman and Wakefield, 2013; Harris, 2012; Nathan et al., 2019). Digital media (topic 1) is the most prevalent topic with a brief exception during the post-dotcom crash period (2003–2005, Figure 4). It has an overall positive trend and its difference with the other topics increases over time. At the end of the study period, digital media is undoubtedly the dominant topic of the business websites geolocated to Shoreditch. Importantly, 2010 is the year of the launch of the East London Tech City programme, which aimed to ‘accelerate’ the cluster (Foord, 2013). In line with other evidence (Nathan et al., 2019), we observe an increase of digital activities a year after the policy intervention.

Business services and finance activities (topic 2) appear to have a competitive relationship with Topic 1 (digital media) as whenever the prevalence of topic 1 increases, the prevalence of topic 2 decreases and vice versa. Moreover, the prevalence of business technology services (topic 4) overcame topic 2 in 2010, consistent with digital technologies gradually shifting the industrial base of Shoreditch and leading to new and related economic activities, a process reflecting branching and recombination of knowledge within economic clusters (Boschma and Frenken, 2011; Boschma and Iammarino, 2009).

Economic activities linked to craft (topic 3) were decreasing in prevalence until 2006 and since then their importance steadily increases illustrating the resurgence of the crafts and art industries (Foord, 2013). A steady but small increase can be observed for fashion and trade (topic 12), which

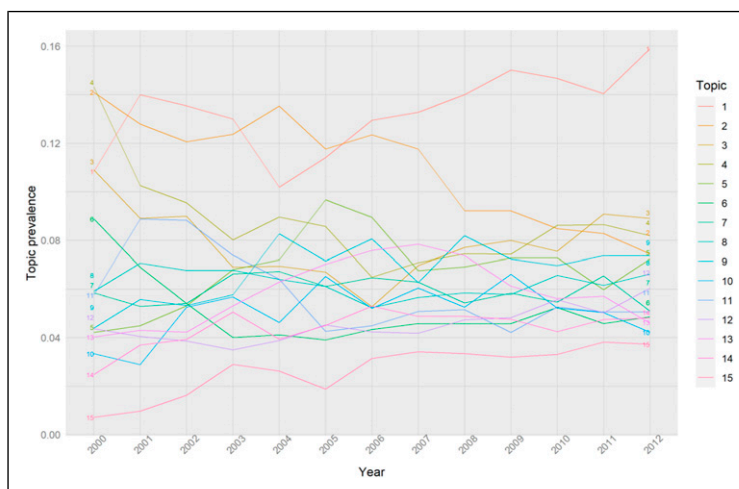


Figure 4. Topic prevalence over time.

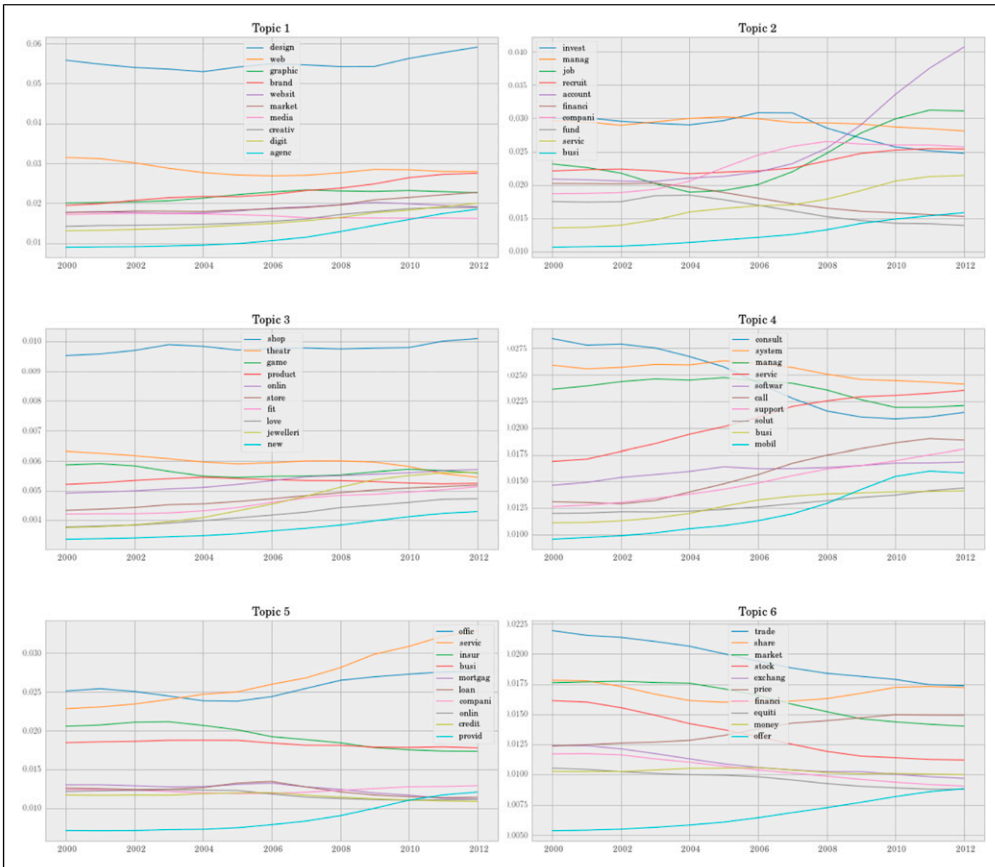


Figure 5. Dynamic term frequency per topic.

can be linked to publicly funded initiatives to support creative sectors such as the 2003–2009 City Growth Programme (Bagwell, 2008).

