Modelling clusters from the ground up: a web data approach

Christoph Stich, Emmanouil Tranos & Max Nathan

University of Bristol, Alan Turing Institute
e.tranos@bristol.ac.uk, @EmmanouilTranos

Introduction

A lot of theoretical and empirical work on clusters

- Urban econ & econ geog (micro-foundations, MAR vs. Jacobs)
- Evolutionary perspectives (path dependency)
- Globalisation scholars (global value chains / production networks)
- Temporary / online collaboration tools

Some basic questions still unresolved

- E.g. MAR vs. Jacobs; feasibility of cluster policy; appropriate policy mix
- Hard-to-fix empirical challenges:
- Data / economic activity scale mismatch (MAUP)
- SIC lag behind real-world industrial evolution
- Defining clusters based on industries instead of activities (e.g. fintech or cleantech)
- Tradeoffs between richness and reach of data

Contribution

- A new approach to analyse clusters from the bottom up
- Over time
- Web data and data science methods
- Empirical cluster research challenges (MAUP, SIC, richness/reach tradeoff)
- Shoreditch: East London Tech City aka Silicon Roundabout

Empirical strategy

Web data

- *Archived*, commercial websites 2000-2012
- Geolocated in Shoreditch, London
- Flexible approach in exploring economic activities and their dynamics
- Readily available, cheap to obtain and extensive in terms of the theme and population coverage
- Under-explored, public domain data

Methods

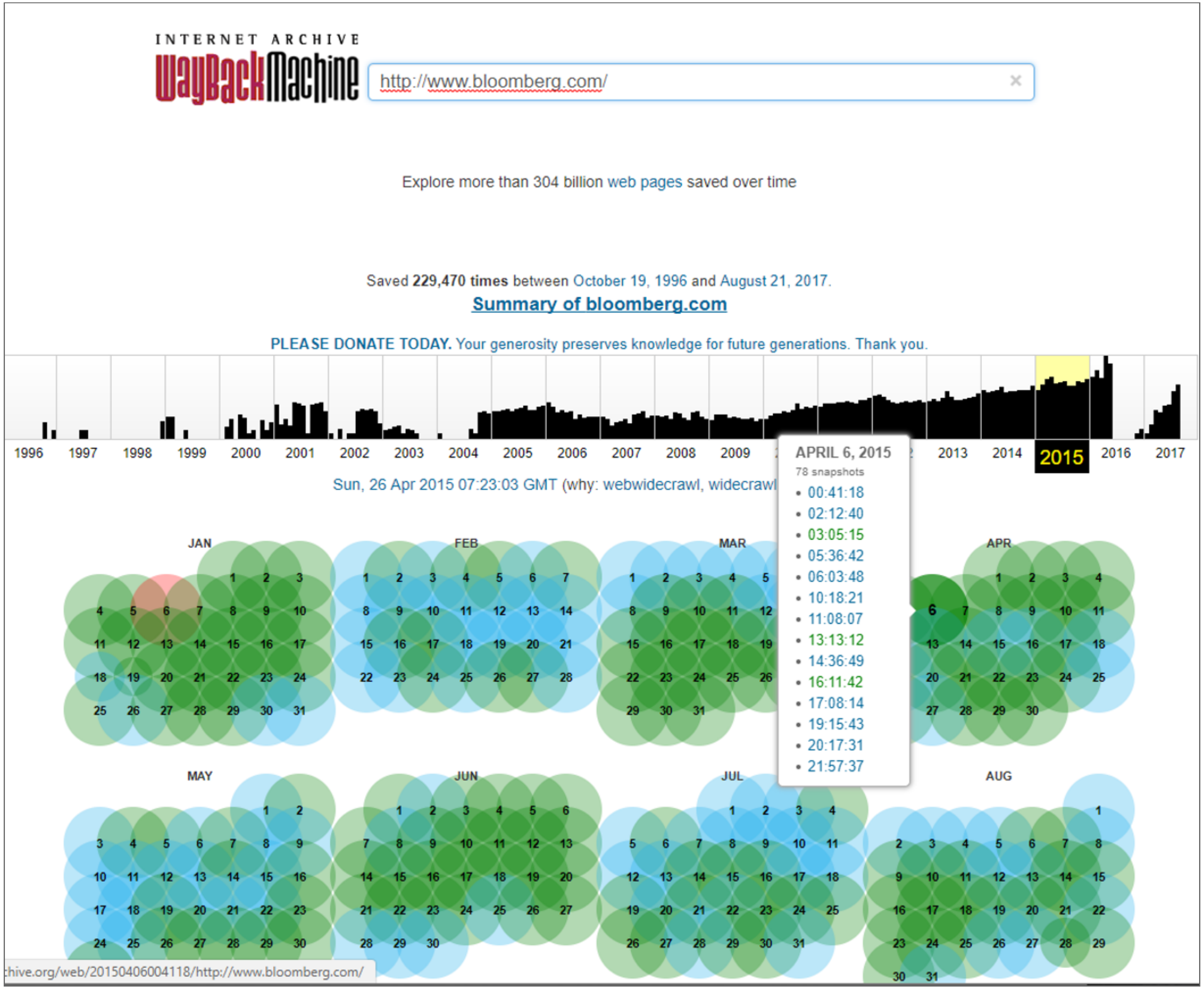
- Data cleaning: create a subset of business websites in Shoreditch
- Spatial analysis for interesting outliers
- Topic modelling: Latent Dirichlet Allocation (LDA)
- Extract bundles of economic activities (topics)
- Extract the key terms of every topic
- Bottom up classification *vs*. top-down SIC

Data

Web data: The Internet Archive

- The largest archive of webpages in the world
- 273 billion webpages from over 361 million websites, 15 petabytes of storage (1996 -)
- A web crawler starts with a list of URLs (a seed list) to crawl and downloads a copy of their content
- Using the hyperlinks included in the crawled URLs, new URLs are identified and crawled (snowball sampling)
- Time-stamp

Web data: The Internet Archive



INTERNET ARCHIVE WaybackMachine

http://www.bloomberg.com/

Explore more than 304 billion web pages saved over time

Saved **229,470 times** between October 19, 1996 and August 21, 2017.

Summary of bloomberg.com

PLEASE DONATE TODAY. Your generosity preserves knowledge for future generations. Thank you.

