UCL CENTRE FOR ADVANCED SPATIAL ANALYSIS

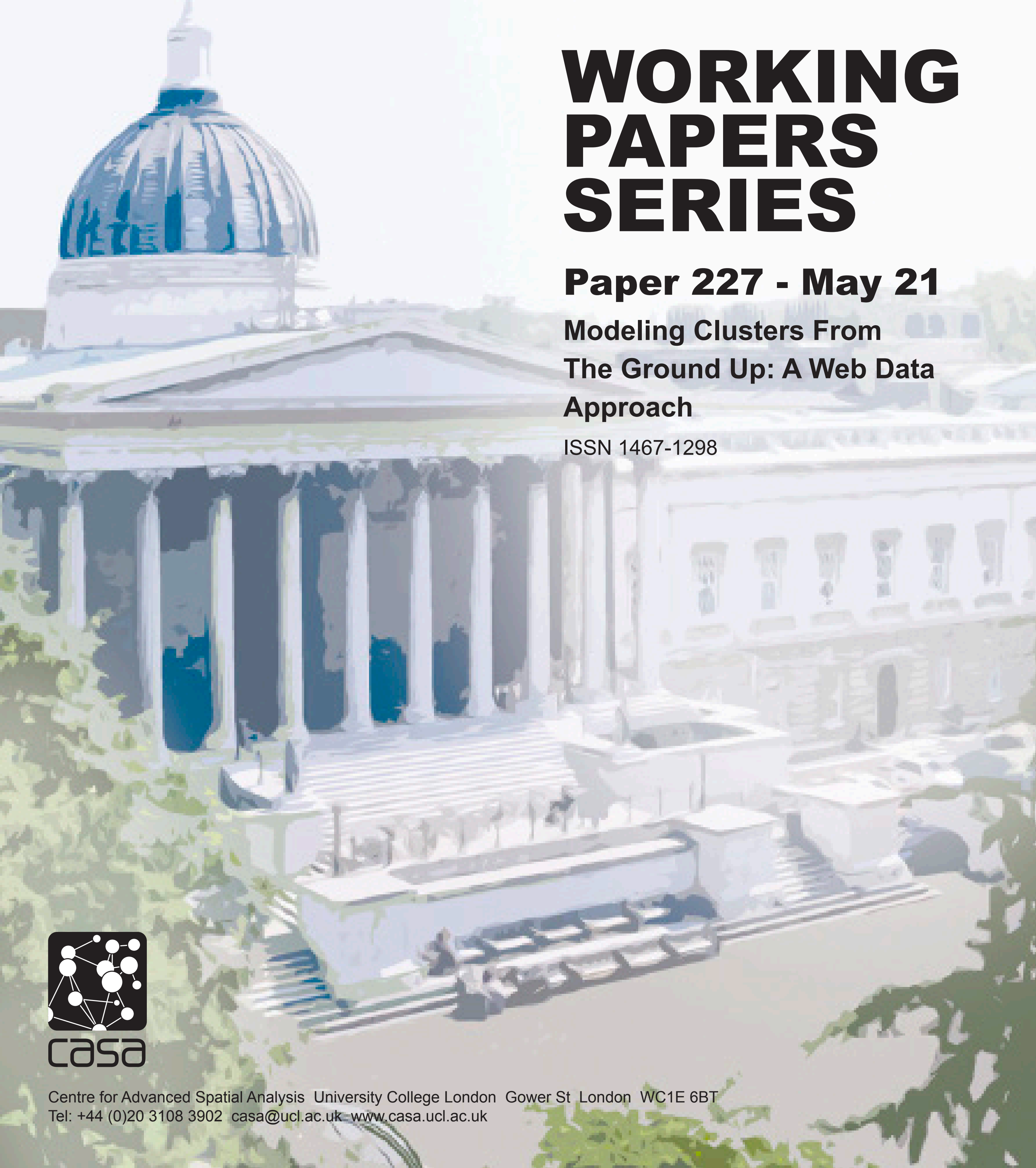
UCL

WORKING PAPERS SERIES

Paper 227 - May 21

Modeling Clusters From The Ground Up: A Web Data Approach

ISSN 1467-1298



casa

Centre for Advanced Spatial Analysis University College London Gower St London WC1E 6BT
Tel: +44 (0)20 3108 3902 casa@ucl.ac.uk www.casa.ucl.ac.uk

MODELING CLUSTERS FROM THE GROUND UP: A WEB DATA APPROACH

Christoph Stich[1], Emmanouil Tranos[2, 3*], Max Nathan[4, 5]

[1] University of Birmingham, [2] University of Bristol, [3] Alan Turing Institute,

[4] University College London, [5] Centre for Economic Performance.

christoph.stich@gmail.com, e.tranos@bristol.ac.uk, max.nathan@ucl.ac.uk.

* Corresponding author.

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

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. The first two authors have received funding from the Consumer Data Research Centre (CDRC) and ESRC. This paper represents the views of the authors, not the funders or data providers.

1/ Introduction

Clusters are most simply understood as physically co-located, interacting groups of firms – as Marshall (1890) first identified them. There is now a vast theoretical and empirical literature on cluster formation, characteristics and dynamics (see Duranton (2011), Uyarra and Ramlogan (2013), Chatterji et al (2014) and Duranton and Kerr (2015) for recent reviews). Within this field we can pick out four main schools of thought. Urban economists and economic geographers have tended to focus on cluster microfoundations, and specifically the relative importance of within-industry localization (Marshall-Arrow-Romer) versus cross-industry (Jacobs) effects (Krugman 1991; Glaeser et al. 1992; Ellison and Glaeser 1997; Fujita, Krugman, and Venables 1999; Henderson 2007). Evolutionary perspectives have highlighted the role of path-dependence and cluster branching in shaping outcomes (Martin and Sunley 2006; Boschma and Frenken 2011). Globalization scholars have explored how clusters sit within larger cross-national production systems such as global value chains or production networks (Gereffi, Humphrey, and Sturgeon 2005; Sturgeon, Van Biesebroeck, and Gereffi 2008; Yeung and Coe 2015). More recently, organizational scholars have argued that temporary and online collaborations complement and substitute for physical co-location (Bathelt 2005; Grabher and Ibert 2014).

Despite this wealth of activity, many basic questions in this field remain unresolved. First, we are still unclear about the relative salience of different cluster microfoundations, in particular the balance between industrial specialization and diversity (Ellison, Glaeser, and Kerr 2010; Kerr and Kominers 2015; Frenken, Cefis, and Stam 2015; Cariagliu, de Dominicis, and de Groot 2016). Frameworks for cluster evolution are still in debate, particularly the desirable level of analytical generalizability (Boschma and Iammarino 2009; Boschma and Fornahl

2011; Martin and Sunley 2011; Neffke, Henning, and Boschma 2011). The feasibility of cluster policy and the appropriate policy mix also remain unclear (Martin and Sunley 2003; Duranton 2011; Duranton and Kerr 2015; McCann and Ortega-Argilés 2013).

These questions are hard in part because of a number of hard-to-fix empirical challenges. Clustering does not always take place at the scale of available data, and working at inappropriate scales can distort results (the Modifiable Areal Unit Problem, or MUAP). Researchers have turned to geocoded plant-level information to tackle this (see inter alia Duranton and Overman (2005), Henderson (2003), Baldwin et al (2010) and Neffke et al (2011)). However, the industrial classifications used in this kind of 'administrative big data' are backward-looking and tend to lag behind 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 and media-based data is one way for researchers to work with companies as they describe themselves (Nathan and Rosso 2015). Third, relations between cluster participants have remained very hard to look at in any structured way (Park et al. (2019) is a recent exception). Fourth, cutting across all of these are tradeoffs between data richness and reach. Firm censuses ask limited questions, while novel 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, Waterworth, and Shapira 2015).

Our paper makes two contributions to the field. First, we propose a novel approach to analyzing clusters over time, based on web data and data science methods. This structured approach tackles a number of the analytical challenges facing empirical cluster research, including MUAP, the industrial classification problem and the richness/reach tradeoff. This enables us to cleanly 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 East London, only hitherto explored through a handful of case studies (Foord 2013a; Nathan and Vandore 2014; Nathan, Vandore, and Voss 2019; Martins 2015b). 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.

Specifically, we exploit a novel cache of archived and geolocated website data 2000-2012 from the Internet Archive, the JISC UK Web Domain Dataset (JISC and the Internet Archive 2013; Jackson 2013). While in the public domain, this dataset has been rarely used by social scientists (Tranos, Kitsos, and Ortega-Argilés 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 then use topic modelling of metadata to bundle activities in economic space, working both across the cluster and within modelled ‘verticals’. We apply this approach to the Shoreditch case study. We expose the micro-geography of sectors within the cluster as we see co-location of related activities at the postcode level; explore cluster-level topics, their granular content and their evolution over time; and provide a detailed breakdown of ‘creative digital’ industry space. Our model reproduces several stylized facts about the cluster, for example picking out the growth of creative digital activities and the uptick of activity after the introduction of the ‘Tech City’ cluster program in 2010. We are able to observe the evolution of the different economic activities within Shoreditch and also processes of branching out of new and technologically related activities.

We contribute to an evolving literature which aims to expose the mechanisms of cluster formation and success in detail, by moving beyond a pre-determined understanding of economic clusters in spatial, temporal and technological terms (Ter Wal and Boschma 2011; Balland, Boschma, and Frenken 2015; Catini et al. 2015; Delgado, Porter, and Stern 2015). We also join a growing literature employing web data for economic analysis (Gentzkow, Kelly, and Taddy 2019) and specifically economic geography research questions (Musso and Merletti 2016; Papagiannidis et al. 2018).

The paper is structured as follows. Section 2 discusses how web data has been utilized in business and innovation studies. Section 3 presents the data and methods, alongside key characteristics of the test study area. Section 4 gives our results. Section 5 concludes and identifies further opportunities for using such data in economic geography research.

2/ Using website data for economic analysis: a review

Just like most economic and social activities, business behaviors, patterns and actions leave digital traces that can be used in order to learn more about firms (Rabari and Storper 2014; Arribas-Bel 2014). One example is website data, which are readily available, cheap to obtain and extensive in terms of the theme and population coverage. The vast majority of businesses in more developed countries maintain websites, which act as self-reporting platforms and include valuable business information. Across OECD countries in 2018, for example, over 81 per cent of firms with 10 or more employees had a website or home page.[1] Coverage for smaller firms is only slightly less: in 2014 75 per cent of all UK companies with at least one employee maintained a website (Gök, Waterworth, and Shapira 2015), and this figure is likely to be similar in other more developed country settings. Business website text typically contains qualitative information on a large array of themes: from the types of economic activity a firm is engaging with and the outputs of the firms (products and services), to export orientation, research and development (R&D) and innovation activities (Blazquez and Domenech (2018a, 2018b). 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). 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 per cent of all websites contain some place reference (Hill 2009). We exploit all of these features in our analysis.

Until recently researchers have made limited use of live and archived web data. HTML text is unstructured. Corporate websites are also highly varied, both in terms of design and the information they contain. These lead to computational challenges that can only be tackled by using tools and methods outside the traditional social science toolkit. We use a number of big data analytics tools in this paper, mainly drawn from Natural Language Processing (NLP).

A handful of recent studies have used web data and data science tools for industry and/or cluster analysis. Blazquez and Domenech (2018c) use web data from corporate websites to test the

[1] OECD.Stat, percentage of businesses with a website or homepage, firms with 10 or more employees. Accessed 8 Februry 2019.

export orientation of a small sample of 350 Spanish companies. They use this to 'nowcast' and track important cluster / regional features. Arora et al. (2013) and Shapira, Gök, and Salehi (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, Waterworth, and Shapira (2015) explore the R&D activities of 296 green goods SMEs based in the UK. Li et al. (2018) focus on a similar size sample of US-based SMEs working on green goods, using information from website content 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) also use these data to respectively, illustrate UK business' adoption of the web in the late 1990s, and to explore the evolution of the .uk country code Top Level Domain (ccTLD) and the linking practices of British university websites. 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 of these studies have important empirical limitations. Typically only a few hundred subjects or less are covered. Most studies only look at a single point in time. The detailed spatial dimension of the web data is also ignored, except to place firms in large administrative units. By contrast, we are able to work with 12 years of data for thousands of business websites, and explore cluster dynamics. We also use postcode level information from self-reported trading addresses, rather than the registration addresses often included in commercial firm data. Importantly, commercial or freely available firm data are not a bias-free source for business websites. Companies House, the UK's registrar of companies, does not include any information about business websites and only 24 per cent of the records that Papagiannidis et al. (2018) used included business URLs.

3/ Data and methods

We make use of a unique source of archived web data, which has 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 subset of the Internet Archive, which is curated by the British Library and includes all the archived webpages under the .uk ccTLD[2], 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 Internet Archive 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 Internet Archive contained 273 billion webpages from 361 million websites, which took up 15 petabytes of storage (Internet Archive 2016).

Our raw data consists of billions of timestamped URLs of .uk webpages, which have been archived during the period 2000-2012, accessed using the Internet Archive GUI (Graphic User

[2] http://data.webarchive.org.uk/opendata/ukwa.ds.2/, accessed 23rd September 2109.

Interface) or programmatically.[3] Our analysis uses 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 Geoindex includes 2.5 billion URLs (Jackson 2013). 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; Tranos, Kitsos, and Ortega-Argilés 2020).

Such data are not without limitations. Ainsworth et al. (2011) find that 35-90 per cent of webpages have been archived globally by any public archives. The Internet Archive, just like any other archive, only captures publicly available webpages and is constrained by robot exclusions.[4] Webpages that attract more traffic also have higher probability of being archived and being archived more often. Nevertheless, the consensus is that the Internet Archive is the most extensive and complete archive in the world (Ainsworth et al. 2011; Holzmann, Nejdl, and Anand 2016). Focusing on a subset of websites close to one used in this paper, Thelwall and Vaughan (2004) indicate that the Internet Archive captures at least one webpage for 92 per cent of all the US commercial websites[5].

3.1/ Data cleaning

[3] An example of an archived webpage using this GUI, which is known as the Wayback Machine, can be found in the Appendix (Figure AX1).

[4] 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 files.

[5] We cannot rule out poor coverage for a small number of individual websites. For instance, Hale, Blank, and Alexander (2017) compare the live and archived TripAdvisor London webpages on the Internet Archive. For this single case, they find that only 24 per cent were archived, with webpage popularity being the main driver for the archival bias.