Figure 5 presents the within topics term frequency to assess how the consistency of topic changes over time. Starting from the digital media topic (topic 1), the term frequency remains stable. The main message is the consistent difference between the two most frequent terms – *design* and *web*. Design was and remained throughout the study period an integral characteristic of the economic activities clustered in Shoreditch. Similar observations can be made for the other related topics. *Shop* is the most frequent term for topic 3 throughout the study period reflecting the retail nature of the economic activities reflected in the craft topic.

Similarly, *music* and *design* are the dominant terms for music and performance arts (topic 8) and visual arts (topic 9). Regarding the fashion and trade topic (topic 12) the difference between *fashion* and *design* steadily increases highlighting the rising role that fashion plays for Shoreditch (Bagwell, 2008; Foord, 2013).

Contrary to the topics linked to digital and creative activities, business and financial activities topics are not as stable during the study period. The frequency of terms like *invest*, *finance* and *fund* drop after the 2008 financial crisis for topic 2 (business services and finance). Similarly, the frequency of terms including *trade* and *stock* decrease over time in topic 6 (investment services), while terms such as *price* and *offer* appear more frequently at the end of the study period. Within

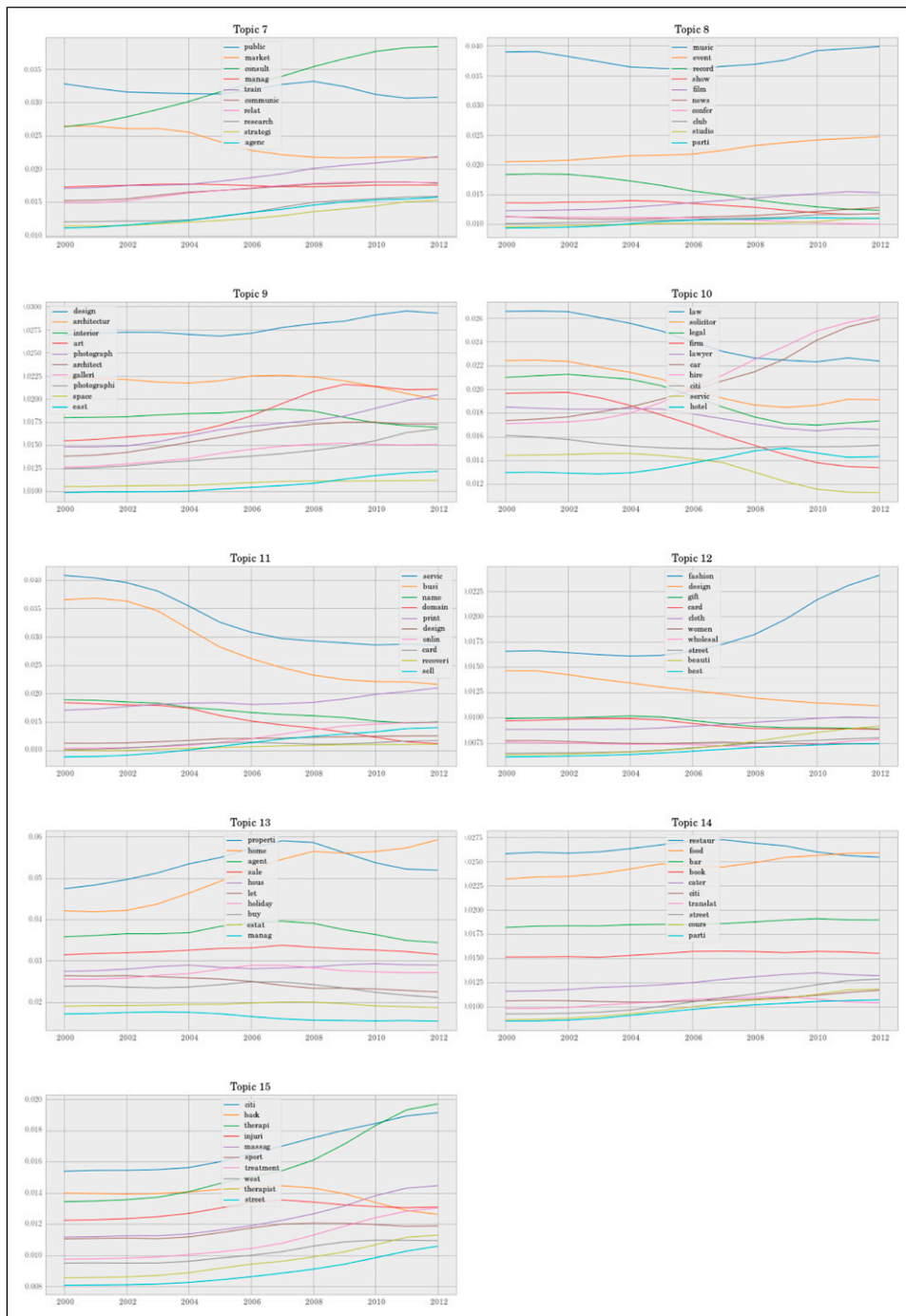


Figure 5. Continued.