Sun, 26 Apr 2015 07:23:03 GMT (why: webwidecrawl, widecrawl

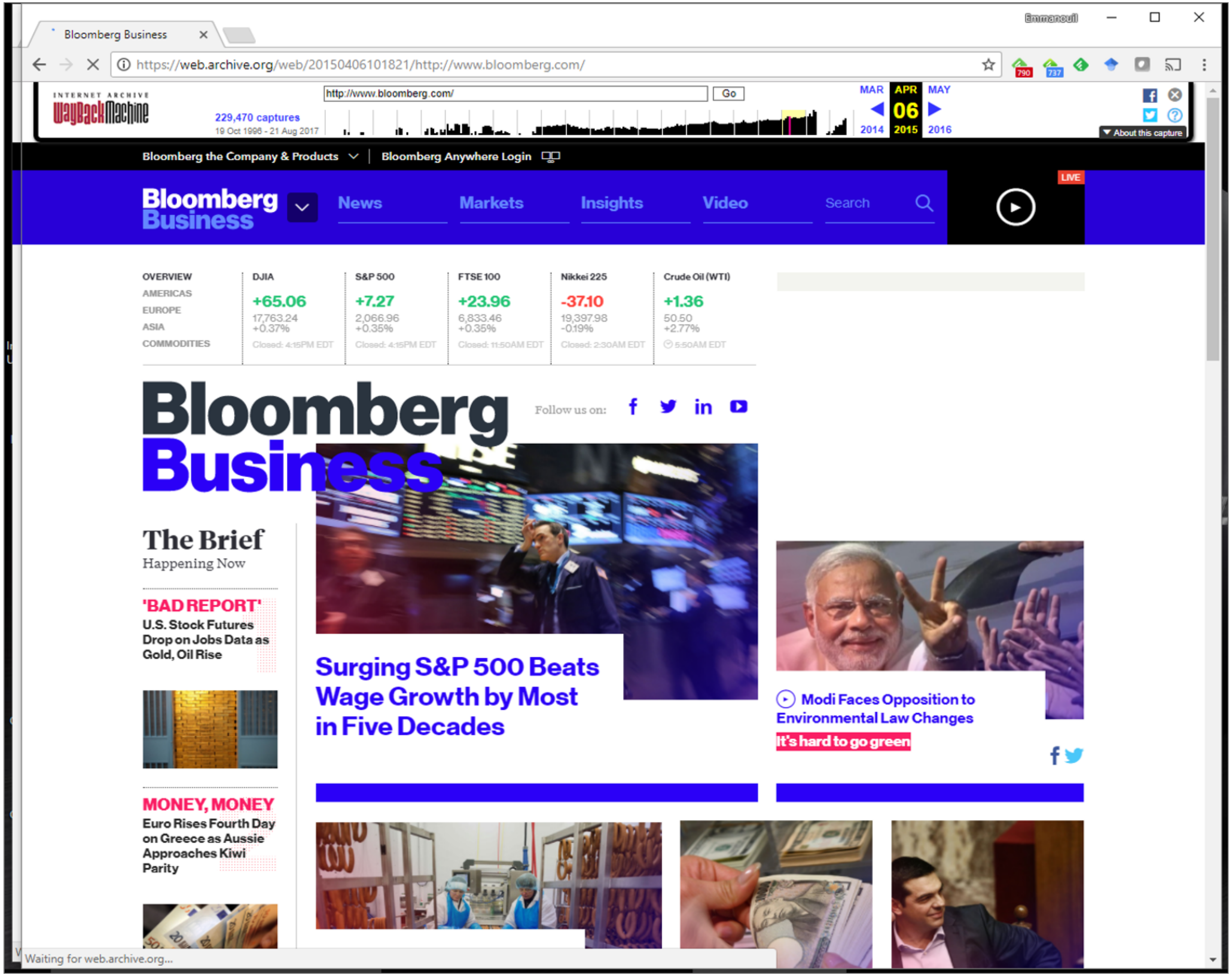
APRIL 6, 2015

78 snapshots

- 00:41:18
- 02:12:40
- 03:05:15
- 05:36:42
- 06:03:48
- 10:18:21
- 11:08:07
- 13:13:12
- 14:36:49
- 16:11:42
- 17:08:14
- 19:15:43
- 20:17:31
- 21:57:37

hive.org/web/20150406004118/http://www.bloomberg.com/

Web data: The Internet Archive



Bloomberg Business
Emmanouil
https://web.archive.org/web/20150406101821/http://www.bloomberg.com/
INTERNET ARCHIVE
WaybackMachine
http://www.bloomberg.com/
Go
229,470 captures
19 Oct 1996 - 21 Aug 2017
MAR
APR
MAY
06
2014
2015
2016
About this capture

Bloomberg the Company & Products

Bloomberg Anywhere Login

OVERVIEW	DJIA	S&P 500	FTSE 100	Nikkei 225	Crude Oil (WTI)
AMERICAS	+65.06	+7.27	+23.96	-37.10	+1.36
EUROPE	17,763.24	2,066.96	6,833.46	19,397.98	50.50
ASIA	+0.37%	+0.35%	+0.35%	-0.19%	+2.77%
COMMODITIES	Closed: 4:15PM EDT	Closed: 4:15PM EDT	Closed: 11:50AM EDT	Closed: 2:30AM EDT	5:50AM EDT

Bloomberg Business

Follow us on:

The Brief

Happening Now

'BAD REPORT'

U.S. Stock Futures Drop on Jobs Data as Gold, Oil Rise

MONEY, MONEY

Euro Rises Fourth Day on Greece as Aussie Approaches Kiwi Parity

Surging S&P 500 Beats Wage Growth by Most in Five Decades

Modi Faces Opposition to Environmental Law Changes

It's hard to go green

Waiting for web.archive.org...