To use website content to robustly model industrial clusters requires a number of cleaning steps. We start with the Geoindex, which contains all the archived .uk webpages with a string in the UK postcode format in the web text. We trim data to 2000-2012, as the archived web data before 2000 is quite sparse and for 2013 we only have data for the first quarter. We drop false positives (webpages with postcode-like strings which do not contain live postcodes in that year). We also keep only webpages under the .co.uk or ltd.uk second level domain, 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 ccTLD 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 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 the cleaned, archived webpages in order 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. For the case study, we further subset these data and only keep webpages with at least one postcode within the Shoreditch area during the 2000-2012 period. Following Nathan et al (2019), we define the Shoreditch cluster as a 1km 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.[6] In 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.

Figure 1 about here

Websites located at the right end of the long tail distribution include a large number of postcodes at least one of which falls within Shoreditch. These sites are typically online directories, which were popular in the beginning of the study period (see Figure AX2 in the Appendix). We drop these types of sites from our analysis 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 falls within the 1km Shoreditch zone (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. Figure 2 illustrates examples of such websites. It includes the home page of commercial websites with

[6] In principle we could also link website data to administrative datasets such as Companies House, to validate information on company status, industry and so on. However, this requires fuzzy matching on name and location, which is problematic in our case. Many firms trade under different names to the registered corporate entity. Relatedly, registered addresses in Companies House often may not correspond with actual trading locations. Further, company owners pick their own industry codes, and a non-trivial share are missing or are uninformative. For more on these issues see (Nathan and Rosso 2015).

a unique postcode within Shoreditch, where usually the economic activity is presented, and the 'contact us' page, where usually the Shoreditch postcode can be found. At a second stage we run a sensitivity check by running the analysis to a larger sample that includes websites with up to 11 postcodes, at least one of which is in the Shoreditch zone (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.

Figure 2 about here

We deal briefly here with two other concerns. Firms use websites in a range of 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 of the way the Internet Archive operates. Specifically, the crawler finds and archives a given website based on the hyperlinks from other websites leading to that site. We would expect 'placeholder' or defunct websites to contain zero or very few valid hyperlinks from other sites. Moreover, we would not expect defunct firms to continue paying domain names fees.

3.2 Topic modelling

We use Latent Dirichlet Allocation (LDA) to analyze the cleaned website data. In particular, we use an extension to LDA by Blei and Lafferty (2006) to explicitly account for the temporal evolution of the dataset. LDA is a widely used tool in natural language processing. More recently, 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), or to identify classes of economic activities in a region (Papagiannidis et al. 2018).

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 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 can 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 frameworks, we can then look at the growth and change of topics over time.

The intuition of LDA is that each website – or *document,* per text analysis terminology – is composed of several different overlapping topics, which together form the overall economic activity space. However, we cannot directly observe the topics, the hidden structure we are interested in; only the words that make up the documents. More 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:

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

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, Fankhauser, and Nejdl 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 (see for example Blei and Lafferty 2006; Lee et al. 2016; Shalit, Weinshall, and Chechik 2013) and has the advantage to allow for 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, in order 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.[8] As we are not trying to predict the topics of future documents, we do not need to keep training and testing data separately.

[7] Running the analysis on full website text substantially increased noise and led to less interpretable topics.

[8] https://www.nltk.org/_modules/nltk/stem/snowball.html

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.

3.3 Spatial analysis

We use the postcode-level information to conduct exploratory spatial analysis within the cluster. We start by analyzing the distribution of website counts at the postcode level. This allows us to identify potential spatial concentration of economic activities. High concentrations of commercial websites in specific postcodes likely illustrates co-location of economic activity, as in a classical Marshallian industrial district. Second, we explore the spatial signatures of different topics and investigate whether they are co-located, using heat maps to take advantage of our very fine-grained data.[9] As cluster location can act in part as a signaling device for firms (Appold 2005, Romanelli and Khessina 2005), very high website/postcode densities outliers may be driven by intermediaries offering 'prestige addresses' to firms actually based elsewhere. We identify one such outlier in our cleaning and remove it from the data.

[9] Our data can include firms with multiple commercial websites, which point to different economic activities, but to the same physical location. Although we are not able to directly identify these cases as we cannot match firms with archived commercial websites, these cases do not pose any concerns given that our focus is on economic activities presented on commercial websites and not on firms.

3.4 / Validation

We deploy three strategies to validate our findings. Crucially, we implement our approach to model a well-known technology cluster in East London, the Shoreditch / ‘Tech City’ ecosystem. This case provides us with theoretical and empirical stylized facts, allowing us to benchmark our results against established ground truth (Pickles 1995). We also apply two more technical checks. We reproduce our cluster-level analysis using a larger set of websites containing up to 11 different postcodes: multi-postcode websites represent larger multi-site firms, including chains. These may be economically important but less embedded in the cluster itself. We also compare results derived from our web-based methods with a more traditional approach based on administrative microdata from Companies House, a comprehensive UK company register. This exercise illustrates how open data and data science methods can complement more established analytical approaches in understanding clusters.

4/ Results

We apply our framework to model a well-known technology cluster in Shoreditch, East London (now widely known as ‘Tech City’). It acts as a good test case, having much in common with urban technology production districts in large cities around the world (such as in New York, San Francisco, Berlin, Stockholm, Tel Aviv and LA), including in its evolution from ‘depressed’ ex-industrial area to ‘vibrant’ post-industrial milieu (Zukin 1982; Scott 1997; Hall 1998; Hutton 2008; Scott 2014). Here we briefly set out some stylized facts, drawing on existing qualitative and quantitative case studies. These form the ground truth which 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. Like many urban clusters it is tightly drawn around key physical landmarks, notably 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 gradually 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 2013a; Nathan, Vandore, and Voss 2019). Proximity to London's main financial district gives the area a body of financial and business services firms, with a number of new office developments in recent years. The area has become a desirable residential neighborhood, with extensive new luxury apartment developments and accompanying real estate and local amenities for well-off incomers. At the same time, a vibrant visitor, 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). In common with similar clusters in other cities, the creative technology community grew 'organically' for many years before coming to the attention of policymakers (Pratt 2009; Foord 2013a; Nathan and Vandore 2014; Jones 2017). 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, Vandore, and Voss 2019).

4.1 / 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, like Shoreditch, they can even consist of a single building. With this in mind 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). And as Figure 4a illustrates, this postcode refers to a relatively small building. An online search for this postcode provided an explanation for this outlier: these are the premises of a virtual address provider, which enables businesses to use their premises as their postal or business registration addresses. Importantly, as Figure 4b demonstrates, on top of businesses with only a virtual presence in Shoreditch, there is also a ‘digital squatting’ phenomenon, as companies use this postal address – also on their websites – without the authorization of the virtual address provider. Both cases are examples of the cluster signaling effect: these companies gain benefits not from actual physical co-location, but instead from presenting as part of the cluster (via postcodes on websites). In this case, we decided to remove the websites anchored to this postcode from the analysis presented in the next sections.

Figure 3 about here

Figure 4a and 4b about here

4.2 / Cluster-level analysis

We present here the LDA results for the 8,801 commercial websites with one unique postcode within the 1km Shoreditch zone. One of the LDA parameters that needs to be exogenously defined is the number of topics. Because we want to analyze the industrial structure of Shoreditch we opted for the highest number of topics up to the point that the derived topics cannot be manually labeled. Hence, Table 1 presents the LDA outputs for k=15 topics. It needs to be highlighted though that even solutions with less topics, which can be provided upon request, lead to similar conclusions when we look at the topic terms.

Table 1 about here

The last column of Table 1 presents the 20 most frequent terms – or, in other words, 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 2013a; Nathan and Vandore 2014; Jones 2017; Nathan, Vandore, and Voss 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* (Foord 2013b).

Topics 3, 8, 9 and 12 depict the art scene of Shoreditch. The pre-WW2 craft tradition of the area 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, etc.*), while visual arts can be found in Topic 9 (*design, art, photograph, architecture, architect, interior etc.*). Topic 12 represents fashion related economic activities (*fashion, design, cloth, watch*). Again, these LDA findings are in accordance with previous research and also reflect past urban economic developments programs, which aimed to support creative industries including fashion, jewelry and furniture makers (Foord 2013a). 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; Taylor et al. 2013). 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, 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*).

4.3 / 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, Henning, and Boschma 2011). Our framework can explore these temporal dynamics by looking at the topic prevalence (Figure 5) and within topics term frequency (Figure 6).

Again, our framework cleanly reproduces existing stylized facts. In line with existing studies of Shoreditch (Cushman and Wakefield 2013; Nathan, Vandore, and Voss 2019; Harris 2012), digital media (Topic 1) is the most prevalent topic in the study area with a brief exception during the post dotcom crash period (2003-2005, Figure 5). 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, Vandore, and Voss 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 consistent with 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 reflecting the resurgence of the crafts and art industries (Foord 2013a). A steady but small increase can also be observed for fashion and trade (Topic 12), which can be linked to publicly funded initiatives to support creative sectors such as the 2003–2009 City Growth Programme (Bagwell 2008).

Figure 5 about here

Figure 6 presents the within topics term frequency to assess how the consistency of topics changes over time. Starting from the digital media topic (Topic 1), the term frequency remains stable over time. 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 2013a).

Contrary to the topics linked to digital and creative activities, business and financial activities topics are not as stable during the study period. For instance, 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 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. For instance, we can observe the growth of term *onlin* in topic 3 (craft) and 11 (business support), and *softwar* and *mobil* in topic 4 (business technology services). The growth of these terms is consistent with both the overall growth of digital technologies during the study period, but also to technological diffusion within Shoreditch, from the dominant economic activities reflected in topic 1 (digital media) to other economic activities.[10]

[10] It would be possible to disentangle these local / global processes by comparing the spread of terms linked to digitization a) within the cluster and b) across our entire corpus of websites. This is a major exercise, arguably out of scope of this paper.

Figure 6 about here

All in all, 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 within Shoreditch, 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.

4.4 / Cluster footprint

The heatmaps of the websites assigned to the different topics derived from the dynamic LDA model (Figure 7) enable us to analyze the spatial structure of the different economic activities within Shoreditch. Interestingly, the topics linked to the digital and creative character of Shoreditch (Topics 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). This should not come as a surprise as this is the area where the Barbican, a large arts center is located. Topic 14, which depicts the hospitality industry, has the same epicenter as the digital media topic reflecting again how interwoven these topics are. Nevertheless, as expected, it captures all of the study area. The same applies to 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 south part of the study area, which is adjacent to the City of London, a world-leading financial cluster. In total, although the maps clearly indicate two distinct poles in the study area – that is the more creative north west 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.

Figure 7 about here

The above analysis 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 methods confirms the results from previous studies, which were based on extensive interviews and fieldwork (Nathan, Vandore, and Voss 2019; Martins 2015b; Foord 2013b), web inquiries on a pre-defined small sample of firms (Taylor et al. 2014; Taylor et al. 2013), or secondary data analysis from propriety data providers (Foord 2013b). 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. The next section provides robustness checks by (i) using an extended sample of archived, commercial websites linked to Shoreditch, and (ii) by comparing the depth of analysis that our proposed research framework can achieve against the use of administrative business records.

4.5 / Robustness checks

We first run a sensitivity test with a larger, but less locally-specific set of websites. Our main analysis uses only archived commercial websites with a unique postcode within Shoreditch. We re-run the analysis using a larger set of websites with up to 11 different postcodes, at least one of which is located within the Shoreditch area. This subset includes 23,412 websites, which represent 50 per cent of the universe of all the archived commercial websites with at least one postcode within the Shoreditch area in 2000-2012. 32 per cent of the unique postcodes in this subset are located outside the broader London area and, as expected, they decrease the sharpness of the LDA topics. In the Appendix, Figure AX3 gives the distribution of postcode distances from Shoreditch roundabout and Table AX1 illustrates the LDA outputs for 15 topics. The topics are noticeably less precise in reproducing established features of the local economy given the more extensive spatial reach of this subset. Nevertheless, the LDA outputs for year 2012, presented in Table AX1, still reveal digital media (Topic 2 and 7), arts and craft (Topic3, 11 and 15), business and financial services (Topic 1, 4, 10 and 13) as well as the hospitality industry (Topic 8), wellbeing (Topic 5 and 14) and real estate (Topic 6). As expected, some new topics also emerged (Topic 9 depicting travel and Topic 12 education). Importantly and despite the different subset, this exercise still highlights the area's locally-based industry mix that is related to, but distinct from a more 'generic' set of activities found across the city.

Next, we run a comparison check setting our web data and text-based results against results from conventional administrative microdata and SIC codes. To do this we use 2012 Companies House microdata provided by OpenCorporates. Companies House is a UK-wide register which includes all private companies and most partnerships; firms are obliged to register details at incorporation and provide regular financial updates. As noted earlier, Companies House industry and location variables have limitations, so we treat this exercise as broad-brush rather than forensic.

Table 2 about here

We identify all companies with a registration address within the Shoreditch zone and active during 2000-2012. We then plot the frequency of 5-digit SIC codes, the most detailed available. Table 2 gives results, color-coded to highlight some of the key points of the comparison. We find that when companies self-describe using SICs, we are unable to pick up some of the main known features of the Shoreditch ecosystem. The most frequently occurring SIC activity is ‘Management consultancy activities *other* than financial management’ (emphasis added), which accounts for 20 per cent of the registered businesses in Shoreditch. A further 36 per cent of all the registered businesses in Shoreditch defined their economic activities as ‘other’ or ‘not elsewhere classified’ (yellow-coded). We further color-coded as green SICs corresponding to digital creative activity identified in the LDA, and in prior empirical studies of the area. SICs partially reproduce some of this activity but miss much foundational detail, in particular economic activities related with branding, design, graphics, web and web services, and their intersections with artistic activity.

To sum up, the robustness checks indicated that our main findings can also be replicated when using a much larger and spatially extended subset. Moreover, our framework reveals more insights about the economic activities of the study area than using administrative data, which tends to be the mainstream for such research and policy-oriented analysis.

5/ Conclusions

Clusters, their formation and evolution are central issues in economic geography. Nevertheless, modelling clusters and their dynamics faces some hard-to-solve empirical challenges. In this paper we introduce a novel approach for analyzing and modeling clusters using public web data and data science methods, including text analysis. This is a powerful and flexible approach which 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 some new insights. We show that this approach is significantly more informative than next-best analysis using open administrative data. This approach has multiple potential applications, not only for re-analyzing existing clusters, but also in detecting unknown or emerging cluster formations.

Specifically, the use of unstructured textual data from the web and our analytical framework enabled us to move beyond the rigid SIC-based understanding of the activity space. 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. Importantly, despite the richness of our results, our methods and data are transferable to different spatial contexts. In addition, the spatial granularity of our data allow us 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. Regarding policy, although the main aim of this paper is not to assess related urban policies, we observe a correspondence between the establishment of the Tech City programme by the UK Government 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 data, which are currently unexplored.

References

Ainsworth, S. G., A. Alsum, H. SalahEldeen, M. Weigle, M. Nelson. 2011. How much of the web is archived? Paper read at Proceedings of the 11th annual international ACM/IEEE joint conference on Digital libraries.