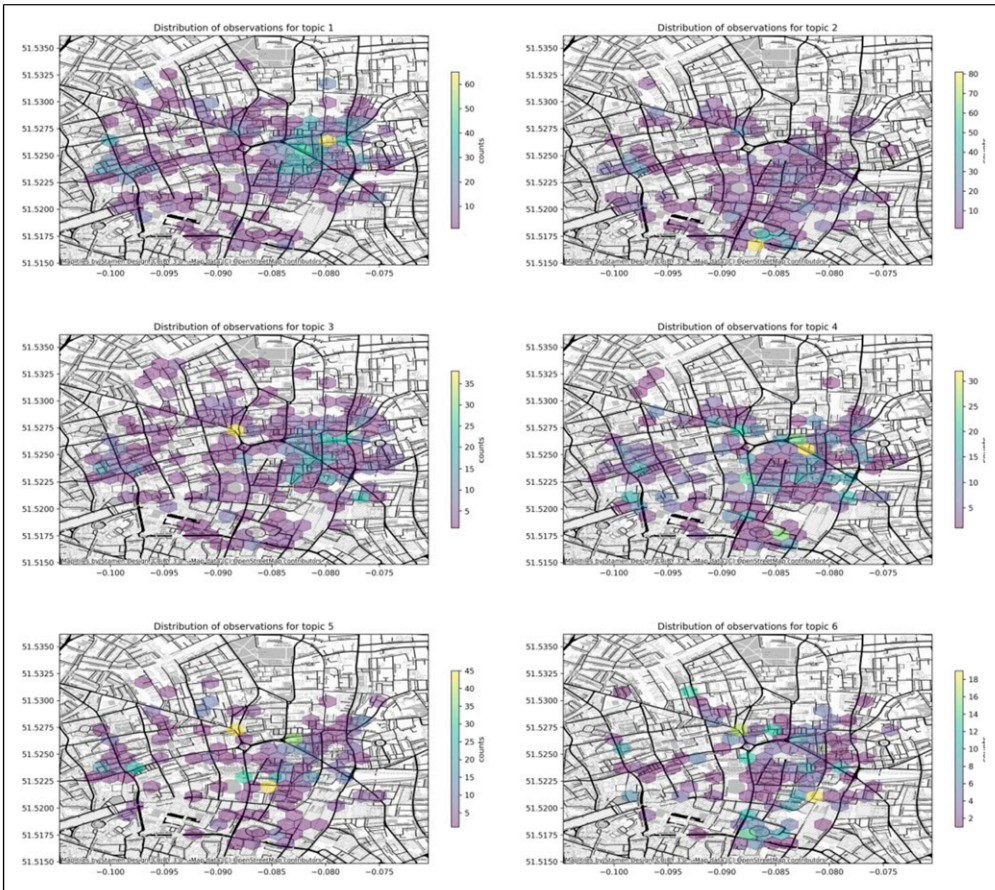


Figure 6. The spatial footprint of the different topics.

topic 4 (business technology services) the frequency of terms such as *servic*, *call*, *support* and *mobil* increases. The topic with the most changes is the one referring to legal services (topic 10). While terms such as *law*, *legal*, *solicitor* and *firm* decrease overtime, the frequency of *car* and *hire* increase.

Interestingly, we see the digital and technology terms associated with topic 1 appearing in other topics with greater frequency over time. We observe the growth of term *onlin* in topic 3 (craft) and 11 (business support), and *softwar* and *mobil* in topic 4 (business technology services), which is consistent with both the overall growth of digital technologies during the study period, and the technological diffusion within Shoreditch, from the dominant economic activities reflected in topic 1 (digital media) to other economic activities.

Our framework highlighted the well-established nature of digital and creative activities rooted in Shoreditch and the more volatile character of business and financial activities, which are present in Shoreditch, but as the next section highlights are spatially linked to adjacent areas. We were able to observe the evolution of economic activities illustrating processes of branching and, to a lesser extent, technological diffusion. Moreover, we associated changes in the prevalence of specific topics with place-based policies during the study period.