Web data: The Internet Archive

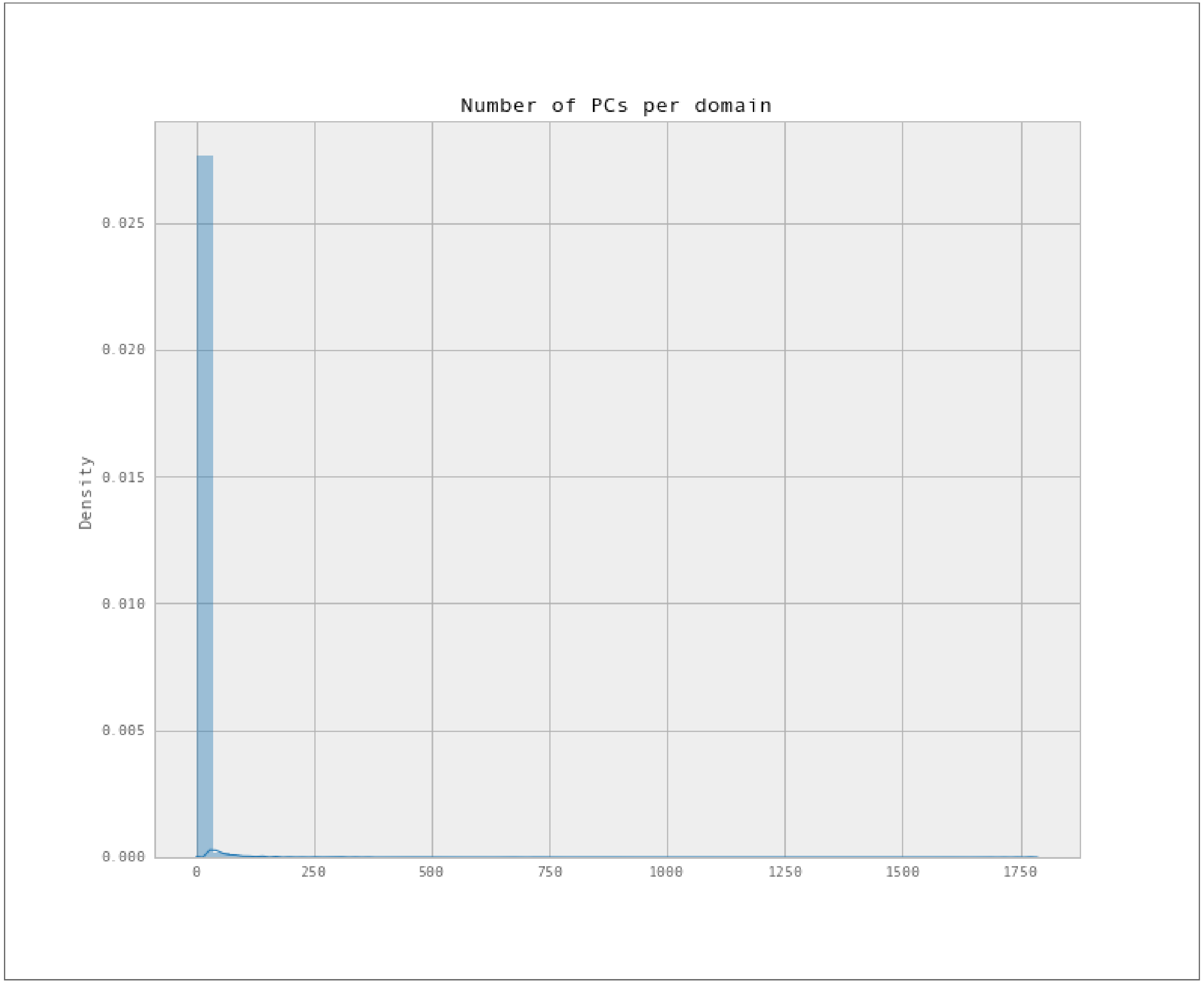
- JISC UK Web Domain Dataset: all archived webpages from the .uk domain 1996-2012
- Curated by the British Library
- Geoindex: a subset of the .uk archived webpages which contain a UK postcode
- circa 0.5 billion URLs with valid UK postcodes

20080509162138/**http://uk.eurogate.co.uk/contact_us** IG8 8HD

Data cleaning

- All the archived .uk webpages
- Archived during 2000-2012
- Commercial webpages (.co.uk & .ltd.uk)
- A postcode *in the web text* within the Shoreditch area
- From webpages to websites: **http://www.website1.co.uk/webpage1** and **http://www.website1.co.uk/webpage2** are part of the **http://www.website1.co.uk**
- 1 *vs*. multuple postcodes in a website

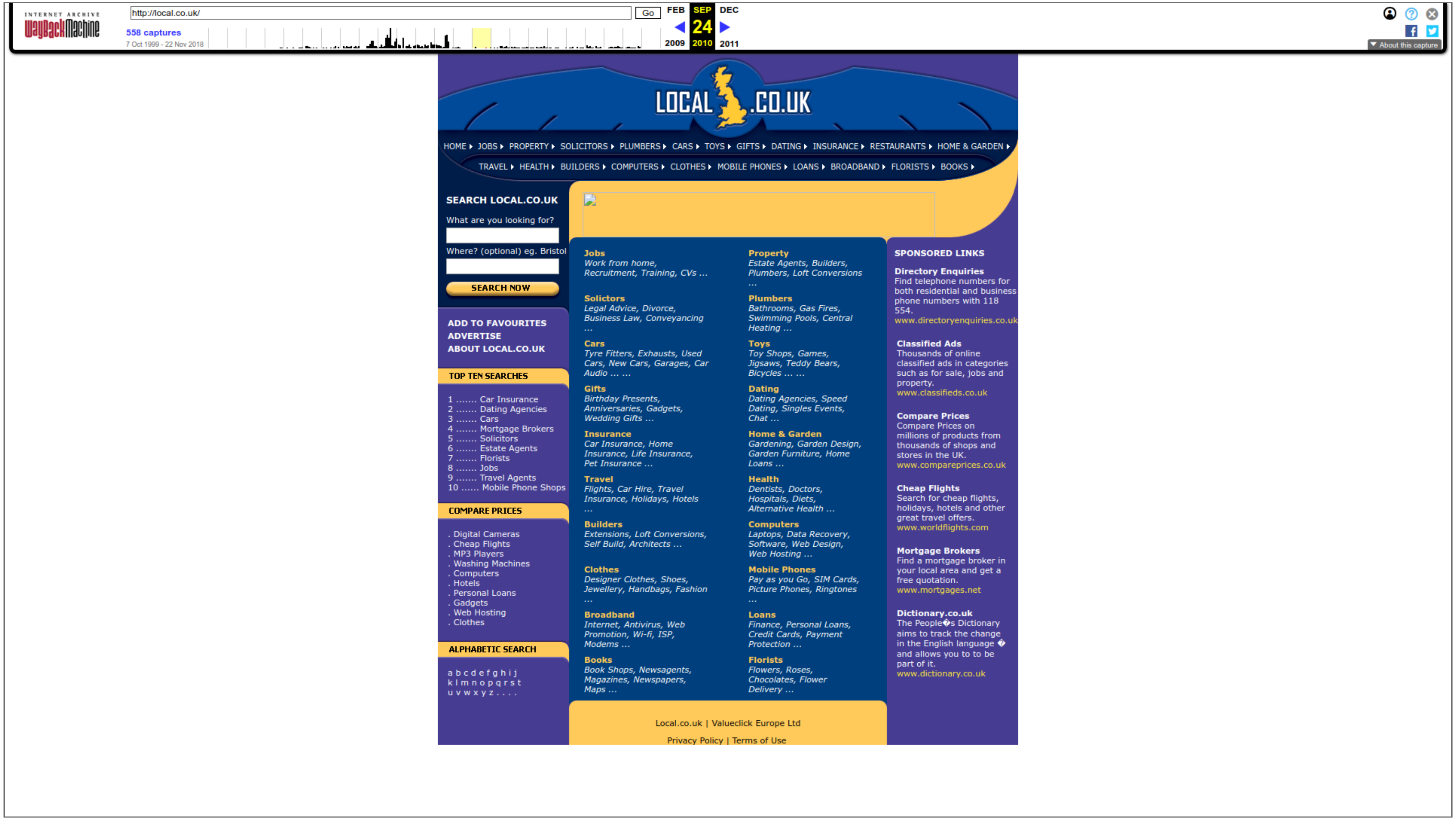
Data cleaning



Data cleaning

- Right side: websites with a large number of postcodes (e.g. directories, real estate websites)
- Left side: websites with a unique postcode in Shoreditch

Directory website with a lot of postcodes


INTERNET ARCHIVE
WaybackMachine
http://local.co.uk/
Go
558 captures
7 Oct 1999 - 22 Nov 2018
FEB
SEP
DEC
24
2009
2010
2011
About this capture
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

SEARCH LOCAL.CO.UK

What are you looking for?