Arora, S., J. Youtie, P. Shapira, L. Gao, T. Ma. 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, J. R., W. M. Brown, D. L. Rigby. 2010. AGGLOMERATION ECONOMIES: MICRODATA PANEL ESTIMATES FROM CANADIAN MANUFACTURING*. *Journal of Regional Science,* 50 (5):915-934.

Balland, P.-A., R. Boschma, K. Frenken. 2015. Proximity and Innovation: From Statics to Dynamics. *Regional Studies,* 49 (6):907-920.

Bathelt, H. 2005. Geographies of production: growth regimes in spatial perspective (II) - knowledge creation and growth in clusters. *Progress in Human Geography,* 29 (2):204-216.

Blazquez, D., J. Domenech. 2018a. Big Data sources and methods for social and economic analyses. *Technological Forecasting and Social Change,* 130:99-113.

———. 2018b. Web data mining for monitoring business export orientation. *Technological and Economic Development of Economy,* 24 (2):406-428.

Blazquez, D., J. Domenech. 2018c. Web data mining for monitoring business export orientation. *Technological Forecasting and Social Change,* 24 (2):406-428.

Blei, D. M. 2012. Probabilistic Topic Models. *Communications of the ACM,* 55:77-84.

Blei, D. M., B. c. B. Edu, A. Y. Ng, A. c. S. Edu, M. I. Jordan, J. c. B. Edu. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research,* 3:993-1022.

Blei, D. M., J. D. Lafferty. 2006. Dynamic topic models. Paper read at Proceedings of the 23rd international conference on Machine learning.

Boschma, R., D. Fornahl. 2011. Cluster Evolution and a Roadmap for Future Research. *Regional Studies,* 45 (10):1295-1298.

Boschma, R., K. Frenken. 2011. The emerging empirics of evolutionary economic geography. *Journal of Economic Geography,* 11 (2):295-307.

Boschma, R., S. Iammarino. 2009. Related Variety, Trade Linkages, and Regional Growth in Italy. *Economic Geography,* 85 (3):289-311.

Cariagliu, A., L. de Dominicis, H. L. F. de Groot. 2016. Both Marshall and Jacobs were Right! AU - Caragliu, Andrea. *Economic Geography,* 92 (1):87-111.

Catini, R., D. Karamshuk, O. Penner, M. Riccaboni. 2015. Identifying geographic clusters: A network analytic approach. *Research Policy,* 44 (9):1749-1762.

Chatterji, A., E. Glaeser, W. Kerr. 2014. Clusters of Entrepreneurship and Innovation. *Innovation Policy and the Economy,* 14 (1):129-166.

Crampton, J. W., M. Graham, A. Poorthuis, T. Shelton, M. Stephens, M. W. Wilson, M. Zook. 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., M. E. Porter, S. Stern. 2015. Defining clusters of related industries. *Journal of Economic Geography,* 16 (1):1-38.

Duranton, G. 2011. California Dreamin': The feeble case for cluster policies. *Review of Economic Analysis,* 3 (1):3-45.

Duranton, G., W. Kerr. 2015. The Logic of Agglomeration. In *NBER Working Paper 21452*. Cambridge, Mass: NBER.

Duranton, G., H. G. Overman. 2005. Testing for Localization Using Micro-Geographic Data. *The Review of Economic Studies,* 72 (4):1077-1106.

Ellison, G., Edward L. Glaeser. 1997. Geographic Concentration in U.S. Manufacturing Industries: A Dartboard Approach. *Journal of Political Economy,* 105 (5):889-927.

Ellison, G., E. L. Glaeser, W. Kerr. 2010. What Causes Industry Agglomeration? Evidence from Coagglomeration Patterns. *The American Economic Review,* 100 (3):1195-1213.

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

Foord, J. C.-. 2013b. The new boomtown? Creative city to Tech City in east London. *Cities,* 33:51-60.

Frenken, K., E. Cefis, E. Stam. 2015. Industrial Dynamics and Clusters: A Survey. *Regional Studies,* 49 (1):10-27.

Fujita, M., P. Krugman, A. J. Venables. 1999. *The Spatial Economy: Cities, Regions, and International Trade*. Cambridge, MA: MIT Press.

Gentzkow, M., B. Kelly, M. Taddy. 2019. Text as Data. *Journal of Economic Literature,* 57 (3):535-74.

Gereffi, G., J. Humphrey, T. Sturgeon. 2005. The governance of global value chains. *Review of International Political Economy,* 12 (1):78-104.

Glaeser, E., H. Kallal, J. Scheinkmann, A. Shleifer. 1992. Growth in Cities. *Journal of Political Economy,* 100 (6):1126.

Gök, A., A. Waterworth, P. Shapira. 2015. Use of web mining in studying innovation. *Scientometrics,* 102 (1):653-671.

Grabher, G., O. Ibert. 2014. Distance as asset? Knowledge collaboration in hybrid virtual communities. *Journal of Economic Geography,* 14 (1):97-123.

Hale, S. A., G. Blank, V. D. Alexander. 2017. Live versus archive: Comparing a web archive to a population of web pages. In *Web as History: Using Web Archives to Understand the Past and the Present,*, eds. N. Brügger and R. Schroeder. London: UCL Press.

Hale, S. A., T. Yasseri, J. Cowls, E. T. Meyer, R. Schroeder, H. Margetts. 2014. Mapping the UK webspace: fifteen years of british universities on the web. In *Proceedings of the 2014 ACM conference on Web science*, 62-70. Bloomington, Indiana, USA: ACM.

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, J. V. 2003. Marshall's scale economies. *Journal of Urban Economics,* 53 (1):1-28.

Henderson, J. V. 2007. Understanding knowledge spillovers. *Regional Science and Urban Economics,* 37 (4):497-508.

Hill, L. L. 2009. *Georeferencing: The geographic associations of information*. Cambridge, MA: MIT Press.

Holzmann, H., W. Nejdl, A. Anand. 2016. The Dawn of today's popular domains: A study of the archived German Web over 18 years. Paper read at Digital Libraries (JCDL), 2016 IEEE/ACM Joint Conference.

Hope, O. 2020. *The changing face of the online world*. Nominet 2017 [cited 26th February 2020]. Available from https://www.nominet.uk/changing-face-online-world/.

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, A. N. 2013. JISC UK Web Domain Dataset (1996-2010) Geoindex, ed. The British Library.

JISC and the Internet Archive. 2013. JISC UK Web Domain Dataset (1996-2013), ed. The British Library.

Jones, E. 2017. Planning for Tech City in post-recession London. London: UCL.

Kerr, W., S. Kominers. 2015. Agglomerative Forces and Cluster Shapes. *Review of Economics and Statistics,* 97 (4):877-899.

Krestel, R., P. Fankhauser, W. Nejdl. 2009. Latent dirichlet allocation for tag recommendation. Paper read at Proceedings of the third ACM conference on Recommender systems.

Krugman, P. 1991. *Geography and Trade*. Cambridge, MA: MIT Press.

Lansley, G., P. A. Longley. 2016. The geography of Twitter topics in London. *Computers, Environment and Urban Systems,* 58:85-96.

Lee, M., Z. Liu, R. Huang, W. Tong. 2016. Application of dynamic topic models to toxicogenomics data. Paper read at BMC bioinformatics.

Li, Y., S. Arora, J. Youtie, P. Shapira. 2018. Using web mining to explore Triple Helix influences on growth in small and mid-size firms. *Technovation,* 76-77:3-14.

Li, Z., C. Wang, X. Xie, X. Wang, W.-Y. Ma. 2007. Exploring LDA-based document model for geographic information retrieval. Paper read at Workshop of the Cross-Language Evaluation Forum for European Languages.

Marshall, A. 1890. *Principles of Economics*. 8th ed. New York: Macmillan.

Martin, M. E., N. Schuurman. 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, R., P. Sunley. 2003. Deconstructing clusters: chaotic concept or policy panacea? *Journal of Economic Geography,* 3 (1):5-35.

———. 2006. Path dependence and regional economic evolution. *Journal of Economic Geography,* 6 (4):395-437.

———. 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.

———. 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.

McCann, P., R. Ortega-Argilés. 2013. Redesigning and Reforming European Regional Policy: The Reasons, the Logic, and the Outcomes. *International Regional Science Review,* 36 (3):424-445.

Musso, M., F. Merletti. 2016. This is the future: A reconstruction of the UK business web space (1996–2001). *New Media & Society,* 18 (7):1120-1142.

Nathan, M., A. Rosso. 2015. Mapping digital businesses with Big Data: some early findings from the UK *Research Policy,* 44 (9):1714-1733.

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

Nathan, M., E. Vandore, G. Voss. 2019. Spatial Imaginaries and Tech Cities: Place-branding East London’s digital economy. *Journal of Economic Geography,* 19 (2):409-432.

Neffke, F., M. Henning, R. Boschma. 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. Paris.

———. 2013. Measuring the Internet Economy: A contribution to the research agenda. In *OECD Digital Economy Papers 226*: OECD Publishing.

Papagiannidis, S., B. Gebka, D. Gertner, F. Stahl. 2015. Diffusion of web technologies and practices: A longitudinal study. *Technological Forecasting and Social Change,* 96:308-321.

Papagiannidis, S., E. W. K. See-To, D. G. Assimakopoulos, Y. Yang. 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.

Park, J., I. B. Wood, E. Jing, A. Nematzadeh, S. Ghosh, M. D. Conover, Y.-Y. Ahn. 2019. Global labor flow network reveals the hierarchical organization and dynamics of geo-industrial clusters. *Nature Communications,* 10 (1):3449.

Pickles, J. ed. 1995. *Ground Truth: The Social Implications of Geographic Information Systems*. London: Guildford Press.

Porter, M. F. 2006. An algorithm for suffix stripping. *Program*.

Pratt, A. C. 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., M. Storper. 2014. The digital skin of cities: urban theory and research in the age of the sensored and metered city, ubiquitous computing and big data. *Cambridge Journal of Regions, Economy and Society*:rsu021.

Scott, A. 2014. Beyond the Creative City: Cognitive-Cultural Capitalism and the New Urbanism. *Regional Studies,* 48 (4):565-578.

Scott, A. J. 1997. The Cultural Economy of Cities. *International Journal of Urban and Regional Research,* 21 (2):323-339.

Shalit, U., D. Weinshall, G. Chechik. 2013. Modeling musical influence with topic models. Paper read at International Conference on Machine Learning.

Shapira, P., A. Gök, F. Salehi. 2016. Graphene enterprise: mapping innovation and business development in a strategic emerging technology. *Journal of Nanoparticle Research,* 18 (9):269.

Sievert, C., K. Shirley. 2014. LDAvis: A method for visualizing and interpreting topics. Paper read at Proceedings of the workshop on interactive language learning, visualization, and interfaces.

Sturgeon, T., J. Van Biesebroeck, G. Gereffi. 2008. Value chains, networks and clusters: reframing the global automotive industry. *Journal of Economic Geography,* 8 (3):297-321.

Taylor, P. J., B. Derudder, J. Faulconbridge, M. Hoyler, P. Ni. 2014. Advanced producer service firms as strategic networks, global cities as strategic places. *Economic Geography,* 90 (3):267-291.

Taylor, P. J., B. Derudder, M. Hoyler, P. Ni. 2013. New regional geographies of the world as practised by leading advanced producer service firms in 2010. *Transactions of the Institute of British Geographers,* 38 (3):497-511.

Ter Wal, A. L. J., R. Boschma. 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., L. Vaughan. 2004. A fair history of the Web? Examining country balance in the Internet Archive. *Library & Information Science Research,* 26 (2):162-176.

Tranos, E., T. Kitsos, R. Ortega-Argilés. 2020. Digital economy in the UK: regional productivity effects of early adoption. *Regional Studies*:1-15.

Tranos, E., C. Stich. 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., R. Ramlogan. 2013. The Effects of Cluster Policy on Innovation. In *MIoIR Compendium of Evidence on Innovation Policy*. London: NESTA / University of Manchester

Yeung, H. W.-c., N. M. Coe. 2015. Toward a Dynamic Theory of Global Production Networks. *Economic Geography,* 91 (1):29-58.

Zook, M. A. 2000. The Web of Production: The Economic Geography of Commercial Internet Content Production in the United States. *Environment and Planning A,* 32:411-26.

Zook, M. A. 2001. Old Hierarchies or New Networks of Centrality? – The Global Geography of the Internet Content Market. *American Behavioral Scientist,* 44 (10):1679-96.

Zukin, S. 1982. *Loft Living: Culture and Capital in Urban Change*. Baltimore: Johns Hopkins University Press.