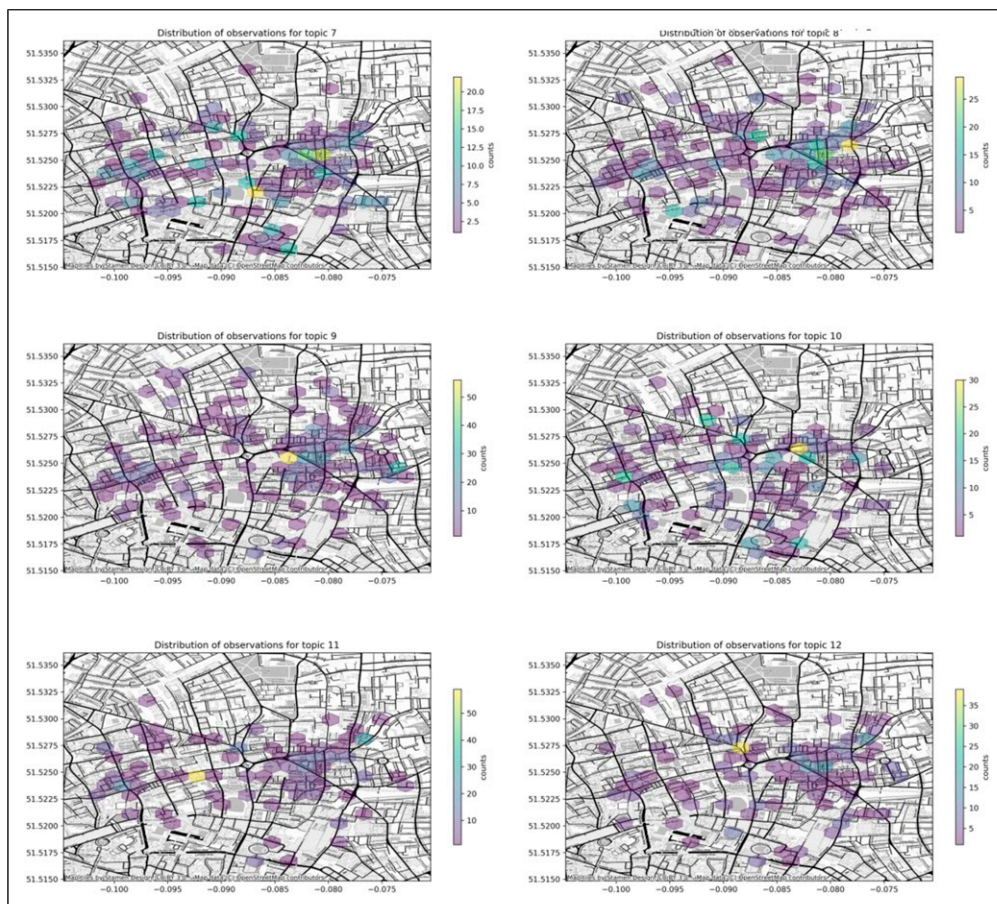


Figure 6. Continued.

Cluster footprint

The heatmaps of the websites assigned to the different topics derived from the dynamic LDA model (Figure 6) expose the spatial structure of the different economic activities. Interestingly, the topics linked to the digital and creative character of Shoreditch (1, 3, 8, 9, 12 and 14) are anchored to the west and north of the Old Street roundabout, which appears in the center of the maps. We also observe some less intense concentrations in the south part of the study area linked to art, fashion and music (e.g. topic 8) as this is the area where the Barbican, a large arts center is located. Topic 14 (hospitality) has the same epicenter as the digital media topic reflecting again how interwoven these topics are. It captures all the study area just like consultancy agents and wellbeing activities (topics 7 and 15). On the contrary, business services and finance and investment services (topics 2 and 6) gravitate towards the City of London, a world-leading financial cluster. Altogether, although the maps clearly indicate two distinct poles in the study area – that is the more creative northwest quarter and the more finance focused south area which is adjacent to the City of London – they also exemplify the spatial mixing of different activities which synthesize the Shoreditch’s identity.

The above draws a detailed picture of the types of economic activities that are present in Shoreditch. Our analysis, which is based on freely available archived web data and data science

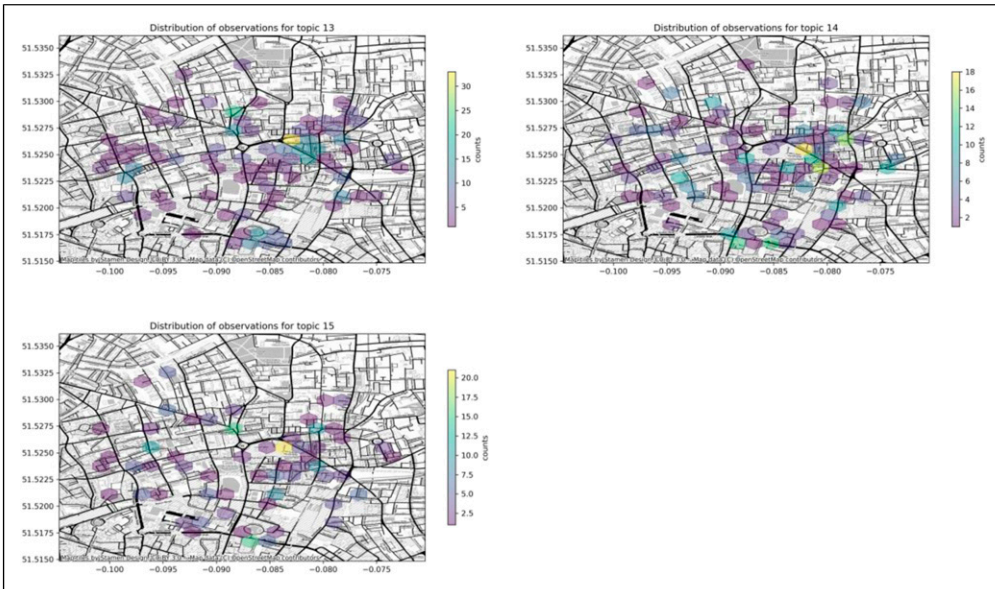


Figure 6. Continued.

methods confirms the results from previous studies, which were based on extensive interviews and fieldwork (Martins, 2015; Nathan et al., 2019), web inquiries on a pre-defined small sample of firms (Taylor et al., 2014), or secondary data analysis from propriety data providers (Foord, 2013). In addition, our approach enables to identify the evolution of these activities over time and provide a more in-depth analysis of the types of the economic activities that have been clustering and growing in Shoreditch.