Where? (optional) eg. Bristol

ADD TO FAVOURITES
ADVERTISE
ABOUT LOCAL.CO.UK

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 ...

Solictors
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 ...

Local.co.uk | Valueclick Europe Ltd

Privacy Policy | Terms of Use

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

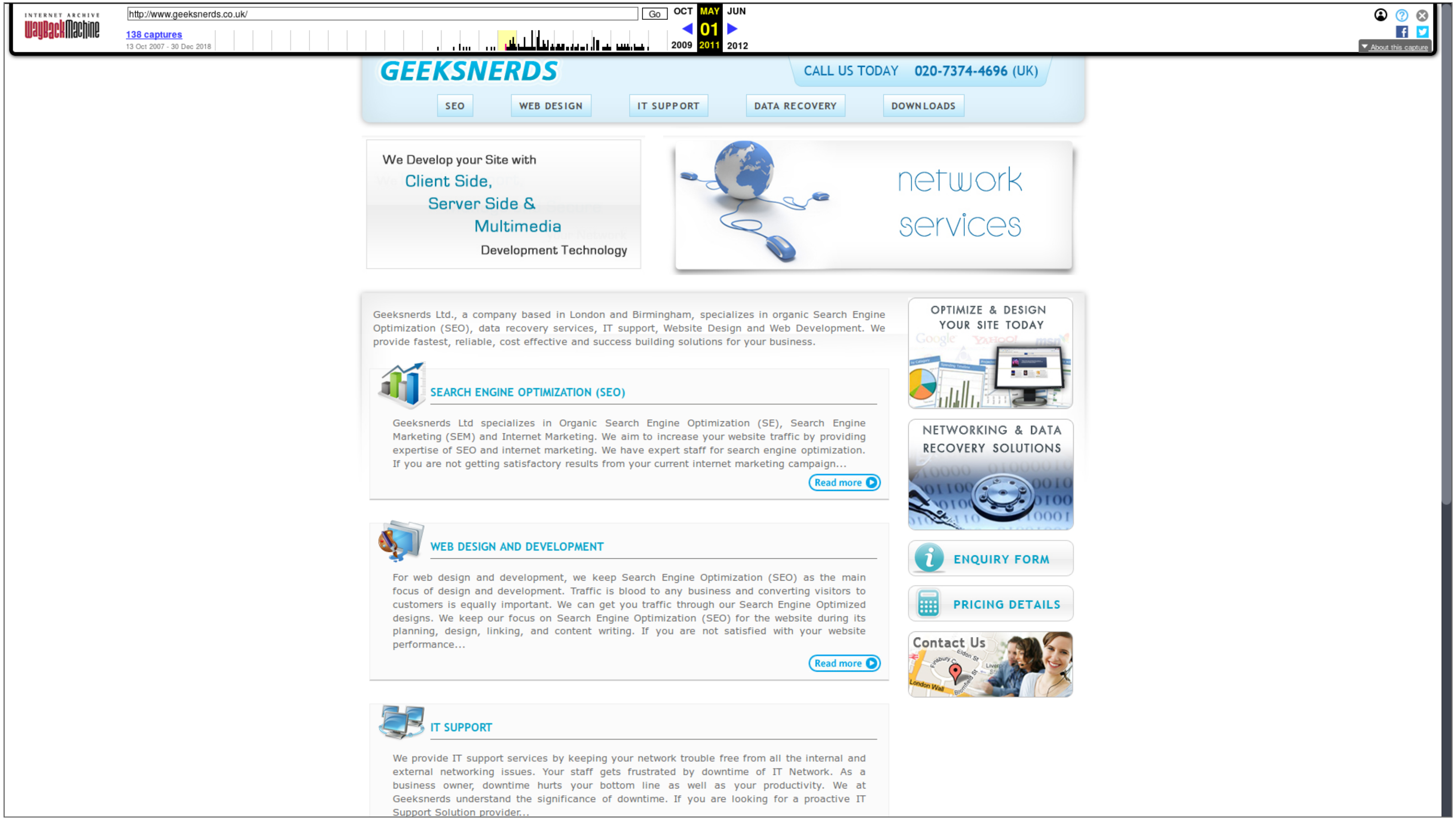
Compare Prices
Compare Prices on millions of products from thousands of shops and stores in the UK.
www.compareprices.co.uk

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

Website with a unique postcode in Shoreditch



We Develop your Site with
Client Side,
Server Side &
Multimedia
Development Technology

Geeksnerds Ltd., a company based in London and Birmingham, specializes in organic Search Engine Optimization (SEO), data recovery services, IT support, Website Design and Web Development. We provide fastest, reliable, cost effective and success building solutions for your business.

SEARCH ENGINE OPTIMIZATION (SEO)

Geeksnerds Ltd specializes in Organic Search Engine Optimization (SE), Search Engine Marketing (SEM) and Internet Marketing. We aim to increase your website traffic by providing expertise of SEO and internet marketing. We have expert staff for search engine optimization. If you are not getting satisfactory results from your current internet marketing campaign...

WEB DESIGN AND DEVELOPMENT

For web design and development, we keep Search Engine Optimization (SEO) as the main focus of design and development. Traffic is blood to any business and converting visitors to customers is equally important. We can get you traffic through our Search Engine Optimized designs. We keep our focus on Search Engine Optimization (SEO) for the website during its planning, design, linking, and content writing. If you are not satisfied with your website performance...