Tables

Table 1: LDA topics

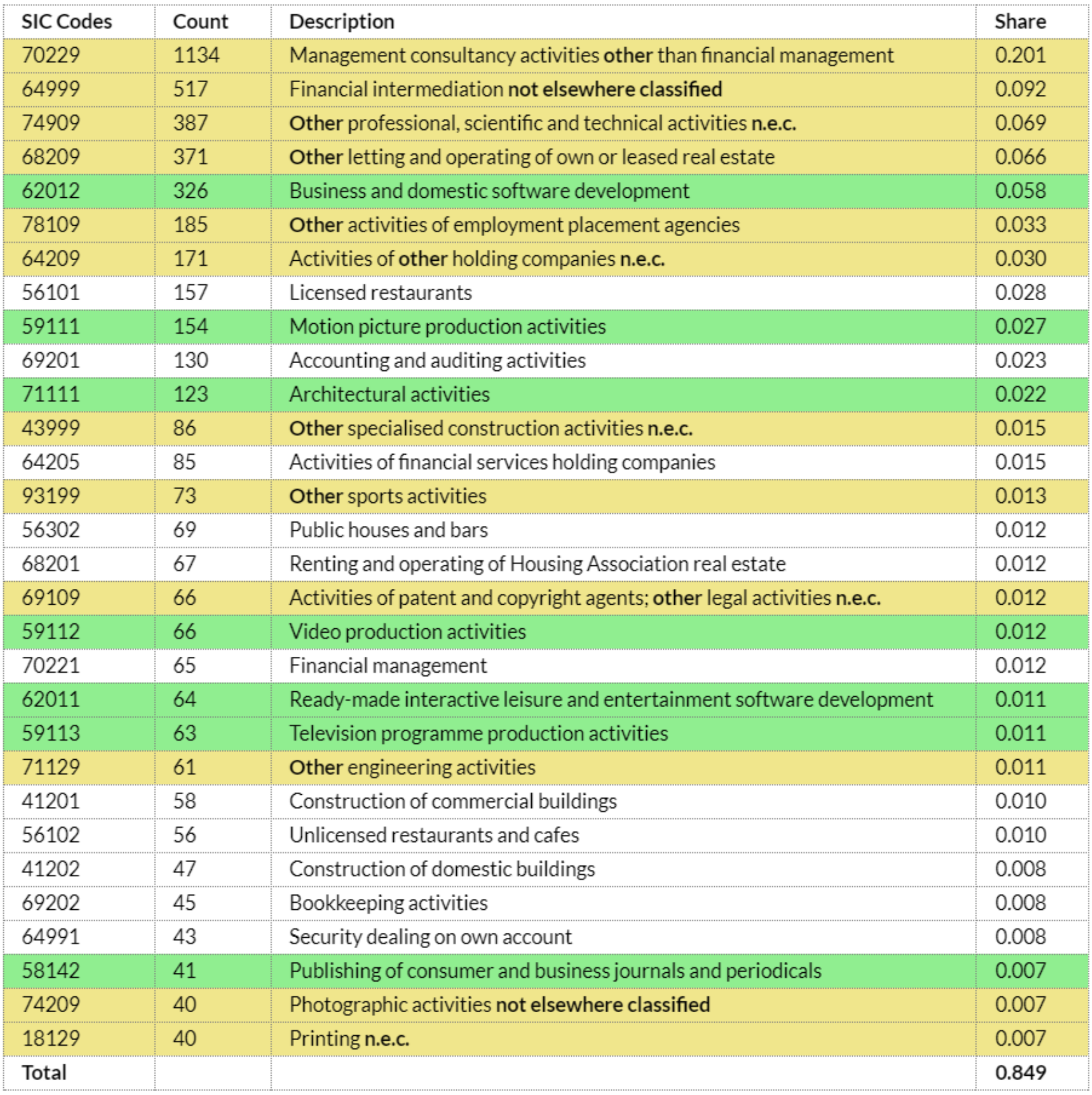
Topic	Label	Term frequency (%)	20 most frequent terms
1	digital media	13.7	design, web, brand, market, graphic, digit, websit, creativ, agenc, media, develop, product, advertis, onlin, print, site, consult, ident, imag, compani
2	business services and finance	10.5	account, job, manag, compani, recruit, 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, mobil, 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, financi, offer, equiti, time, invest, day, rate, deal, inform, book, free, cash, analysi

7	consultancy agents	6.2	consult, public, train, market, relat, communic, manag, research, 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, servic, room, discount, investig, clinic, commerci, litig, employ, 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, therapist, west, street, pain, central, ship, stress, care, south, get, well, cargo, hill

Note: terms are stemmed

Table 2: SIC frequency in Shoreditch



SIC Codes	Count	Description	Share
70229	1134	Management consultancy activities **other** than financial management	0.201
64999	517	Financial intermediation **not elsewhere classified**	0.092
74909	387	**Other** professional, scientific and technical activities **n.e.c.**	0.069
68209	371	**Other** letting and operating of own or leased real estate	0.066
62012	326	Business and domestic software development	0.058
78109	185	**Other** activities of employment placement agencies	0.033
64209	171	Activities of **other** holding companies **n.e.c.**	0.030
56101	157	Licensed restaurants	0.028
59111	154	Motion picture production activities	0.027
69201	130	Accounting and auditing activities	0.023
71111	123	Architectural activities	0.022
43999	86	**Other** specialised construction activities **n.e.c.**	0.015
64205	85	Activities of financial services holding companies	0.015
93199	73	**Other** sports activities	0.013
56302	69	Public houses and bars	0.012
68201	67	Renting and operating of Housing Association real estate	0.012
69109	66	Activities of patent and copyright agents; **other** legal activities **n.e.c.**	0.012
59112	66	Video production activities	0.012
70221	65	Financial management	0.012
62011	64	Ready-made interactive leisure and entertainment software development	0.011
59113	63	Television programme production activities	0.011
71129	61	**Other** engineering activities	0.011
41201	58	Construction of commercial buildings	0.010
56102	56	Unlicensed restaurants and cafes	0.010
41202	47	Construction of domestic buildings	0.008
69202	45	Bookkeeping activities	0.008
64991	43	Security dealing on own account	0.008
58142	41	Publishing of consumer and business journals and periodicals	0.007
74209	40	Photographic activities **not elsewhere classified**	0.007
18129	40	Printing **n.e.c.**	0.007
Total			**0.849**

Figures

Figure 1: N. of postcodes per website distribution 2000-2012

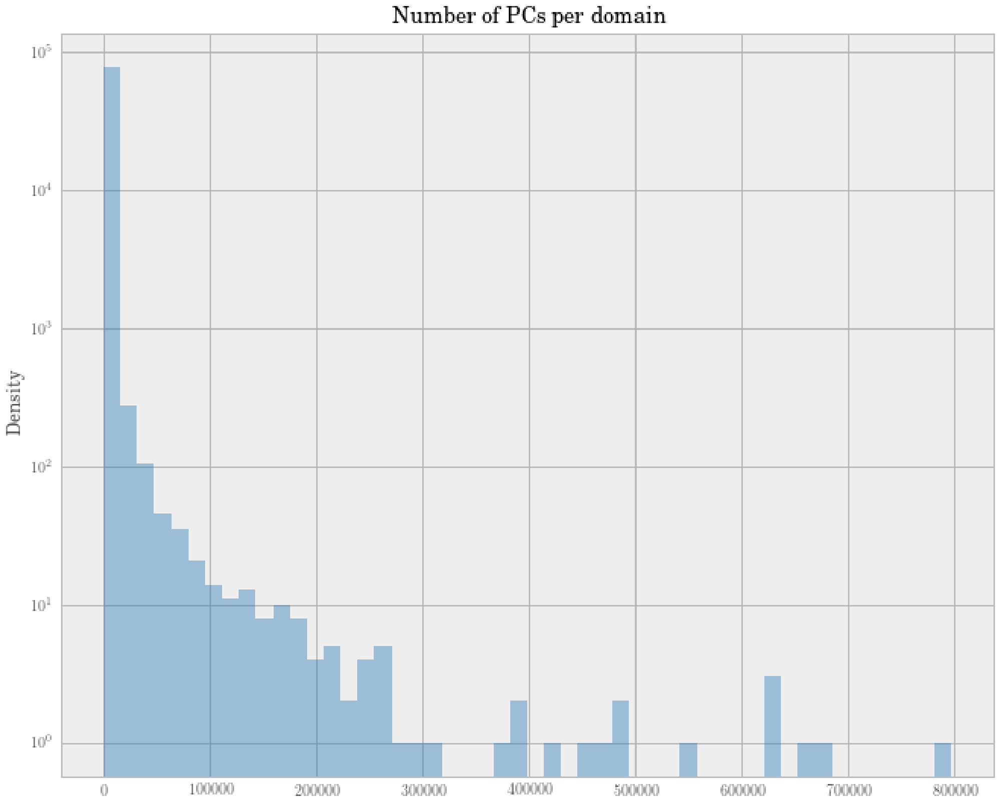
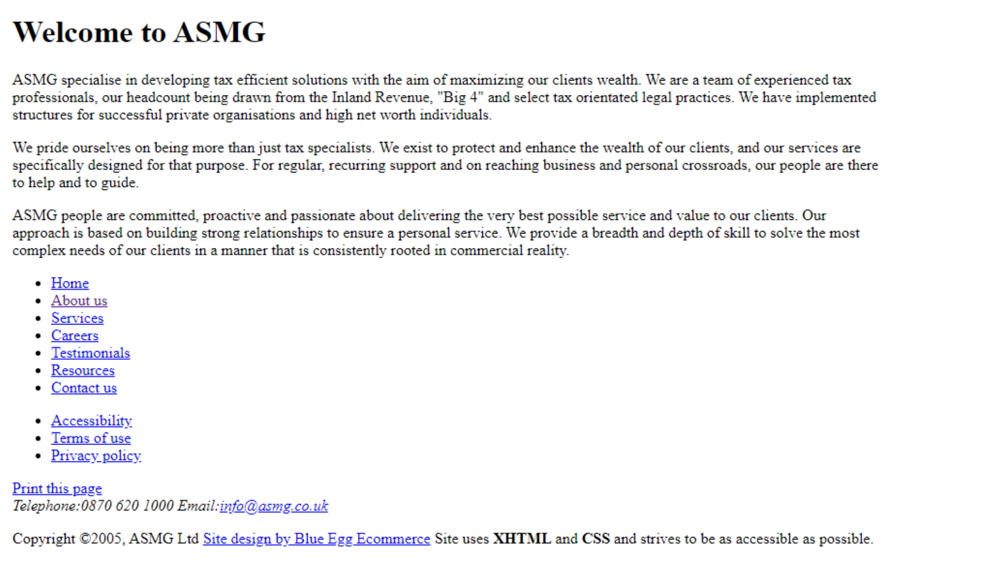
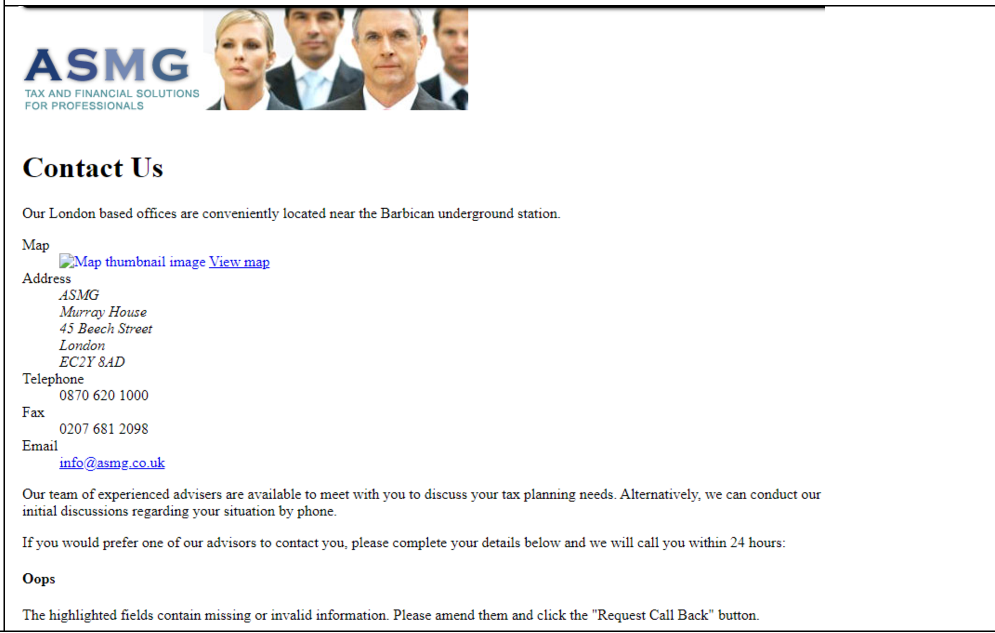
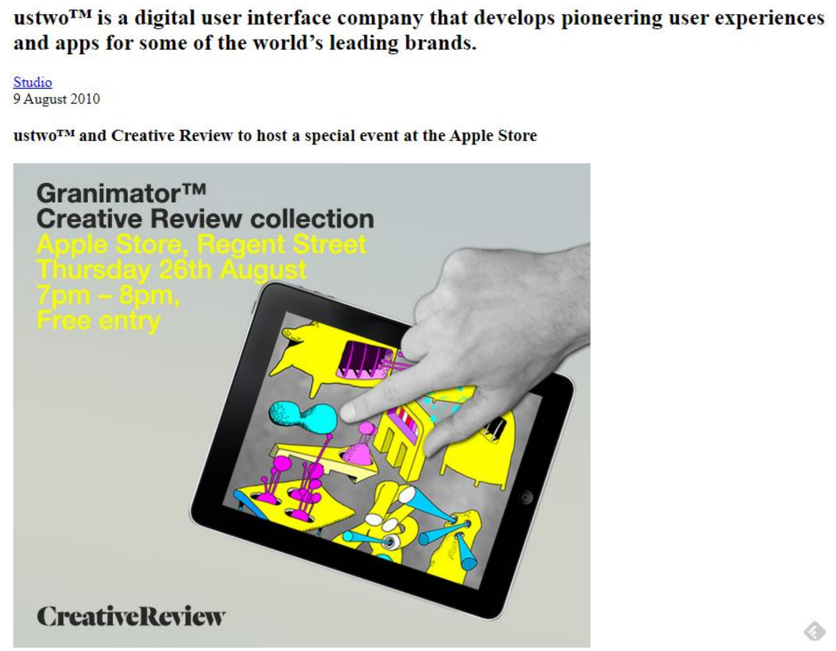
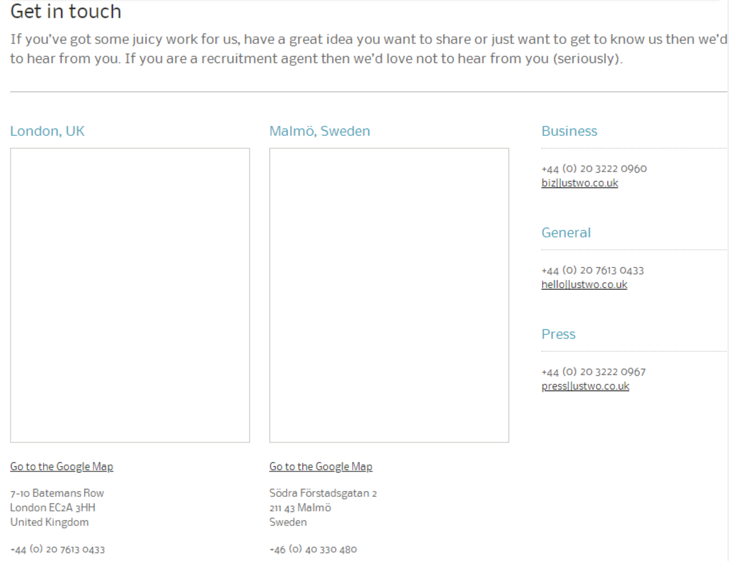
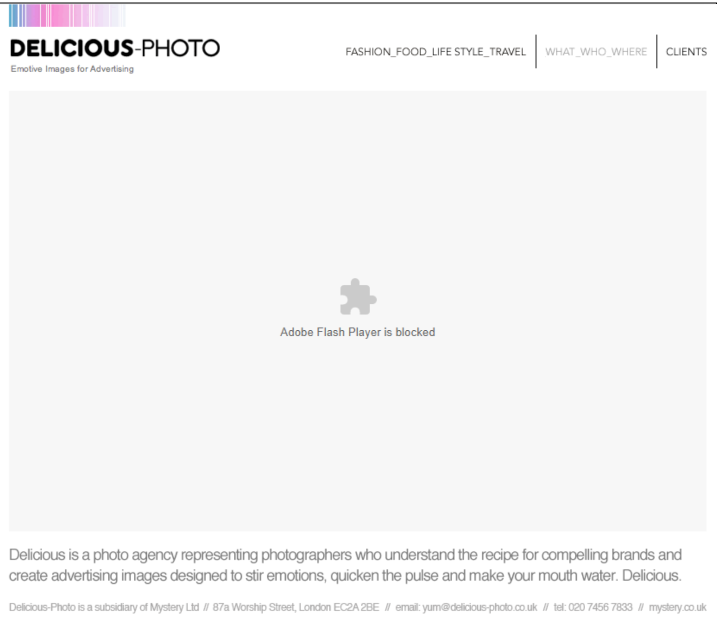