In the [supplementary material](#) we provide two important extensions, specifically (i) a robustness check using an extended sample of archived, commercial websites linked to Shoreditch, and (ii) comparing the depth of analysis that our proposed research framework can achieve against the use of Companies House administrative business records. The first exercise confirms that our findings can also be replicated when using a much larger and spatially extended subset. The second shows that our approach reveals more insights about the economic activities of the study area than using administrative data, which tend to be the mainstream for such research and policy-oriented analysis.

Conclusions

Clusters, their formation and evolution are central issues in geography and urban science. Nevertheless, modelling clusters and their dynamics faces some hard-to-solve empirical challenges. This paper introduces a novel approach for analyzing and modeling clusters using public web data and data science methods. Our powerful and flexible approach, which is aligned with developments in qualitative GIScience, enables us to directly tackle some of these empirical challenges and implement many key theoretical concepts in cluster research, including within-cluster co-location patterns, local distinctiveness, related/unrelated variety of activity, and cluster evolution. We use this approach to analyze a well-known tech cluster in London, reproducing key stylized facts and generating new insights. We show that this approach is significantly more informative than next-best analysis using open administrative data. Our approach has multiple potential applications, not only for re-analyzing existing clusters, but also in detecting unknown or emerging cluster formations.

The use of unstructured textual data from the web enables us to move beyond the rigid SIC- based understanding of the activity space. Business websites typically accurately describe business outputs (Blazquez and Domenech, 2018a; Hernández et al., 2009). Using website metadata – HTML keywords which aim to accurately represent the activities behind a website in a concise manner – we depict the economic activities and their evolution in Shoreditch at a level of detail akin to the ones produced by qualitative studies based on lengthy participant observation and interviews, and greater than the one we obtained when we employed widely used administrative data. Despite the richness of our results, our methods and data are transferable to different spatial and temporal contexts given the current broad availability of web archives combined with tools and the computational capacity to analyze big volumes of textual data. Also, the spatial granularity of our data allow to overcome MAUP linked to the availability of only aggregated data about economic activities. Moreover, instead of focusing on firm registration addresses – which is a common fallacy of business administration data – the web data enables us to better approximate actual trading locations.

Our empirical findings are linked to key theoretical discussion within the cluster literature. Regarding the MAR/Jacobs debate, our analysis clearly indicates the role of specialization (digital content creation), but we also find evidence regarding the importance of diversity including the spillovers from the City of London and the importance of related ancillary activities. Despite the potential footloose nature of digital activities, co-location remains important for these firms, including tight co-location patterns *within* cluster space. From an evolutionary perspective, our analysis illustrates how the digital content activities have become dominant in the area, and how this specialization has led to the creation of new related economic activities. Although our aim is not to assess related urban policies, we observe a correspondence between the establishment of the Tech City programme and digital economic activities becoming dominant in Shoreditch.

The research framework proposed here is transferable to other clusters, for which we do not have enough data to study their evolution and specialization. It can also provide the basis for building algorithms to detect cluster formation on a near real-time manner and, therefore, directly support urban policy makers. The above exemplify the need to enrich the economic geography methodological toolkit with methods outside its traditional core including, among others, NLP which enables researchers to extract meaningful knowledge about places, their economic activities and relations utilizing the vast amounts of textual web data, which are currently unexplored.

Acknowledgements

We are thankful to the participants of the following conferences: Building Better Methods in Economic Geography session in AAG 2019, ERSA 2019, RSAI-BIS 2019 and GfR2020. Thanks to the British Library for curation of the JISC UK Web Domain Archive and to OpenCorporates for Companies House archive data. Thanks also to the anonymous reviewers and the editor for their valueable comments and suggestions.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The authors acknowledge funding from the Consumer Data Research Centre (CDRC) and Engineering and Physical Sciences Research Council (ESRC). This paper represents the views of the authors, not the funders or data providers.

ORCID iDs

Emmanouil Tranos  <https://orcid.org/0000-0002-9620-6542>

Max Nathan  <https://orcid.org/0000-0002-4617-4300>

Supplemental Material

Supplemental material for this article is available online.

Notes

1. Within this field we can pick out four broad types of work. Economic geographers focused on cluster micro-foundations, and specifically the relative importance of within-industry localization (Marshall-Arrow-Romer) versus cross-industry (Jacobs) effects (Ellison and Glaeser, 1997; Glaeser et al., 1992; Henderson, 2007). Evolutionary perspectives have highlighted the role of path-dependence and cluster branching in shaping outcomes (Boschma and Frenken, 2011; Martin and Sunley, 2006). Globalization scholars have explored how clusters sit within larger cross-national production systems such as global value chains or production networks (Yeung and Coe, 2015). Organizational scholars have argued that temporary and online collaborations complement and substitute for physical co-location (Grabher and Ibert, 2014).
2. OECD.Stat, percentage of businesses with a website or homepage, firms with 10 or more employees. Accessed 8 February 2019.
3. <http://data.webarchive.org.uk/opendata/ukwa.ds.2/>, accessed 23 September 2019.
4. E.g. the Common Crawl: <https://commoncrawl.org/>, accessed 4 June 2021.
5. See Figure S1 in Supplementary Material for an example of an archived webpage using the Internet Archive GUI, known as the Wayback Machine.
6. These are standard exclusions policies used by websites to define their interactions with other websites and web crawlers such as search engines and are included in a robots.txt file.
7. Running the analysis on full website text substantially increased noise and led to less interpretable topics. See relevant section in the Supplementary Material.