IT SUPPORT

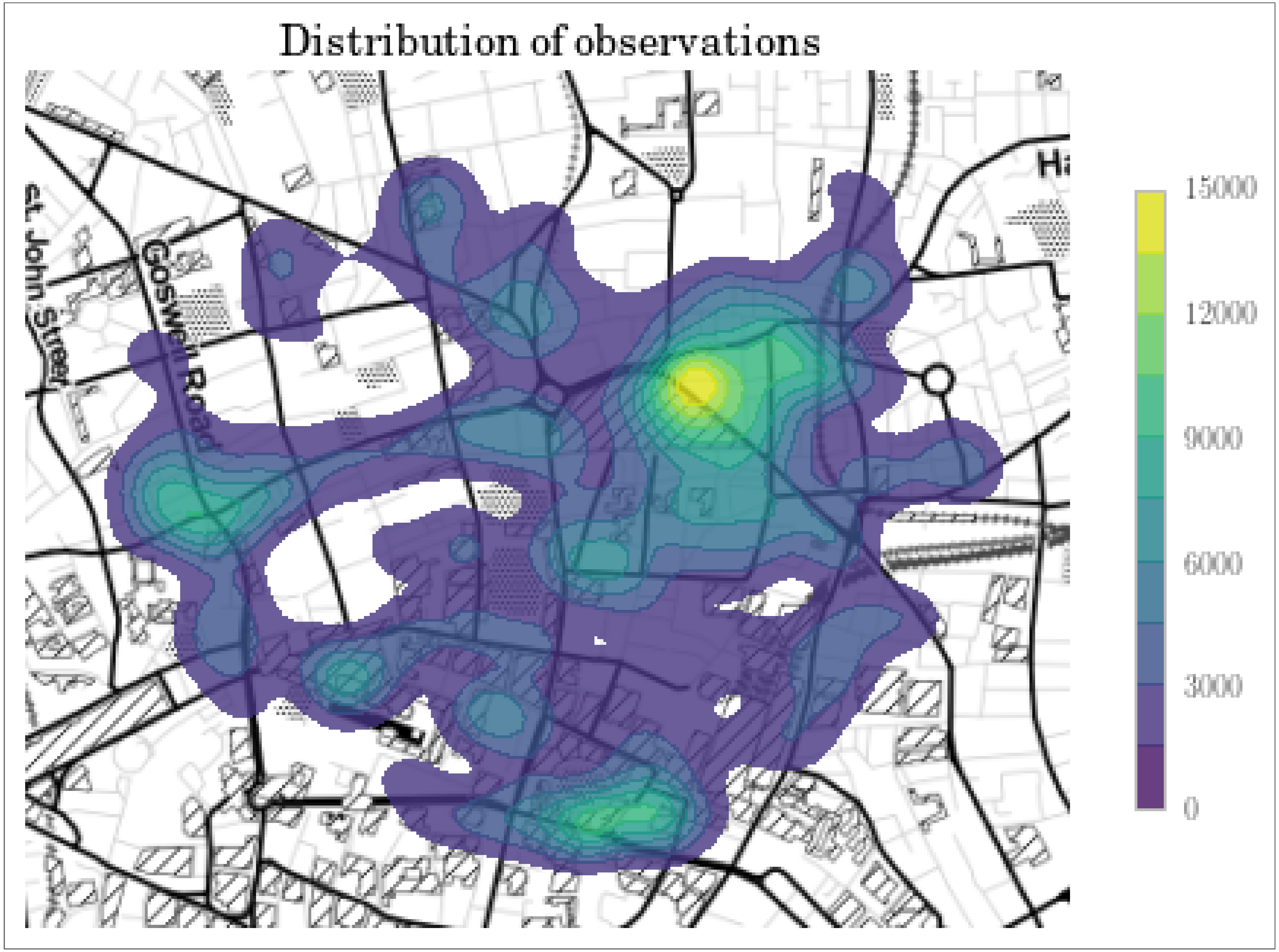
We provide IT support services by keeping your network trouble free from all the internal and external networking issues. Your staff gets frustrated by downtime of IT Network. As a business owner, downtime hurts your bottom line as well as your productivity. We at Geeksnerds understand the significance of downtime. If you are looking for a proactive IT Support Solution provider...

Data cleaning

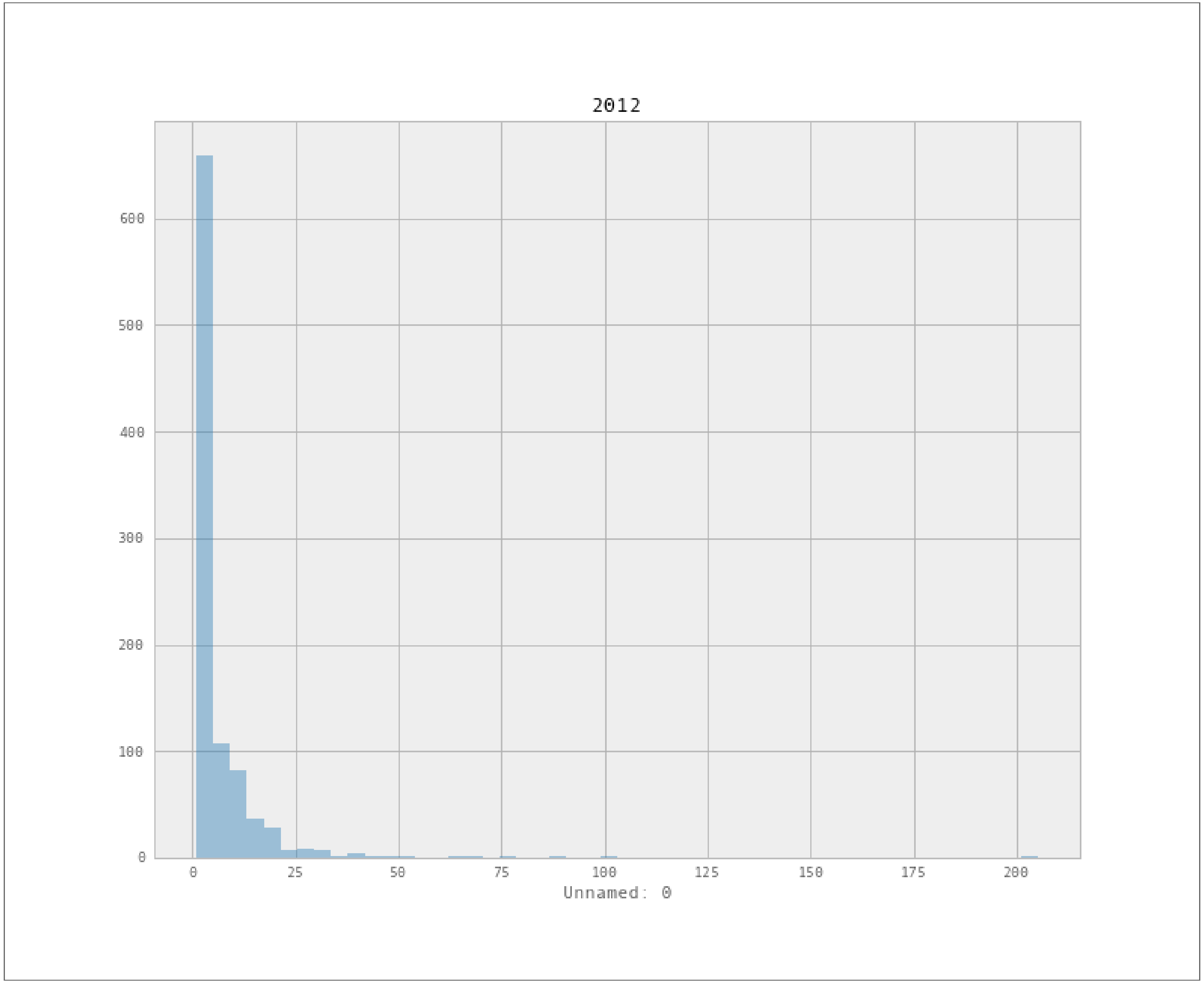
- Current analysis: website with a *unique* postcode in Shoreditch
- 71% of all the archived, commercial, geolocated websites for 2010
- Sensitivity: repeat the analysis including websites with up to 9 postcodes, at least one within Shoreditch
- 95% of all the archived, commercial, geolocated websites for 2010
- We observe **economic activities** and **not firms** within industries
- Websites do not necessarily correspond to firm entities