Figure 2: Snapshots of examples of websites with a unique postcode in Shoreditch

Home and contact us webpage snapshots		Source URL
1	 Welcome to ASMG ASMG specialise in developing tax efficient solutions with the aim of maximizing our clients wealth. We are a team of experienced tax professionals, our headcount being drawn from the Inland Revenue, "Big 4" and select tax orientated legal practices. We have implemented structures for successful private organisations and high net worth individuals. 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. 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 rooted in commercial reality. • Home • About us • Services • Careers • Testimonials • Resources • Contact us • Accessibility • Terms of use • Privacy policy Print this page *Telephone:0870 620 1000 Email:info@asmg.co.uk* Copyright ©2005, ASMG Ltd Site design by Blue Egg Ecommerce Site uses **XHTML** and **CSS** and strives to be as accessible as possible.	http://web.archive.org/web/20060621095920/http://www.asmg.co.uk:80/
	 ASMG TAX AND FINANCIAL SOLUTIONS FOR PROFESSIONALS Contact Us Our London based offices are conveniently located near the Barbican underground station. Map Map thumbnail image View map Address *ASMG* *Murray House* *45 Beech Street* *London* *EC2Y 8AD* Telephone 0870 620 1000 Fax 0207 681 2098 Email info@asmg.co.uk Our team of experienced advisers are available to meet with you to discuss your tax planning needs. Alternatively, we can conduct our initial discussions regarding your situation by phone. If you would prefer one of our advisors to contact you, please complete your details below and we will call you within 24 hours: **Oops** The highlighted fields contain missing or invalid information. Please amend them and click the "Request Call Back" button.	http://web.archive.org/web/20060720182815/http://www.asmg.co.uk/contact-us.htm

2	 **ustwo™ is a digital user interface company that develops pioneering user experiences and apps for some of the world's leading brands.** Studio 9 August 2010 **ustwo™ and Creative Review to host a special event at the Apple Store** Granimator™ Creative Review collection Apple Store, Regent Street Thursday 26th August 7pm – 8pm, Free entry CreativeReview	http://web.archive.org/web/20100813113036/http://www.ustwo.co.uk/
	 Get in touch 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 to hear from you. If you are a recruitment agent then we'd love not to hear from you (seriously). London, UK Go to the Google Map 7-10 Batemans Row London EC2A 3HH United Kingdom +44 (0) 20 7613 0433 Malmö, Sweden Go to the Google Map Södra Förstadsgatan 2 211 43 Malmö Sweden +46 (0) 40 330 480 Business +44 (0) 20 3222 0960 biz||ustwo.co.uk General +44 (0) 20 7613 0433 hello||ustwo.co.uk Press +44 (0) 20 3222 0967 press||ustwo.co.uk	http://web.archive.org/web/20100813113036/http://www.ustwo.co.uk/contact/
3	 DELICIOUS-PHOTO Emotive Images for Advertising FASHION_FOOD_LIFE STYLE_TRAVEL | WHAT_WHO_WHERE | CLIENTS Adobe Flash Player is blocked 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. Delicious-Photo is a subsidiary of Mystery Ltd // 87a Worship Street, London EC2A 2BE // email: yum@delicious-photo.co.uk // tel: 020 7456 7833 // mystery.co.uk	http://web.archive.org/web/20100722203026/http://www.delicious-photo.co.uk:80/

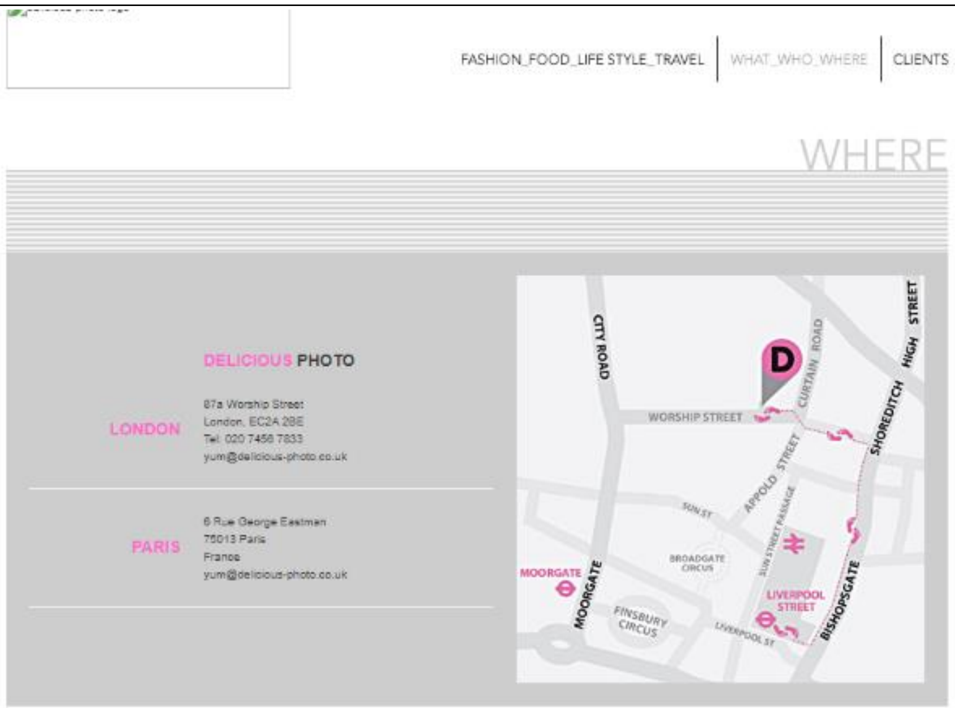
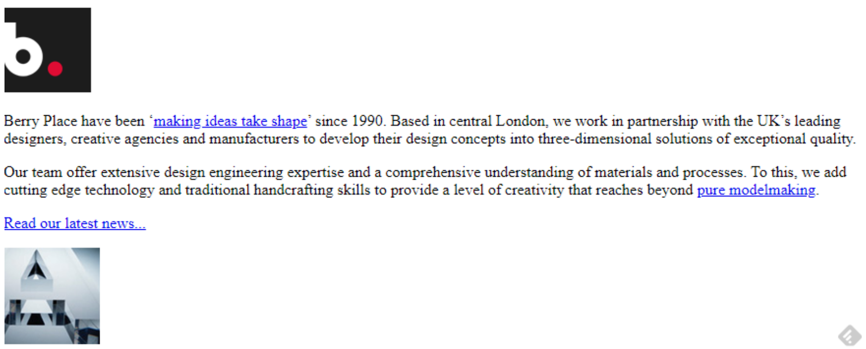
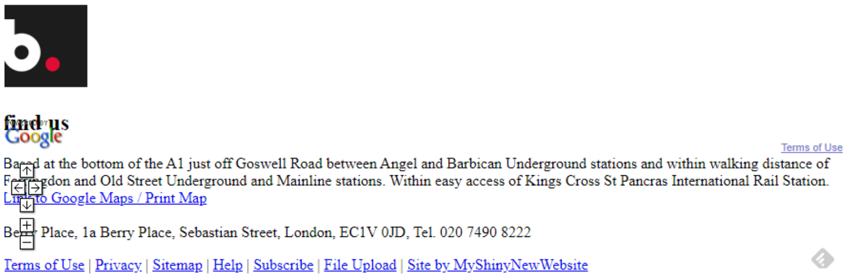
	 FASHION_FOOD_LIFE STYLE_TRAVEL | WHAT_WHO_WHERE | CLIENTS WHERE DELICIOUS PHOTO LONDON 87a Worship Street London, EC2A 2BE Tel: 020 7456 7833 yum@delicious-photo.co.uk PARIS 6 Rue George Eastman 75013 Paris France yum@delicious-photo.co.uk CITY ROAD WORSHIP STREET CURTAIN ROAD SHOREDITCH HIGH STREET APPOLD STREET SUN ST SUN STREET PASSAGE BROADGATE CIRCUS MOORGATE FINSBURY CIRCUS LIVERPOOL STREET LIVERPOOL ST BISHOPSGATE D	http://web.archive.org/web/20100722025731/http://www.delicious-photo.co.uk/who.shtml
4	 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. 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 pure modelmaking. Read our latest news...	http://web.archive.org/web/20100706150800/http://www.berry-
	 find us Based at the bottom of the A1 just off Goswell Road between Angel and Barbican Underground stations and within walking distance of Farringdon and Old Street Underground and Mainline stations. Within easy access of Kings Cross St Pancras International Rail Station. Link to Google Maps / Print Map Berry Place, 1a Berry Place, Sebastian Street, London, EC1V 0JD, Tel. 020 7490 8222 Terms of Use | Privacy | Sitemap | Help | Subscribe | File Upload | Site by MyShinyNewWebsite 	http://web.archive.org/web/20100706150749/http://www.ber-

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

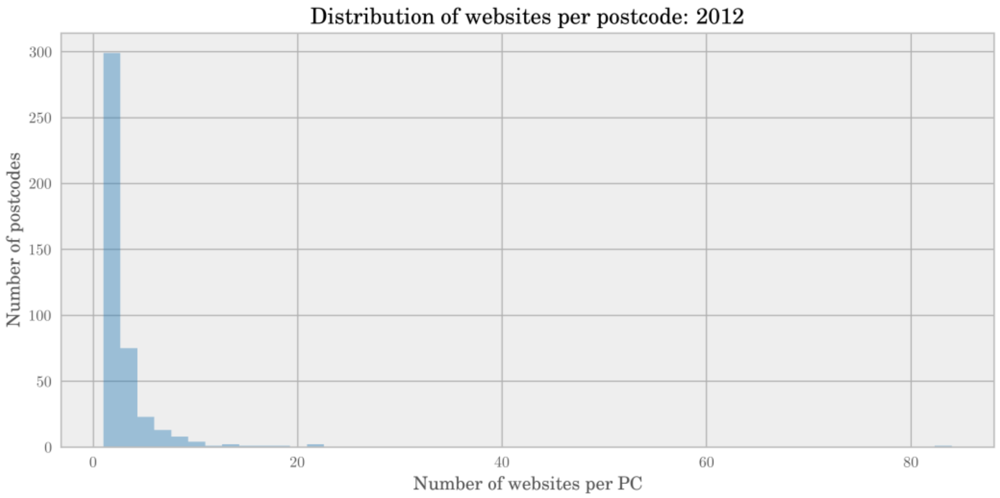
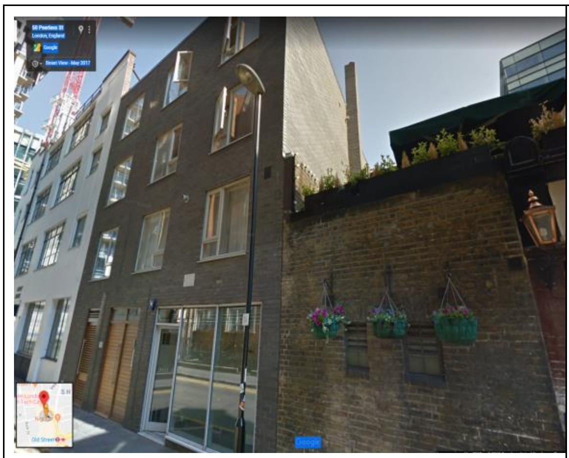
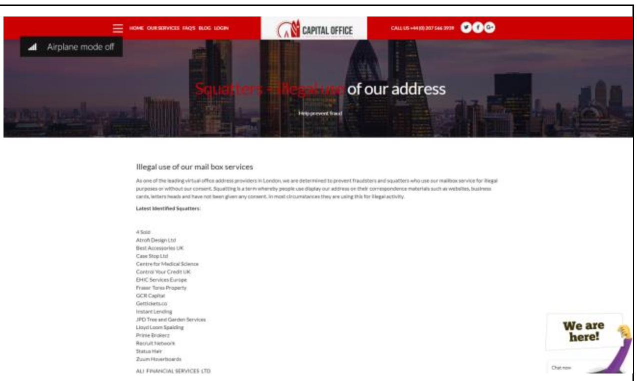


Figure 4: Digital ‘squatting’ in Shoreditch



A



HOME OUR SERVICES FAQ'S BLOG LOGIN

CAPITAL OFFICE

CALL US +44 (0) 207 566 3939

Airplane mode off

Squatters - Illegal use of our address

Help prevent fraud

Illegal use of our mail box services

As one of the leading virtual office address providers in London, we are determined to prevent fraudsters and squatters who use our mailbox service for illegal purposes or without our consent. Squatting is a term whereby people use display our address on their correspondence materials such as websites, business cards, letters heads and have not been given any consent. In most circumstances they are using this for illegal activity.

Latest Identified Squatters:

4 Solo
Atrofi Design Ltd
Best Accessories UK
Case Stop Ltd
Centre for Medical Science
Control Your Credit UK
EHIC Services Europe
Fraser Tores Property
GCR Capital
Gettickets.co
Instant Lending
JPD Tree and Garden Services
Lloyd Loom Spalding
Prime Brokerz
Recruit Network
Status Hair
Zuum Hoverboards

ALI FINANCIAL SERVICES LTD

We are here!