References

- Ainsworth SG, Alsum A, SalahEldeen H, et al. (2011) How much of the web is archived? In: *Proceedings of the 11th Annual International ACM/IEEE Joint Conference on Digital Libraries*. ACM, pp. 133–136.
- Arora S, Youtie J, Shapira P, et al. (2013) Entry strategies in an emerging technology: a pilot web-based study of graphene firms. *Scientometrics* 95(3): 1189–1207.
- Arribas-Bel D (2014) Accidental, open and everywhere: Emerging data sources for the understanding of cities. *Applied Geography* 49: 45–53.
- Bagwell S (2008) Creative clusters and city growth. *Creative Industries Journal* 1(1): 31–46.
- Baldwin JR, Brown WM and Rigby DL (2010) Agglomeration economies: microdata panel estimates from Canadian manufacturing. *Journal of Regional Science* 50(5): 915–934.
- Balland P-A, Boschma R and Frenken K (2015) Proximity and Innovation: From Statics to Dynamics. *Regional Studies* 49(6): 907–920.
- Blazquez D and Domenech J (2018a) Big Data sources and methods for social and economic analyses. *Technological Forecasting and Social Change* 130: 99–113.
- Blazquez D and Domenech J (2018b) Web data mining for monitoring business export orientation. *Technological and Economic Development of Economy* 24(2): 406–428.
- Blei DM (2012) Probabilistic topic models. *Communications of the ACM* 55: 77–84.
- Blei DM, BcB Edu, Ng AY, et al. (2003) Latent Dirichlet allocation. *Journal of Machine Learning Research* 3: 993–1022.

- Blei DM and Lafferty JD (2006) Dynamic topic models. *Proceedings of the 23rd International Conference on Machine Learning*: 113–120.
- Boschma R and Frenken K (2011) The emerging empirics of evolutionary economic geography. *Journal of Economic Geography* 11(2): 295–307.
- Boschma R and Iammarino S (2009) Related variety, trade linkages, and regional growth in Italy. *Economic Geography* 85(3): 289–311.
- Cariagliu A, de Dominicis L and de Groot HLF (2016) Both Marshall and Jacobs were right! AU - Caragliu, Andrea. *Economic Geography* 92(1): 87–111.
- Catini R, Karamshuk D, Penner O, et al. (2015) Identifying geographic clusters: a network analytic approach. *Research Policy* 44(9): 1749–1762.
- Chang J, Gerrish S, Wang C, et al. (2009) Reading tea leaves: how humans interpret topic models. *Advances in Neural Information Processing Systems*: 288–296.
- Crampton JW, Graham M, Poorthuis A, et al. (2013) Beyond the geotag: situating ‘big data’ and leveraging the potential of the geoweb. *Cartography and Geographic Information Science* 40(2): 130–139.
- Currid E (2007) *The Warhol Economy: How Fashion, Art, and Music Drive New York City*. Princeton: Princeton University Press.
- Cushman and Wakefield (2013) *From Goldman to Google*. London: Cushman & Wakefield.
- Delgado M, Porter ME and Stern S (2015) Defining clusters of related industries. *Journal of Economic Geography* 16(1): 1–38.
- Durantón G (2011) California dreamin’: the feeble case for cluster policies. *Review of Economic Analysis* 3(1): 3–45.
- Durantón G and Kerr W (2015) The Logic of Agglomeration. Reportno. Report Number[, Date. Place Published]: Institution|.
- Ellison G and Glaeser EL (1997) Geographic concentration in U.S. manufacturing industries: a dashboard approach. *Journal of Political Economy* 105(5): 889–927.
- Foord J (2013) The new boomtown? *Creative City to Tech City in East London Cities* 33(August): 51–60.
- Frenken K, Cefis E and Stam E (2015) Industrial dynamics and clusters: a survey. *Regional Studies* 49(1): 10–27.
- Glaeser E, Kallal H, Scheinkmann J, et al. (1992) Growth in cities. *Journal of Political Economy* 100(6): 1126–1152.
- Gök A, Waterworth A and Shapira P (2015) Use of web mining in studying innovation. *Scientometrics* 102(1): 653–671.
- Grabher G and Ibert O (2014) Distance as asset? Knowledge collaboration in hybrid virtual communities. *Journal of Economic Geography* 14(1): 97–123.
- Hale SA, Blank G and Alexander VD (2017) Live versus archive: Comparing a web archive to a population of web pages. In: Brügger N and Schroeder R (eds). *Web as History: Using Web Archives to Understand the Past and the Present*. London: UCL Press, 45–61.
- Hale SA, Yasseri T, Cowls J, et al. (2014) Mapping the UK webspace: fifteen years of british universities on the web. In: *Proceedings of the 2014 ACM conference on Web science*. Bloomington, Indiana, USA: ACM, pp. 62–70. 2615691.
- Hall P (1998) *Cities in Civilisation: Culture, Innovation and Urban Order*. London: Weidenfeld and Nicholson.
- Hamnett C (2003) Gentrification and the Middle-class Remaking of Inner London, 1961–2001. *Urban Studies* 40(12): 2401–2426.
- Harris A (2012) Art and gentrification: pursuing the urban pastoral in Hoxton, London. *Transactions of the Institute of British Geographers* 37(2): 226–241.
- Henderson JV (2007) Understanding knowledge spillovers. *Regional Science and Urban Economics* 37(4): 497–508.
- Hernández B, Jiménez J and Martín MJ (2009) Key website factors in e-business strategy. *International Journal of Information Management* 29(5): 362–371.