Results

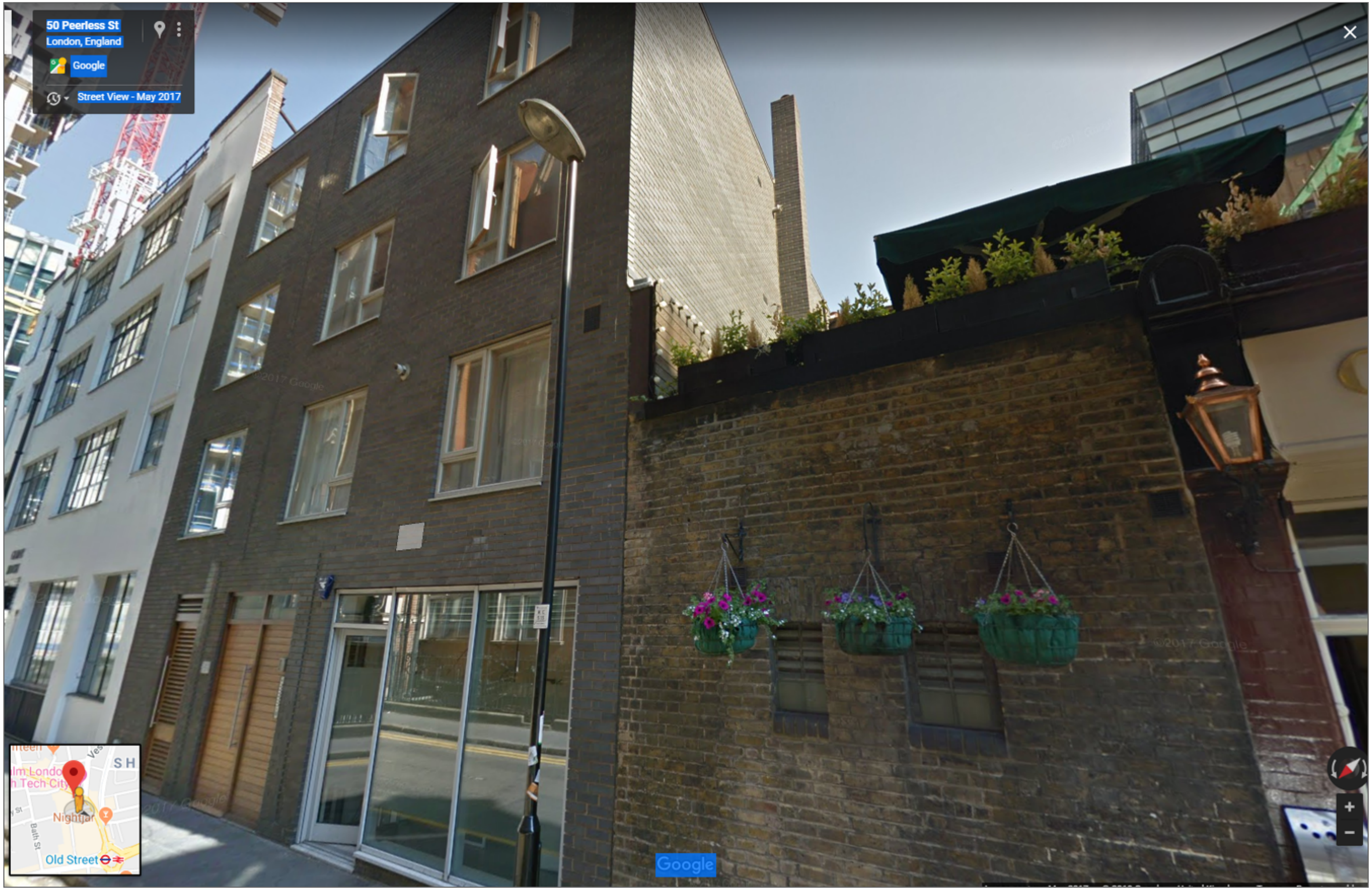
Spatial concentration



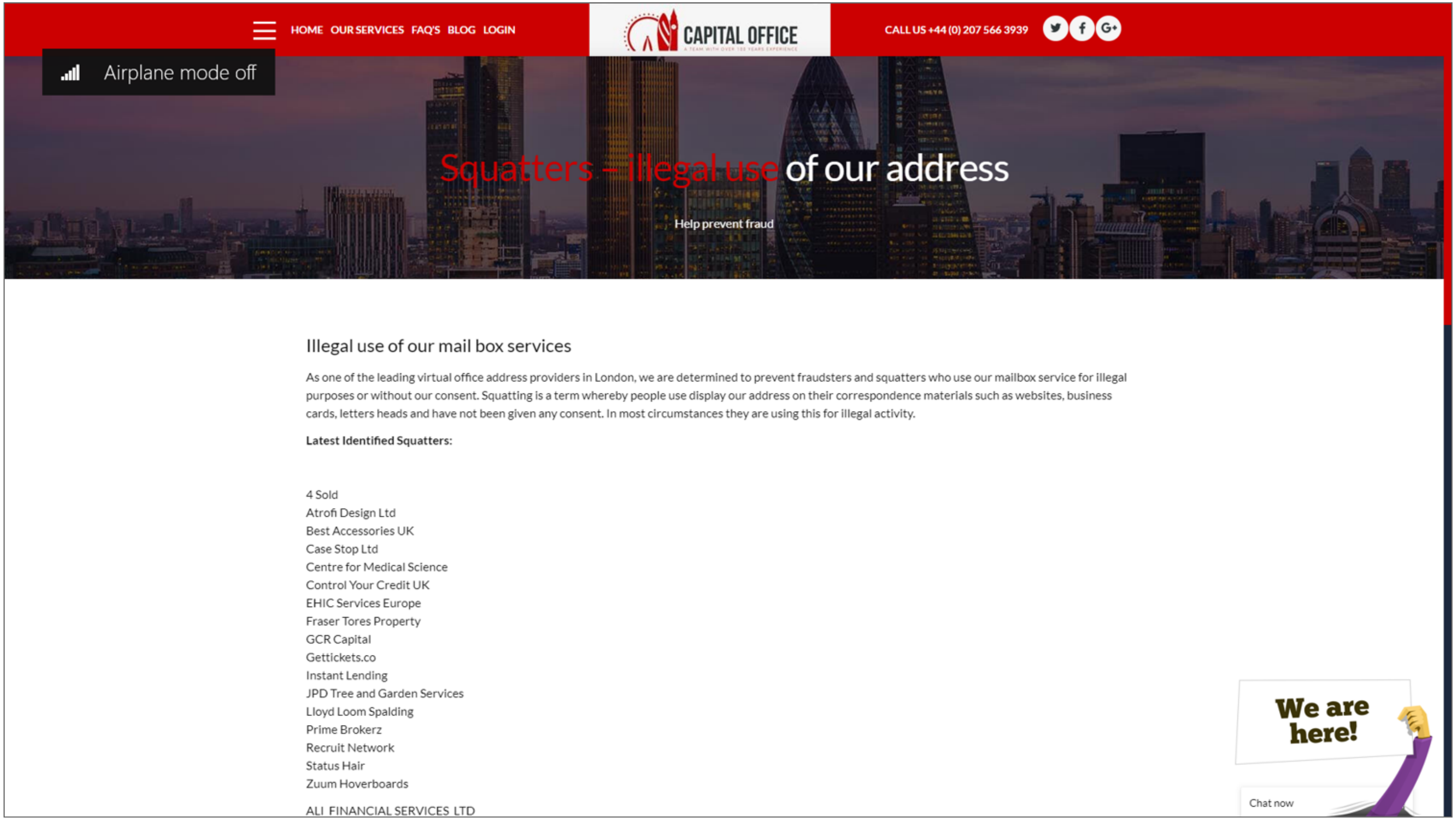
Websites per postcode



Outlier


50 Peerless St
London, England
Google
Street View - May 2017
Old Street
Nightjar
Bath St
S H
Google

Digital squatting



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