Chat now

B

Figure 5: Topic prevalence over time

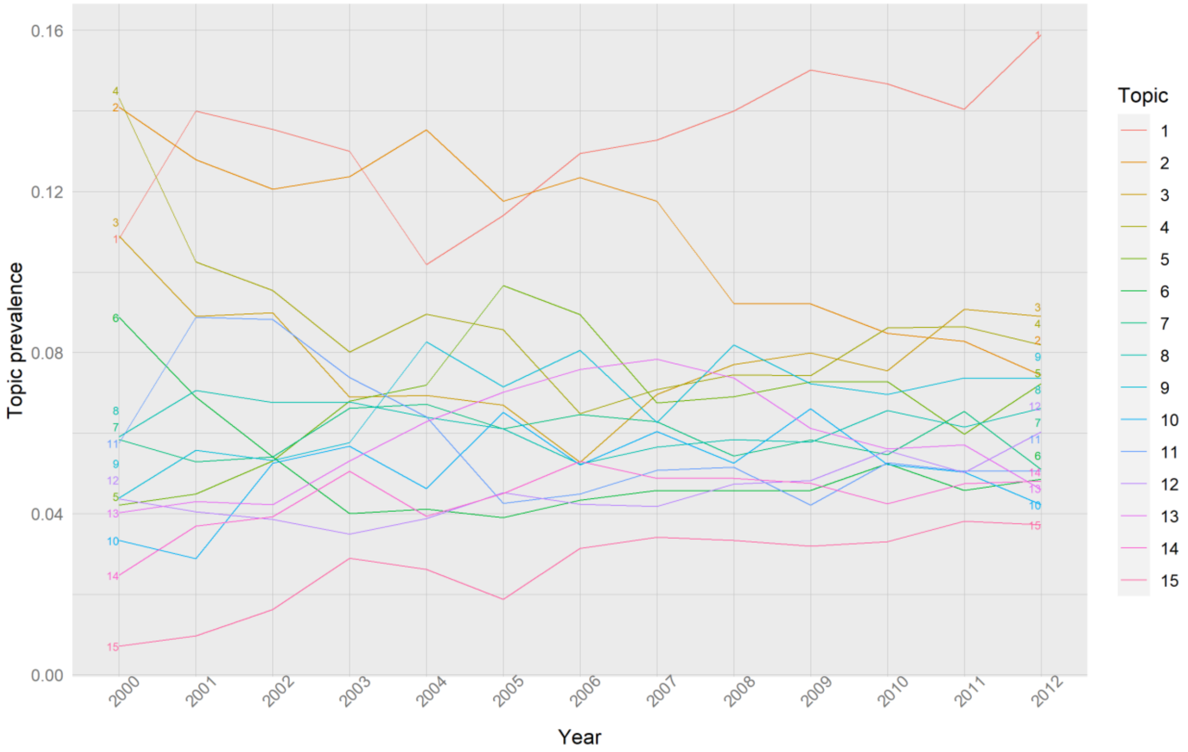
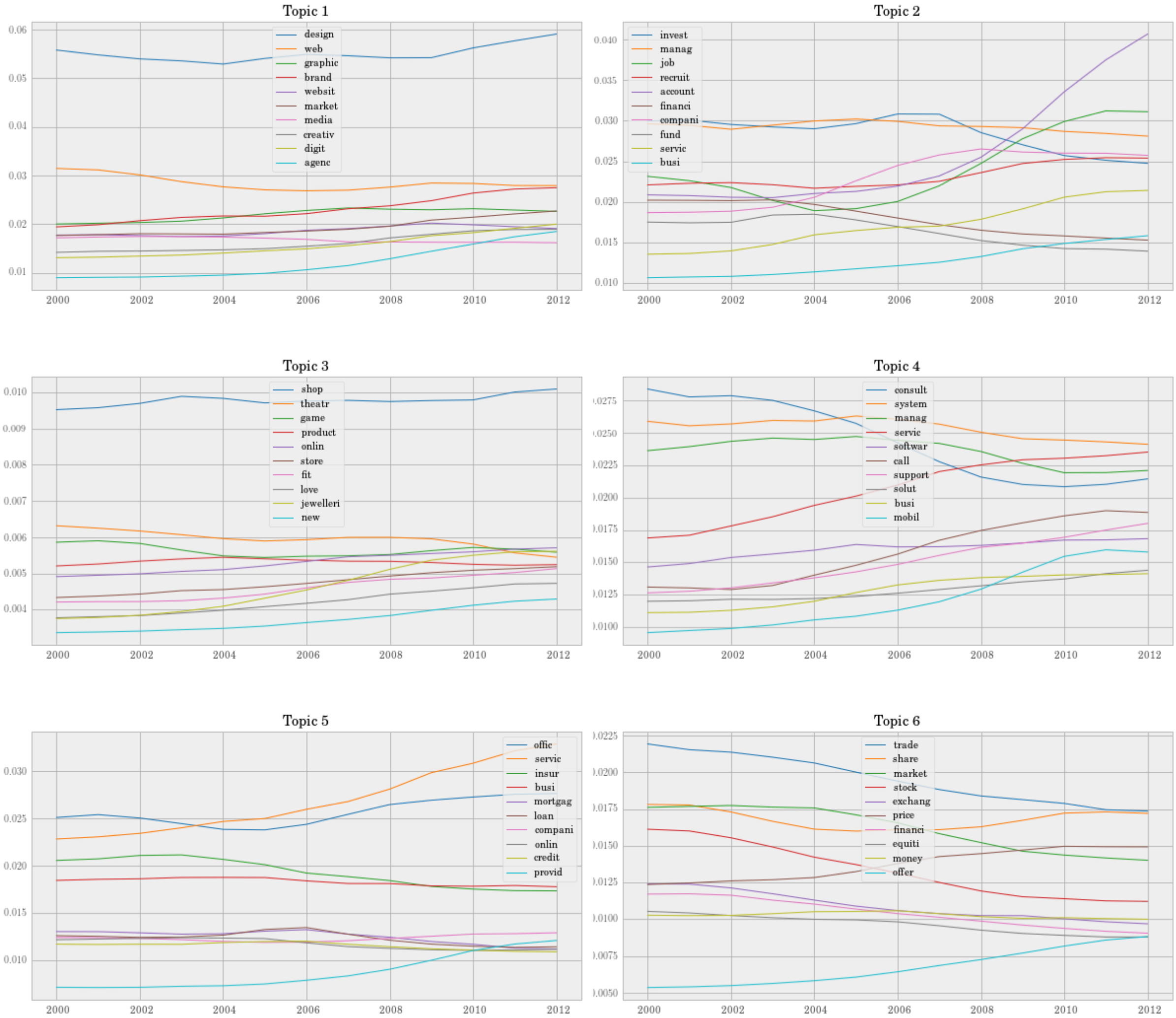
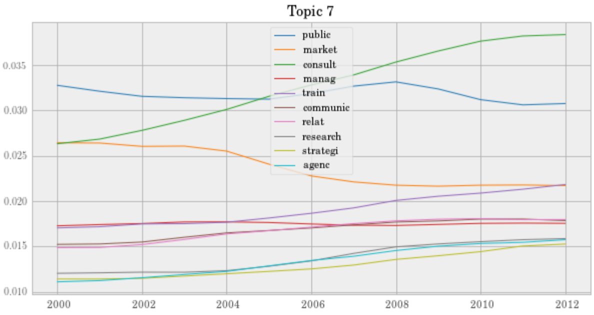
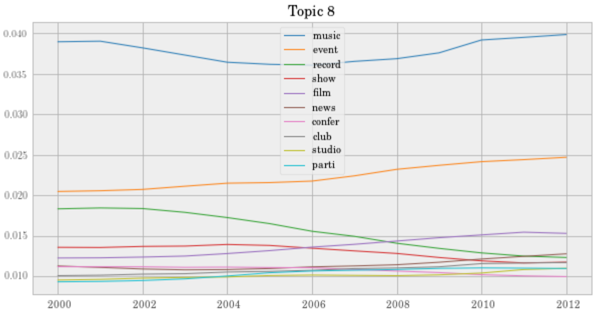