- Hill LL (2009) *Georeferencing: The Geographic Associations of Information*. Cambridge, MA: MIT Press.
- Holzmann H, Nejdil W and Anand A (2016) The dawn of today's popular domains: a study of the archived German web over 18 years. In: *Digital Libraries (JCDL), 2016 IEEE/ACM Joint Conference*. IEEE, pp. 73–82.
- Hope O (2017) The changing face of the online world. Available at:(accessed on 26 February)<https://www.nominet.uk/changing-face-online-world/>
- Hu Y, Deng C and Zhou Z (2019) A semantic and sentiment analysis on online neighborhood reviews for understanding the perceptions of people toward their living environments. *Annals of the American Association of Geographers* 109(4): 1052–1073.
- Hutton T (2008) *The New Economy of the Inner City: Restructuring, Regeneration and Dislocation in the Twenty-First Century Metropolis*. Abingdon: Routledge.
- Internet Archive (2016) *Internet Archive Blogs*.
- Jackson AN (2017) JISC UK web domain dataset (1996–2010) geoindex. Available at: DOI: [10.5259/ukwa.ds.2/geo/1](https://doi.org/10.5259/ukwa.ds.2/geo/1)
- JISC and the Internet Archive (2013) JISC UK web domain dataset (1996–2013). Available at: DOI: [10.5259/ukwa.ds.2/1](https://doi.org/10.5259/ukwa.ds.2/1)
- Jones E (2017) *Planning for Tech City in Post-recession London*. London: UCL.
- Kerr W and Kominers S (2015) Agglomerative Forces and Cluster Shapes. *Review of Economics and Statistics* 97(4): 877–899.
- Kinne J and Axenbeck J (2020) Web mining for innovation ecosystem mapping: a framework and a large-scale pilot study. *Scientometrics* 125(3): 2011–2041.
- Kinne J and Resch B (2018) Generating big spatial data on firm innovation activity from text-mined firm websites. *GI Forum* 1: 82–89.
- Krestel R, Fankhauser P and Nejdil W (2009) Latent dirichlet allocation for tag recommendation. In: *Proceedings of the Third ACM Conference on Recommender Systems*. ACM, pp. 61–68.
- Lansley G and Longley PA (2016) The geography of Twitter topics in London. *Computers, Environment and Urban Systems* 58: 85–96.
- Lee M, Liu Z, Huang R, et al. (2016) *BMC Bioinformatics*. Springer, pp. 153–162. Application of dynamic topic models to toxicogenomics data.
- Li Y, Arora S, Youtie J, et al. (2018) Using web mining to explore triple helix influences on growth in small and mid-size firms. *Technovation* 76-77: 3–14.
- Li Z, Wang C, Xie X, et al. (2007) Exploring LDA-based document model for geographic information retrieval. In: *Workshop of the Cross-Language Evaluation Forum for European Languages*. Springer, pp. 842–849.
- Marshall A (1890) *Principles of Economics*. New York: Macmillan.
- Martin ME and Schuurman N (2017) Area-based topic modeling and visualization of social media for qualitative GIS. *Annals of the American Association of Geographers* 107(5): 1028–1039.
- Martin ME and Schuurman N (2020) Social media big data acquisition and analysis for qualitative GIScience: Challenges and opportunities. *Annals of the American Association of Geographers* 110(5): 1335–1352.
- Martin R and Sunley P (2006) Path dependence and regional economic evolution. *Journal of Economic Geography* 6(4): 395–437.
- Martin R and Sunley P (2011) Conceptualizing cluster evolution: beyond the life cycle model? *Regional Studies* 45(10): 1299–1318.
- Martins J (2015a) The extended workplace in a creative cluster: exploring space (s) of digital work in silicon roundabout. *Journal of Urban Design* 20(1): 125–145.
- Martins J (2015b) The extended workplace in a creative cluster: exploring space(s) of digital work in silicon roundabout. *Journal of Urban Design* 20(1): 25–145.
- Musso M and Merletti F (2016) This is the future: a reconstruction of the UK business web space (1996–2001). *New Media & Society* 18(7): 1120–1142.