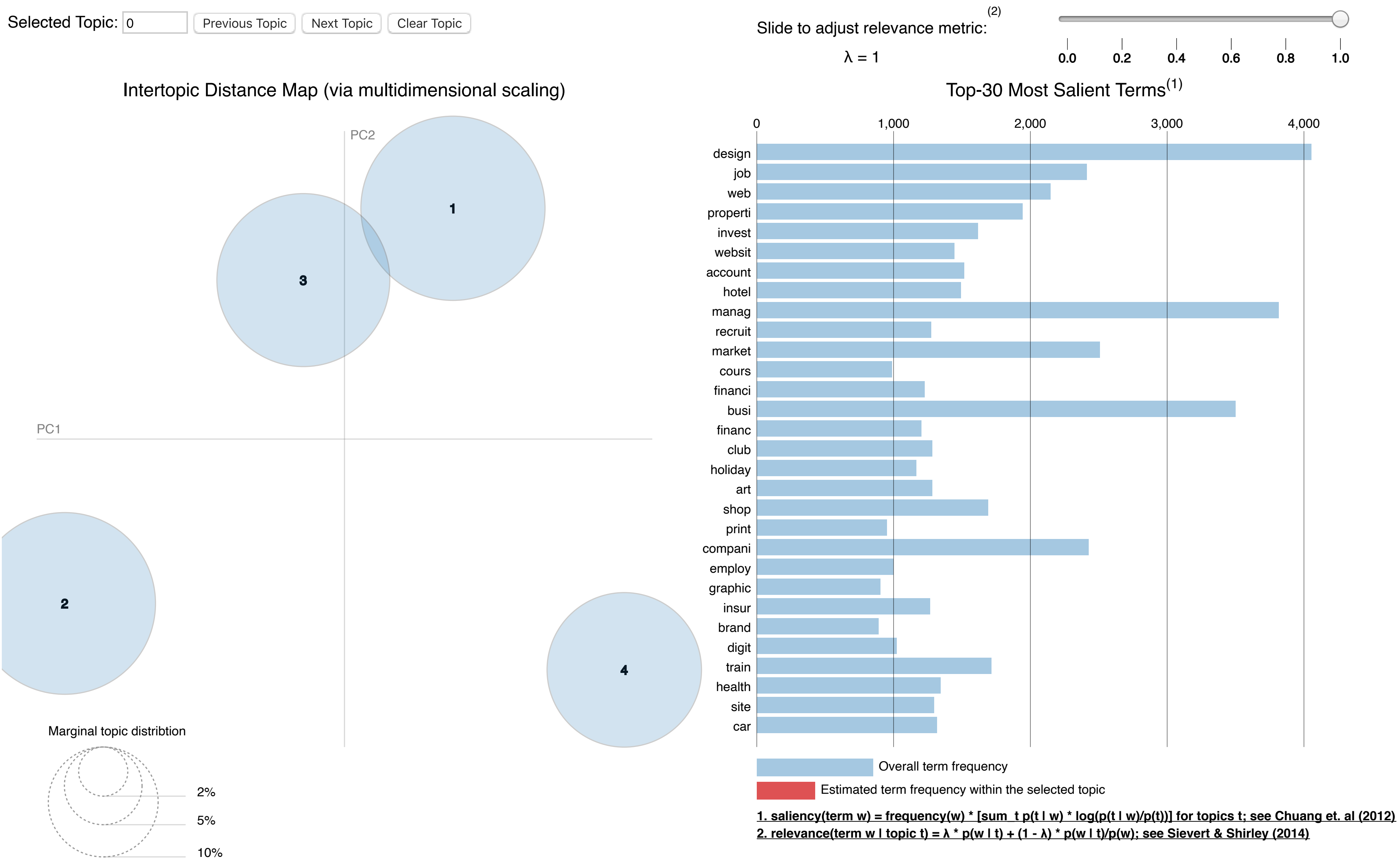
Help prevent fraud

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 Sold
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



Selected Topic: 0

Slide to adjust relevance metric:(2)

λ = 1

Intertopic Distance Map (via multidimensional scaling)

Top-30 Most Salient Terms(1)