Figure 6: Dynamic term frequency per topic

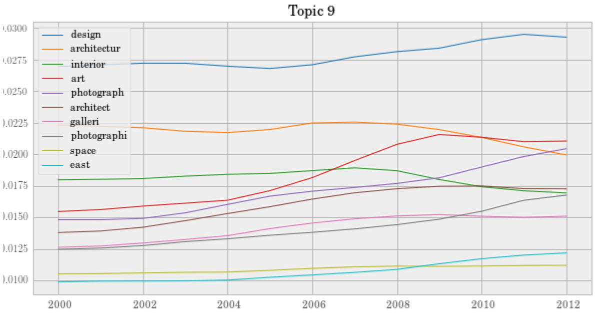




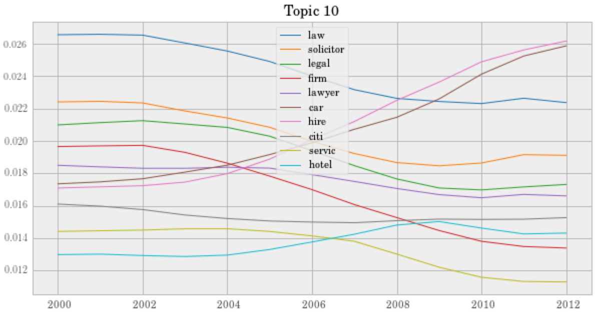
Topic 7



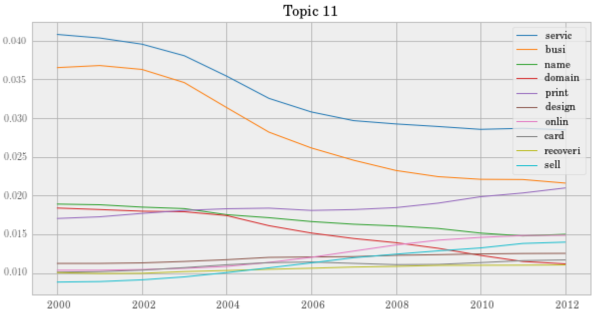
Topic 8



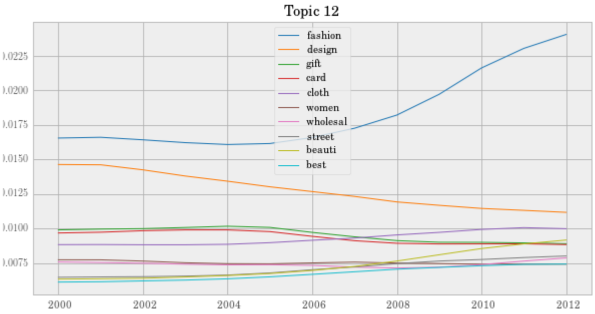
Topic 9



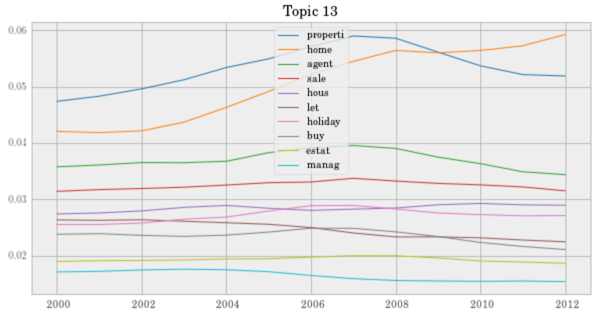
Topic 10



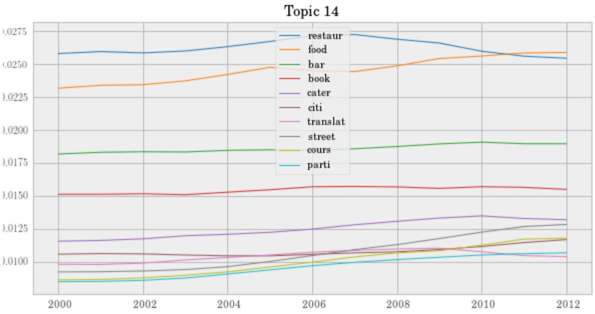
Topic 11



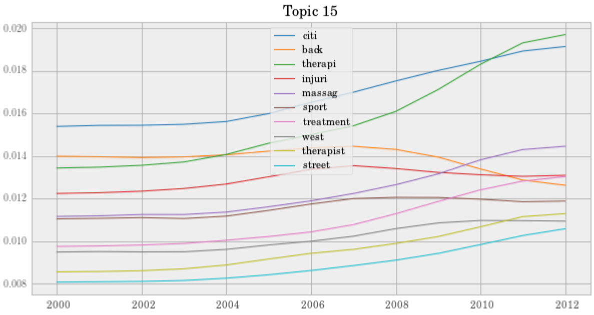
Topic 12



Topic 13

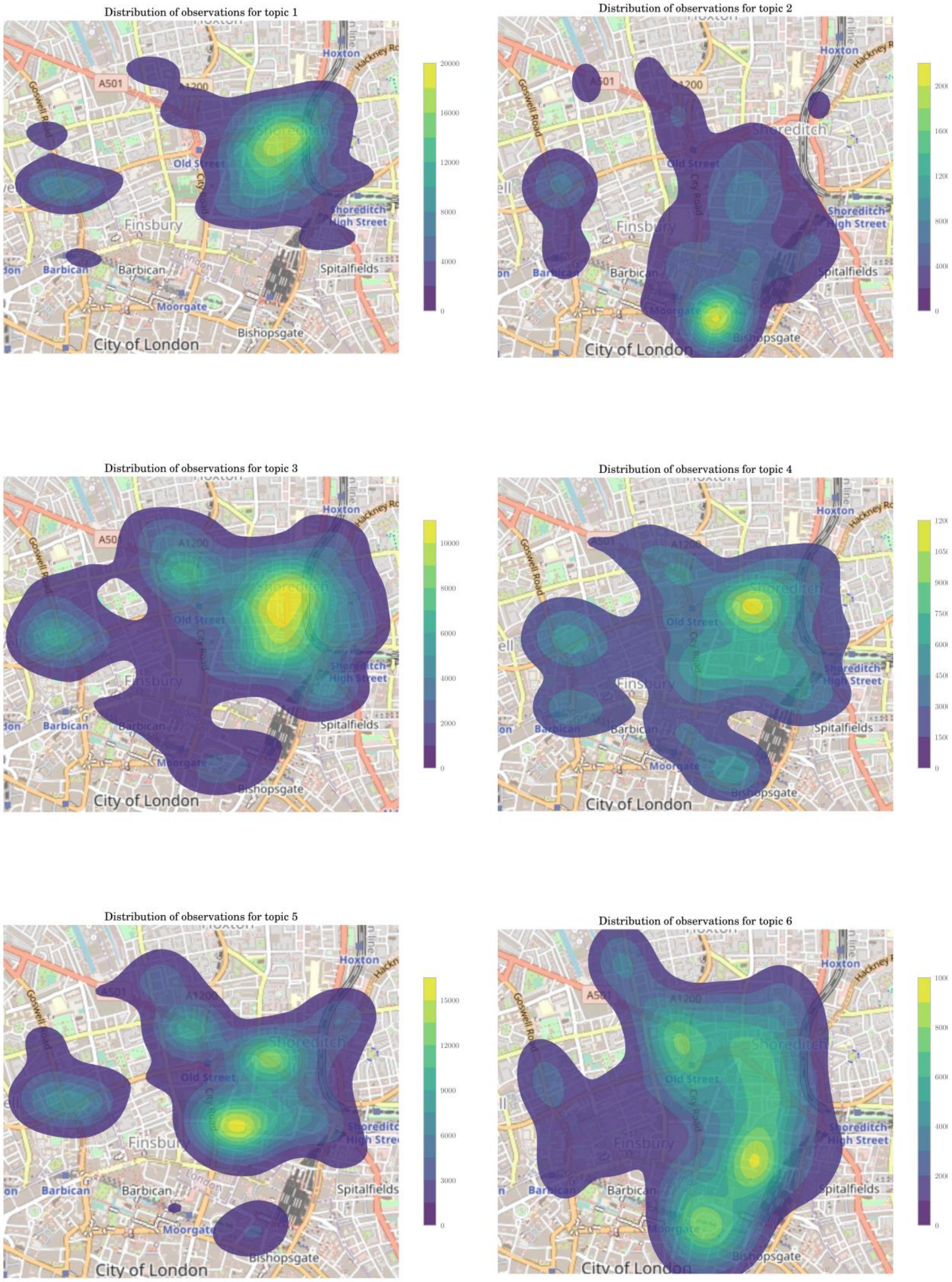


Topic 14

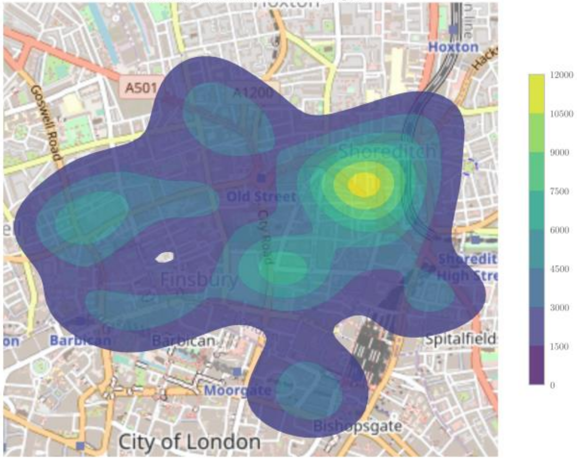


Topic 15

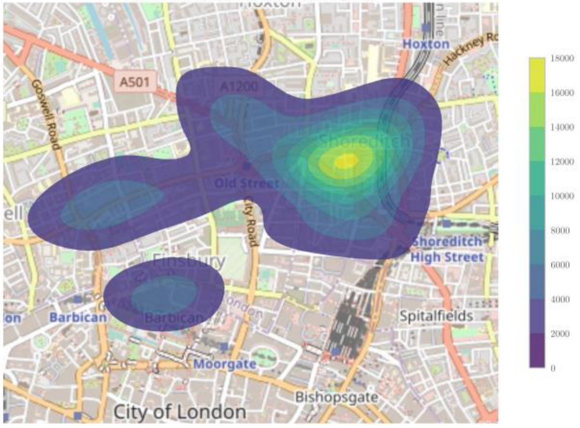
Figure 7: The spatial footprint of the different topics



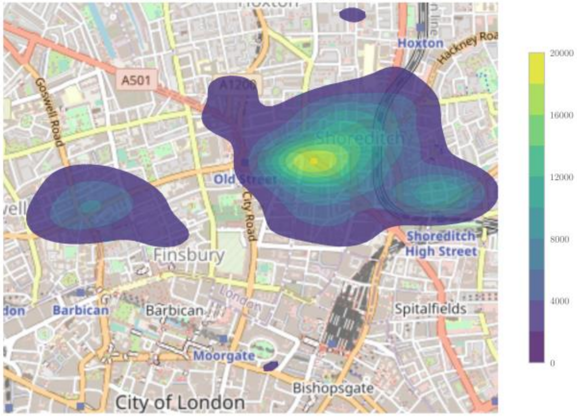
Distribution of observations for topic 7



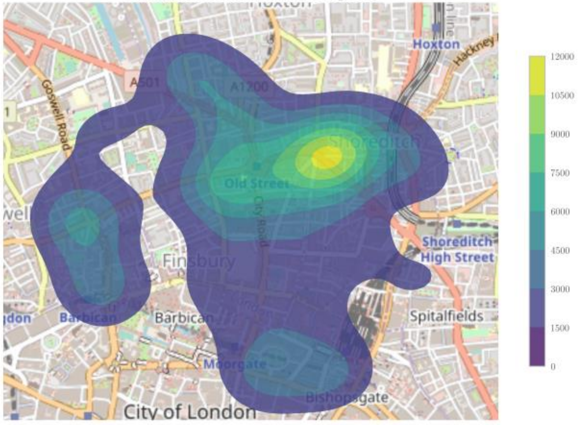
Distribution of observations for topic 8



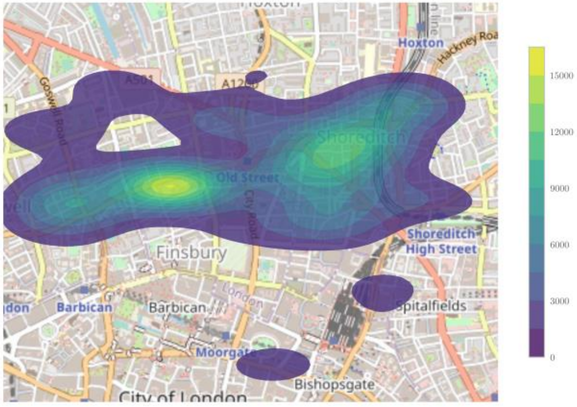
Distribution of observations for topic 9



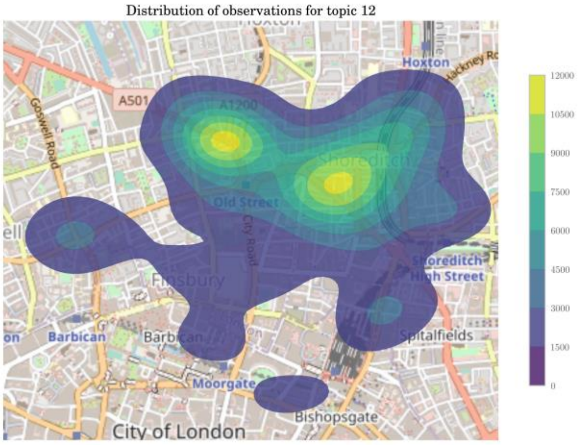
Distribution of observations for topic 10

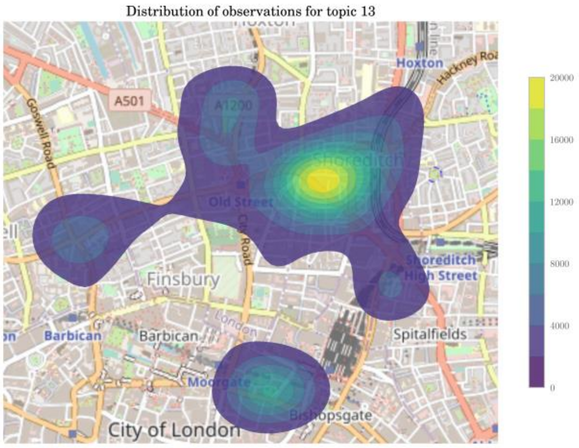


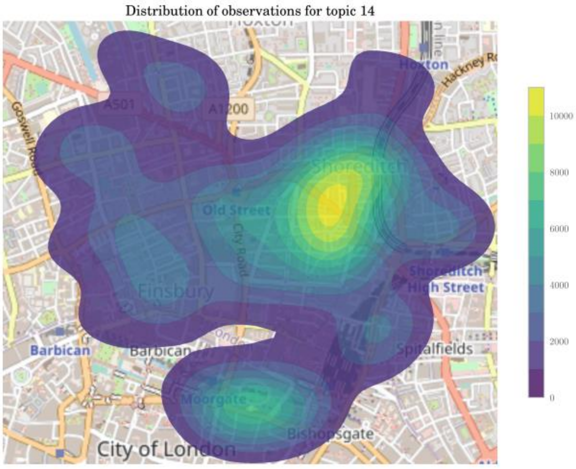
Distribution of observations for topic 11

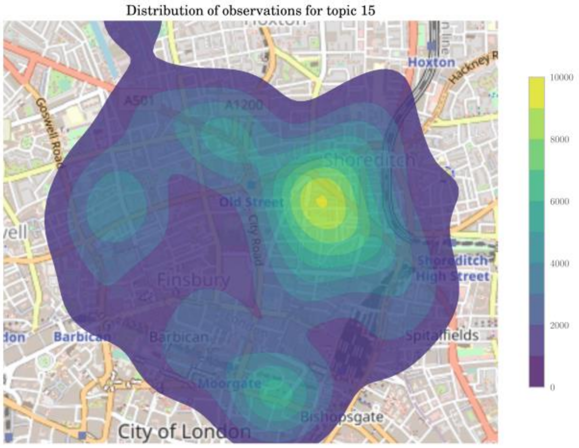


Distribution of observations for topic 12



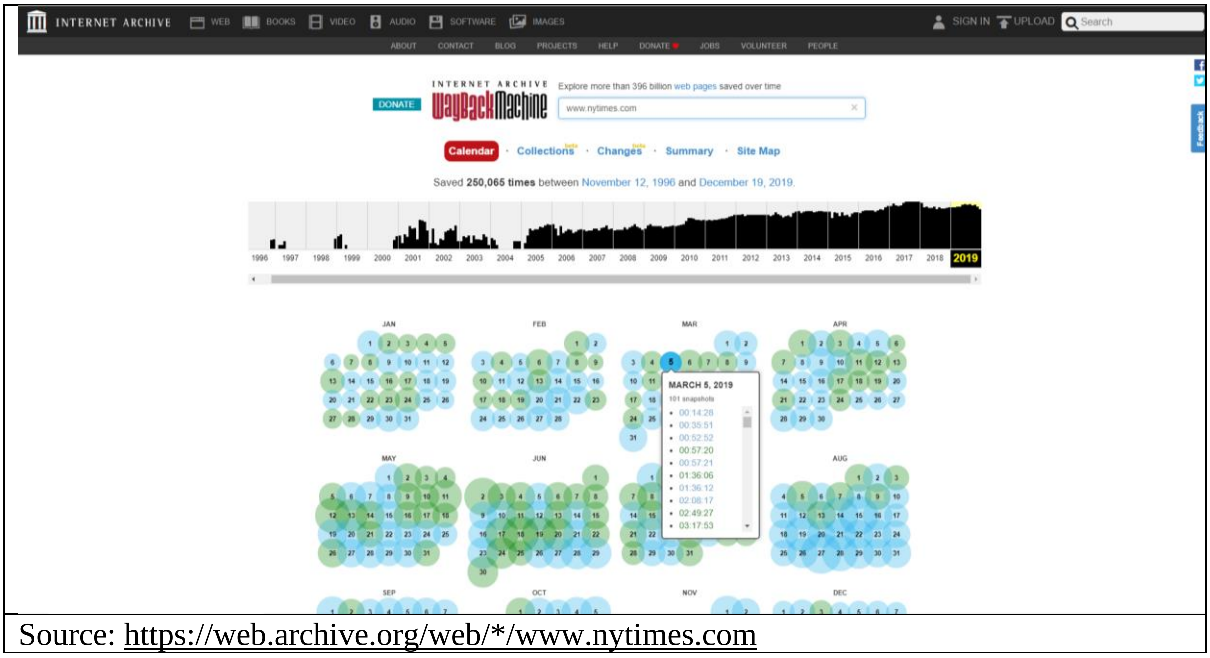

Distribution of observations for topic 13


Distribution of observations for topic 14


Distribution of observations for topic 15

Appendix material

Figure AX1: The Wayback Machine



Explore more than 396 billion web pages saved over time

www.nytimes.com

Saved **250,065 times** between November 12, 1996 and December 19, 2019.

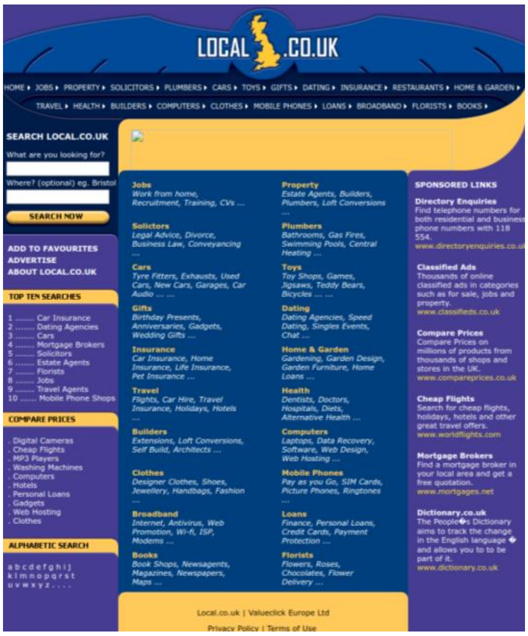
MARCH 5, 2019

101 snapshots

- 00:14:28
- 00:35:51
- 00:52:52
- 00:57:20
- 00:57:21
- 01:36:06
- 01:36:12
- 02:08:17
- 02:49:27
- 03:17:53

Source: https://web.archive.org/web/*/www.nytimes.com

Figure AX2: Snapshots of listings websites


LOCAL.CO.UK
HOME ▸ JOBS ▸ PROPERTY ▸ SOLICITORS ▸ PLUMBERS ▸ CARS ▸ TOYS ▸ GIFTS ▸ DATING ▸ INSURANCE ▸ RESTAURANTS ▸ HOME & GARDEN ▸
TRAVEL ▸ HEALTH ▸ BUILDERS ▸ COMPUTERS ▸ CLOTHES ▸ MOBILE PHONES ▸ LOANS ▸ BROADBAND ▸ FLORISTS ▸ BOOKS ▸
SEARCH NOW
ADD TO FAVOURITES
ADVERTISE
ABOUT LOCAL.CO.UK

SEARCH LOCAL.CO.UK

What are you looking for?

Where? (optional) eg. Bristol

TOP TEN SEARCHES

1 .... Car Insurance
2 .... Dating Agencies
3 .... Cars
4 .... Mortgage Brokers
5 .... Solicitors
6 .... Estate Agents
7 .... Florists
8 .... Jobs
9 .... Travel Agents
10 .... Mobile Phone Shops

COMPARE PRICES

. Digital Cameras
. Cheap Flights
. MP3 Players
. Washing Machines
. Computers
. Hotels
. Personal Loans
. Gadgets
. Web Hosting
. Clothes

ALPHABETIC SEARCH

a b c d e f g h i j
k l m n o p q r s t
u v w x y z . . . .