- Nathan M and Rosso A (2015) Mapping digital businesses with big data: some early findings from the UK. *Research Policy* 44(9): 1714–1733.
- Nathan M and Vandore E (2014) Here be startups: exploring London's 'Tech City' digital cluster. *Environment and Planning A* 46(10): 2283–2299.
- Nathan M, Vandore E and Voss G (2019) Spatial imaginaries and tech cities: place-branding East London's digital economy. *Journal of Economic Geography* 19(2): 409–432.
- Neffke F, Henning M and Boschma R (2011) How do regions diversify over time? Industry relatedness and the development of new growth paths in regions. *Economic Geography* 87(3): 237–265.
- OECD (2001) OECD Communications Outlook 2001. Reportno. Report Number[, Date. Place Published]: Institution|.
- OECD (2013) Measuring the Internet Economy: A contribution to the research agenda. Reportno. Report Number[, Date. Place Published]: Institution|.
- Papagiannidis S, Gebka B, Gertner D, et al. (2015) Diffusion of web technologies and practices: A longitudinal study. *Technological Forecasting and Social Change* 96: 308–321.
- Papagiannidis S, See-To EWK, Assimakopoulos DG, et al. (2018) Identifying industrial clusters with a novel big-data methodology: Are SIC codes (not) fit for purpose in the Internet age? *Computers & Operations Research* 98: 355–366.
- Pickles J (1995) *Ground Truth: The Social Implications of Geographic Information Systems*. London: Guildford Press.
- Porter MF (2006) *An Algorithm for Suffix Stripping*. Program.
- Pratt AC (2009) Urban regeneration: from the artsfeel good' factor to the cultural economy: A case study of Hoxton, London. *Urban Studies* 46(5–6): 1041–1061.
- Rabari C and Storper M (2014) The digital skin of cities: urban theory and research in the age of the sensed and metered city, ubiquitous computing and big data. *Cambridge Journal of Regions, Economy and Society*: rsu021.
- Rehurek R and Sojka P (2010) Software framework for topic modelling with large corpora. In: *Proceedings of the LREC 2010 workshop on new challenges for NLP frameworks*. Citeseer.
- Schroeder R and Brügger N (2017) *The Web as History: Using Web Archives to Understand the Past and the Present*. UCL Press.
- Scott A (2014) Beyond the creative city: cognitive-cultural capitalism and the new urbanism. *Regional Studies* 48(4): 565–578.
- Scott AJ (1997) The cultural economy of cities. *International Journal of Urban and Regional Research* 21(2): 323–339.
- Shalit U, Weinshall D and Chechik G (2013) Modeling musical influence with topic models. In: *International Conference on Machine Learning*. PMLR, pp. 244–252.
- Shapira P, Gök A and Salehi F (2016) Graphene enterprise: mapping innovation and business development in a strategic emerging technology. *Journal of Nanoparticle Research* 18(9): 269.
- Sievert C and Shirley K (2014) LDAvis: a method for visualizing and interpreting topics. In: *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*, pp. 63–70.
- Summers E (2020) Appraisal talk in web archives. *Archivaria* 89(1): 70–102.
- Taylor PJ, Derudder B, Faulconbridge J, et al. (2014) Advanced producer service firms as strategic networks, global cities as strategic places. *Economic Geography* 90(3): 267–291.
- Ter Wal ALJ and Boschma R (2011) Co-evolution of firms, industries and networks in space. *Regional Studies* 45(7): 919–933.
- Thelwall M (2000) Who is using the .co.uk domain? Professional and media adoption of the web. *International Journal of Information Management* 20(6): 441–453.
- Thelwall M and Vaughan L (2004) A fair history of the web? Examining country balance in the internet archive. *Library & Information Science Research* 26(2): 162–176.

- Tranos E, Kitsos T and Ortega-Argilés R (2020) Digital economy in the UK: regional productivity effects of early adoption. *Regional Studies* 55: 1–15.
- Tranos E and Stich C (2020) Individual internet usage and the availability of online content of local interest: a multilevel approach. *Computers, Environment and Urban Systems* 79: 101371.
- Uyarra E and Ramlogan R (2013) The Effects of Cluster Policy on Innovation. Reportno. Report Number[, Date. Place Published]: Institution].
- Yeung HW-c and Coe NM (2015) Toward a dynamic theory of global production networks. *Economic Geography* 91(1): 29–58.
- Zook MA (2000) The web of production: the economic geography of commercial internet content production in the United States. *Environment and Planning A* 32: 411–426.
- Zook MA (2001) Old hierarchies or new networks of centrality? – The global geography of the internet content market. *American Behavioral Scientist* 44(10): 1679–1696.
- Zukin S (1982) *Loft Living: Culture and Capital in Urban Change*. Baltimore: Johns Hopkins University Press.

Author Biographies

Dr Christoph Stich is an experienced data scientist with a strong quantitative and interdisciplinary background. He has extensive experience in utilising big data, applied machine learning, and statistics. He is passionate about explaining data science to non-technical audiences. He also has extensive experience in managing and conducting data science projects.

Dr Emmanouil Tranos a Reader in Quantitative Human Geography at the University of Bristol and a Fellow at The Alan Turing Institute. His research has been exposing the spatial dimensions of digital technologies and the digital economy from their early stages until today. He has published on issues related to the geography of the internet, the economic impacts that digital technologies can generate on cities and regions and the position of cities within spatial, complex networks. He has a strong interest and expertise on the use of new sources of big data to better understand the complexities of smart cities and urban systems.

Dr Max Nathan is Associate Professor in Applied Urban Sciences at UCL. He is also an affiliate in the Urban Programme at the Centre for Economic Performance. He completed his PhD at LSE in 2011. His academic research focuses on the economics of cultural diversity, in particular the performance of diverse cities, communities and teams; innovation systems and clusters, especially in tech and creative industries; and public policy for cities, especially policy design and evaluation. He co-founded the Centre for Cities think tank and the What Works Centre for Local Economic Growth. He is also affiliated to CAGE, IZA, NIESR and the Centre for London. He has worked in Whitehall at the Department of Communities and Local Government on issues like localism, regeneration and economic development.