Jobs *Work from home, Recruitment, Training, CVs ...*

Solicitors *Legal Advice, Divorce, Business Law, Conveyancing ...*

Cars *Tyre Fitters, Exhausts, Used Cars, New Cars, Garages, Car Audio ....*

Gifts *Birthday Presents, Anniversaries, Gadgets, Wedding Gifts ...*

Insurance *Car Insurance, Home Insurance, Life Insurance, Pet Insurance ...*

Travel *Flights, Car Hire, Travel Insurance, Holidays, Hotels ...*

Builders *Extensions, Loft Conversions, Self Build, Architects ...*

Clothes *Designer Clothes, Shoes, Jewellery, Handbags, Fashion ...*

Broadband *Internet, Antivirus, Web Promotion, Wi-fi, ISP, Modems ...*

Books *Book Shops, Newsagents, Magazines, Newspapers, Maps ...*

Property *Estate Agents, Builders, Plumbers, Loft Conversions ...*

Plumbers *Bathrooms, Gas Fires, Swimming Pools, Central Heating ...*

Toys *Toy Shops, Games, Jigsaws, Teddy Bears, Bicycles ....*

Dating *Dating Agencies, Speed Dating, Singles Events, Chat ...*

Home & Garden *Gardening, Garden Design, Garden Furniture, Home Loans ...*

Health *Dentists, Doctors, Hospitals, Diets, Alternative Health ...*

Computers *Laptops, Data Recovery, Software, Web Design, Web Hosting ...*

Mobile Phones *Pay as you Go, SIM Cards, Picture Phones, Ringtones ...*

Loans *Finance, Personal Loans, Credit Cards, Payment Protection ...*

Florists *Flowers, Roses, Chocolates, Flower Delivery ...*

SPONSORED LINKS

Directory Enquiries Find telephone numbers for both residential and business phone numbers with 118 554. www.directoryenquiries.co.uk

Classified Ads Thousands of online classified ads in categories such as for sale, jobs and property. www.classifieds.co.uk

Compare Prices Compare Prices on millions of products from thousands of shops and stores in the UK. www.compareprices.co.uk

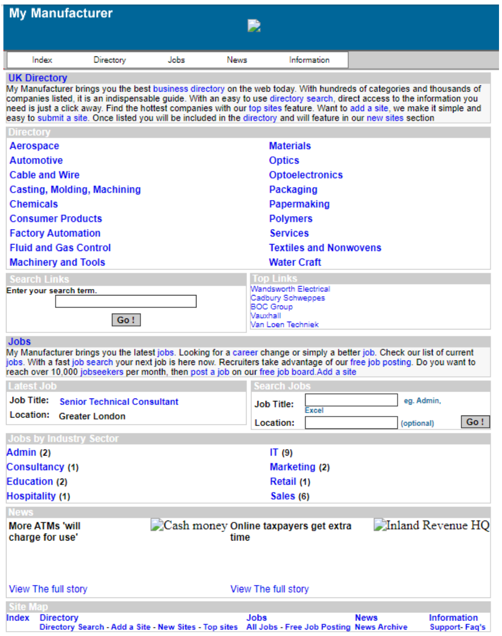
Cheap Flights Search for cheap flights, holidays, hotels and other great travel offers. www.worldflights.com

Mortgage Brokers Find a mortgage broker in your local area and get a free quotation. www.mortgages.net

Dictionary.co.uk The People�s Dictionary aims to track the change in the English language � and allows you to to be part of it. www.dictionary.co.uk

Local.co.uk | Valueclick Europe Ltd

Privacy Policy | Terms of Use



UK Directory

My Manufacturer brings you the best business directory on the web today. With hundreds of categories and thousands of companies listed, it is an indispensable guide. With an easy to use directory search, direct access to the information you need is just a click away. Find the hottest companies with our top sites feature. Want to add a site, we make it simple and easy to submit a site. Once listed you will be included in the directory and will feature in our new sites section

Directory

Aerospace	Materials
Automotive	Optics
Cable and Wire	Optoelectronics
Casting, Molding, Machining	Packaging
Chemicals	Papermaking
Consumer Products	Polymers
Factory Automation	Services
Fluid and Gas Control	Textiles and Nonwovens
Machinery and Tools	Water Craft

Search Links

Enter your search term.

Top Links

Wandsworth Electrical
Cadbury Schweppes
BOC Group
Vauxhall
Van Loen Techniek

Jobs

My Manufacturer brings you the latest jobs. Looking for a career change or simply a better job. Check our list of current jobs. With a fast job search your next job is here now. Recruiters take advantage of our free job posting. Do you want to reach over 10,000 jobseekers per month, then post a job on our free job board.Add a site

Latest Job

Job Title: Senior Technical Consultant
Location: Greater London

Search Jobs

Job Title: eg. Admin, Excel
Location: (optional)

Jobs by Industry Sector

Admin (2)	IT (9)
Consultancy (1)	Marketing (2)
Education (2)	Retail (1)
Hospitality (1)	Sales (6)

News

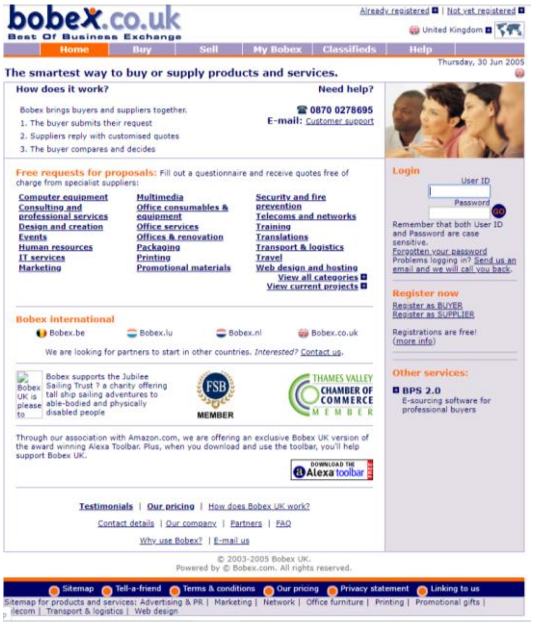
More ATMs 'will charge for use' — View The full story

Cash money Online taxpayers get extra time — View The full story

Inland Revenue HQ

Site Map

Index	Directory	Jobs	News	Information
	Directory Search - Add a Site - New Sites - Top sites	All Jobs - Free Job Posting	News Archive	Support- Faq's



Thursday, 30 Jun 2005

The smartest way to buy or supply products and services.

How does it work?

Bobex brings buyers and suppliers together.

1. The buyer submits their request
2. Suppliers reply with customised quotes
3. The buyer compares and decides

Need help?

☎ 0870 0278695
E-mail: Customer support

Free requests for proposals: Fill out a questionnaire and receive quotes free of charge from specialist suppliers:

Computer equipment	Multimedia	Security and fire prevention
Consulting and professional services	Office consumables & equipment	Telecoms and networks
Design and creation	Office services	Training
Events	Offices & renovation	Translations
Human resources	Packaging	Transport & logistics
IT services	Printing	Travel
Marketing	Promotional materials	Web design and hosting

View all categories
View current projects

Bobex international

Bobex.be Bobex.lu Bobex.nl Bobex.co.uk

We are looking for partners to start in other countries. *Interested?* Contact us.

Bobex UK is please to

Bobex supports the Jubilee Sailing Trust ? a charity offering tall ship sailing adventures to able-bodied and physically disabled people

Through our association with Amazon.com, we are offering an exclusive Bobex UK version of the award winning Alexa Toolbar. Plus, when you download and use the toolbar, you'll help support Bobex UK.

Testimonials | Our pricing | How does Bobex UK work?

Contact details | Our company | Partners | FAQ

Why use Bobex? | E-mail us

Login

User ID

Password

Remember that both User ID and Password are case sensitive.
Forgotten your password
Problems logging in? Send us an email and we will call you back.

Register now

Register as BUYER
Register as SUPPLIER

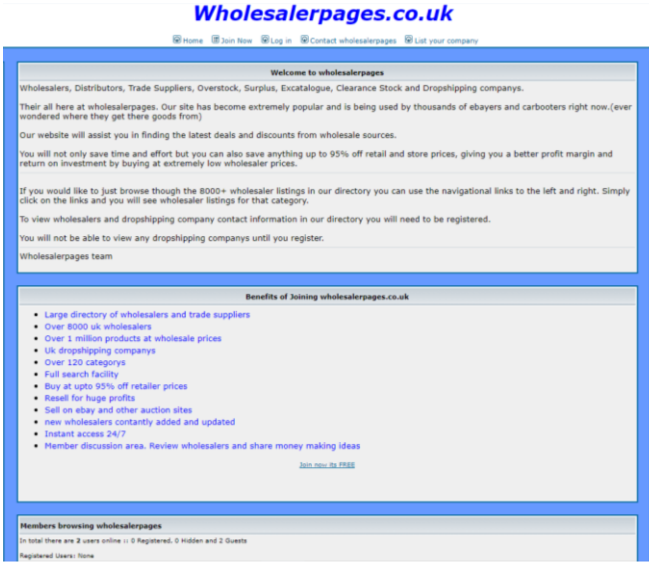
Registrations are free! (more info)

Other services:

BPS 2.0
E-sourcing software for professional buyers

© 2003-2005 Bobex UK.
Powered by © Bobex.com. All rights reserved.

Sitemap for products and services: Advertising & PR | Marketing | Network | Office furniture | Printing | Promotional gifts | lecom | Transport & logistics | Web design



Welcome to wholesalerpages

Wholesalers, Distributors, Trade Suppliers, Overstock, Surplus, Excatalogue, Clearance Stock and Dropshipping companys.

Their all here at wholesalerpages. Our site has become extremely popular and is being used by thousands of ebayers and carbooters right now.(ever wondered where they get there goods from)

Our website will assist you in finding the latest deals and discounts from wholesale sources.

You will not only save time and effort but you can also save anything up to 95% off retail and store prices, giving you a better profit margin and return on investment by buying at extremely low wholesaler prices.

If you would like to just browse though the 8000+ wholesaler listings in our directory you can use the navigational links to the left and right. Simply click on the links and you will see wholesaler listings for that category.

To view wholesalers and dropshipping company contact information in our directory you will need to be registered.

You will not be able to view any dropshipping companys until you register.

Wholesalerpages team

Benefits of Joining wholesalerpages.co.uk

- Large directory of wholesalers and trade suppliers
- Over 8000 uk wholesalers
- Over 1 million products at wholesale prices
- Uk dropshipping companys
- Over 120 categorys
- Full search facility
- Buy at upto 95% off retailer prices
- Resell for huge profits
- Sell on ebay and other auction sites
- new wholesalers contantly added and updated
- Instant access 24/7
- Member discussion area. Review wholesalers and share money making ideas

Join now its FREE

Members browsing wholesalerpages

In total there are 2 users online :: 0 Registered, 0 Hidden and 2 Guests

Registered Users: None

Source URLs:

http://web.archive.org/web/20050601023734/http://www.local.co.uk/

http://web.archive.org/web/20050420234401/http://www.mymanufacturer.co.uk:80/

http://web.archive.org/web/20050630030539/http://www.bobex.co.uk/bobexuk/control/home

http://web.archive.org/web/20060702215651/http://www.wholesalerpages.co.uk:80/

Figure AX3: Distribution of postcode distances from Shoreditch (Old Street roundabout) for websites with ≤ 11 postcodes

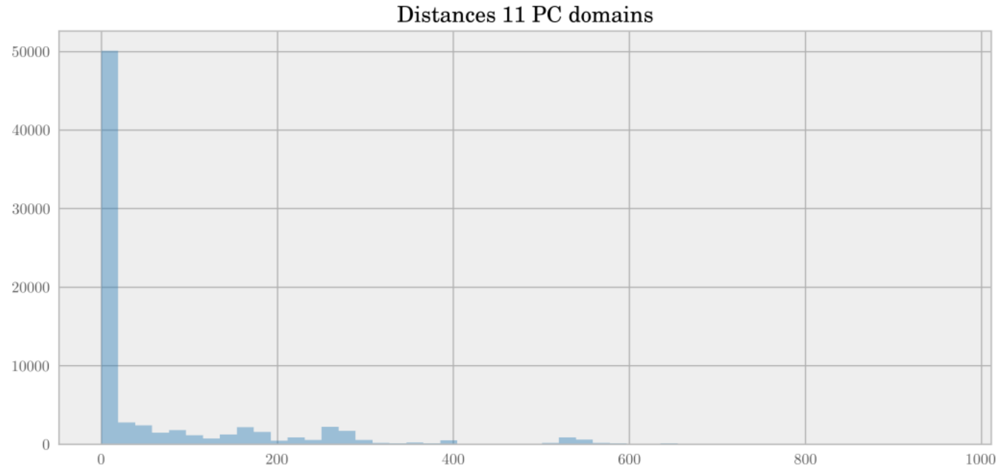


Table AX1: Topics based on the extended subset

Topic	Label	Term frequency (%)	20 most frequent terms
1	business technology services	10.9%	servic, busi, network, provid, system, solut, data, call, manag, mobil, support, phone, softwar, secur, host, centr, internet, server, comput, free
2	digital media services	9.8%	consult, develop, design, manag, web, websit, market, engin, train, site, busi, servic, project, strategi, commerc, search, profession, social, build, compani
3	arts and craft	9.5%	design, gift, cloth, shop, fashion, accessori, furnitur, clean, hand, offic, jewelleri, store, wed, onlin, made, bag, product, bespok, manufactur, wholesal
4	financial services	7.9%	account, financi, financ, servic, trade, market, invest, busi, bank, manag, compani, share, stock, tax, corpor, price, exchang, advic, equiti, report
5	wellbeing	7.2%	music, care, health, record, doctor, well, medic, blog, rock, nurs, home, children, hous, natur, live, soul, shirt, social, peopl, danc
6	real estate	6.9%	properti, home, sale, law, hous, servic, agent, solicitor, estat, buy, commerci, legal, let, busi, rent, lawyer, compani, sell, flat, offic
7	digital media	6.8%	design, brand, digit, agenc, print, graphic, creativ, media, market, photographi, art, photograph, web, advertis, product, studio, onlin, communic, model, illustr
8	hospitality industry	6.7%	news, club, bar, parti, event, restaur, food, review, magazin, venu, music, shop, sport, wed, danc, book, corpor, guid, night, drink
9	travel	5.9%	insur, holiday, travel, car, mortgag, cheap, discount, hotel, onlin, offer, deal, rate, loan, broker, life, quot, person, low, home, hire
10	investment services	5.7%	invest, fund, manag, chariti, pension, trust, independ, financi, advic, servic, publish, market, group, investor, compani, asset, advis, save, capit, money
11	architecture services	5.7%	street, design, build, architectur, architect, hous, interior, east, offic, sustain, ton, space, plan, green, art, urban, park, construct, pub, citi
12	education	5.4%	cours, train, class, school, learn, student, test, lesson, electr, remov, servic, fire, water, system, colleg, certif, control, oil, languag, energi
13	recruitment services	4.2%	job, recruit, career, agenc, employ, servic, vacanc, work, bike, search, sale, new, car, cycl, use, execut, manag, graduat, resourc, citi
14	medical and wellbeing services	3.9%	treatment, therapi, clinic, citi, pain, massag, dentist, health, stress, therapist, counsel, street, injuri, back, cosmet, sport, dental, problem, depress, bodi
15	performing arts	3.6%	art, event, music, artist, exhibit, theatr, confer, perform, show, hire, road, film, galleri, download, includ, festiv, product, cinema, entertain, ferri

Note: terms are stemmed