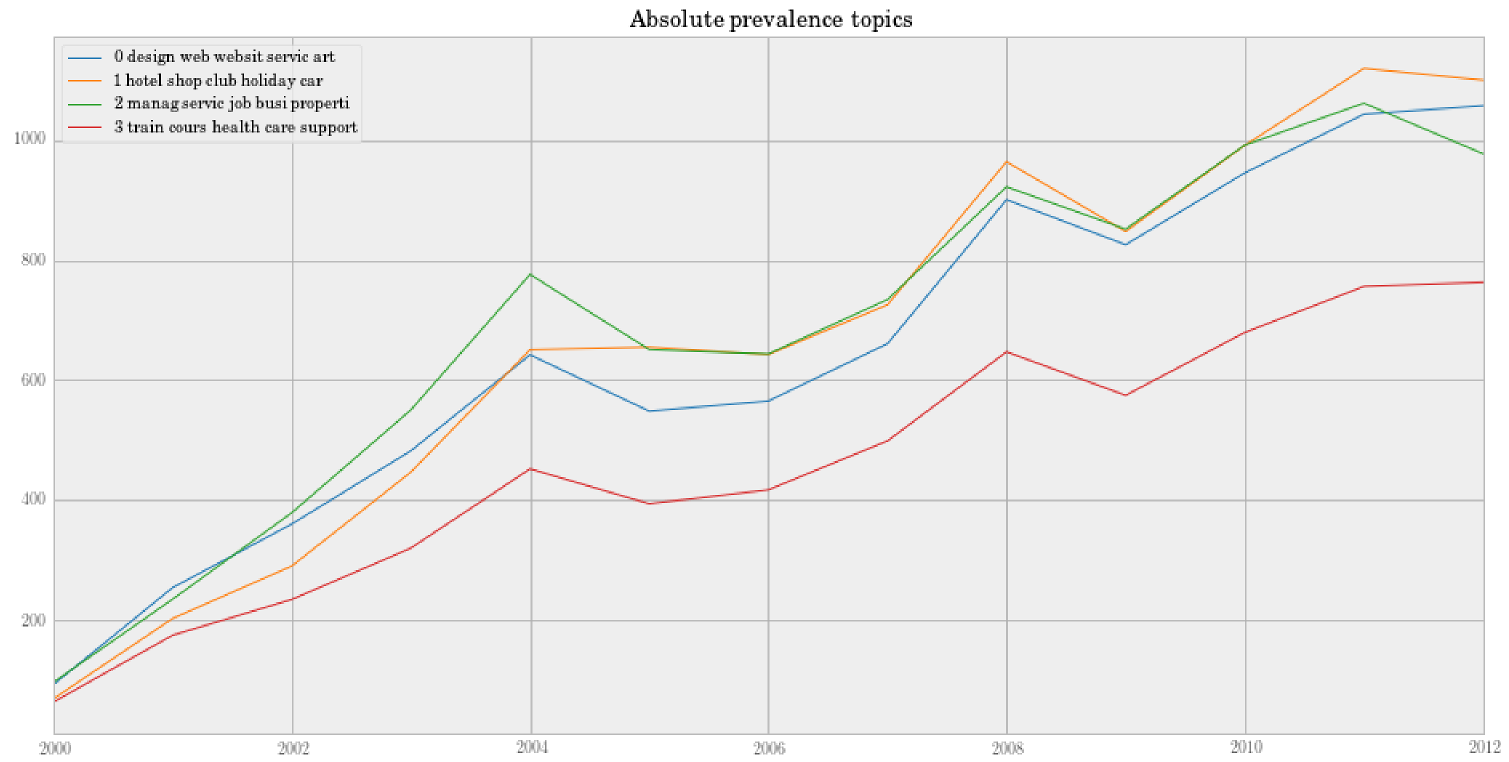
Overall term frequency

Estimated term frequency within the selected topic

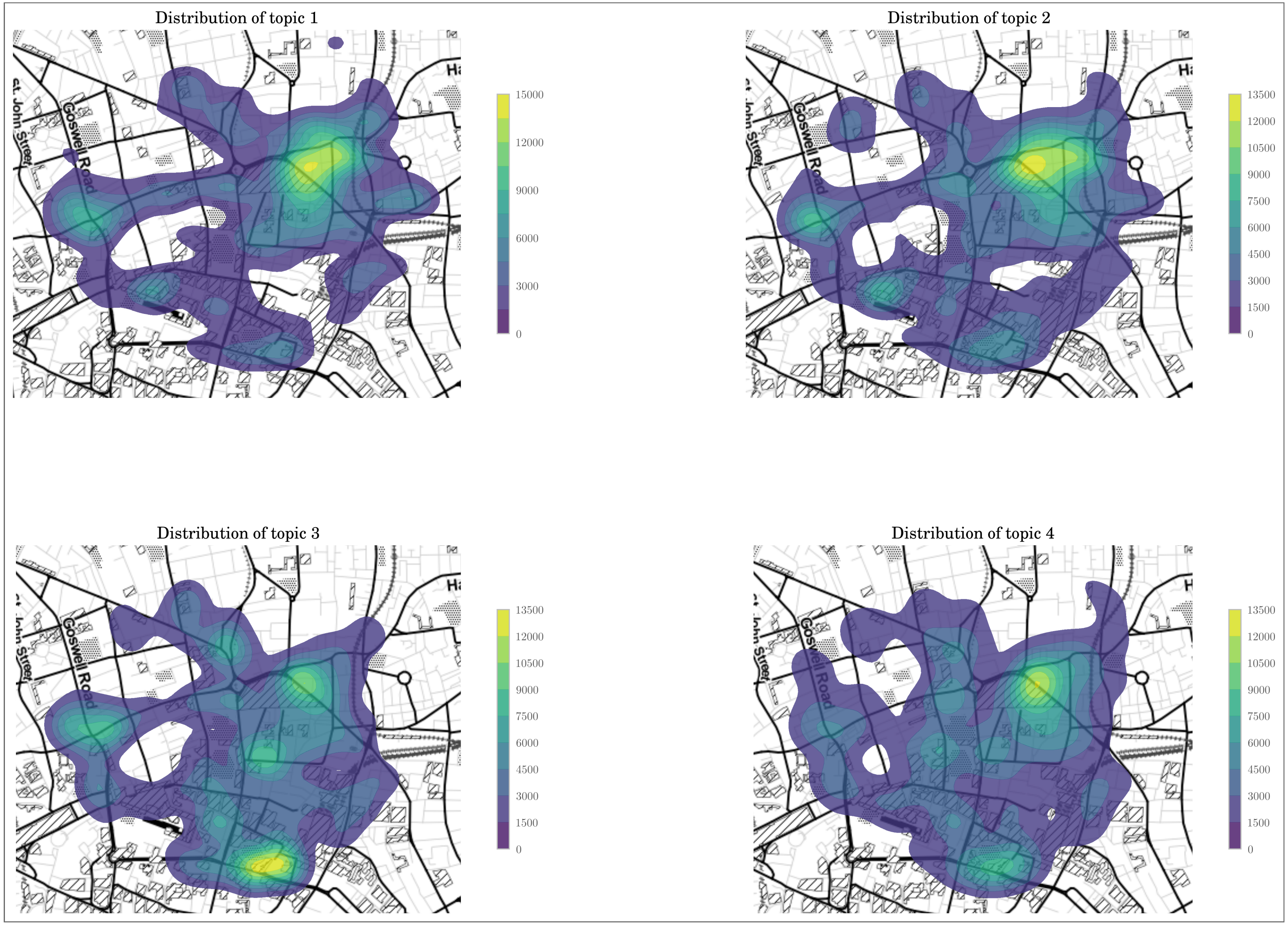
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)

2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Topics over time



Topics over space

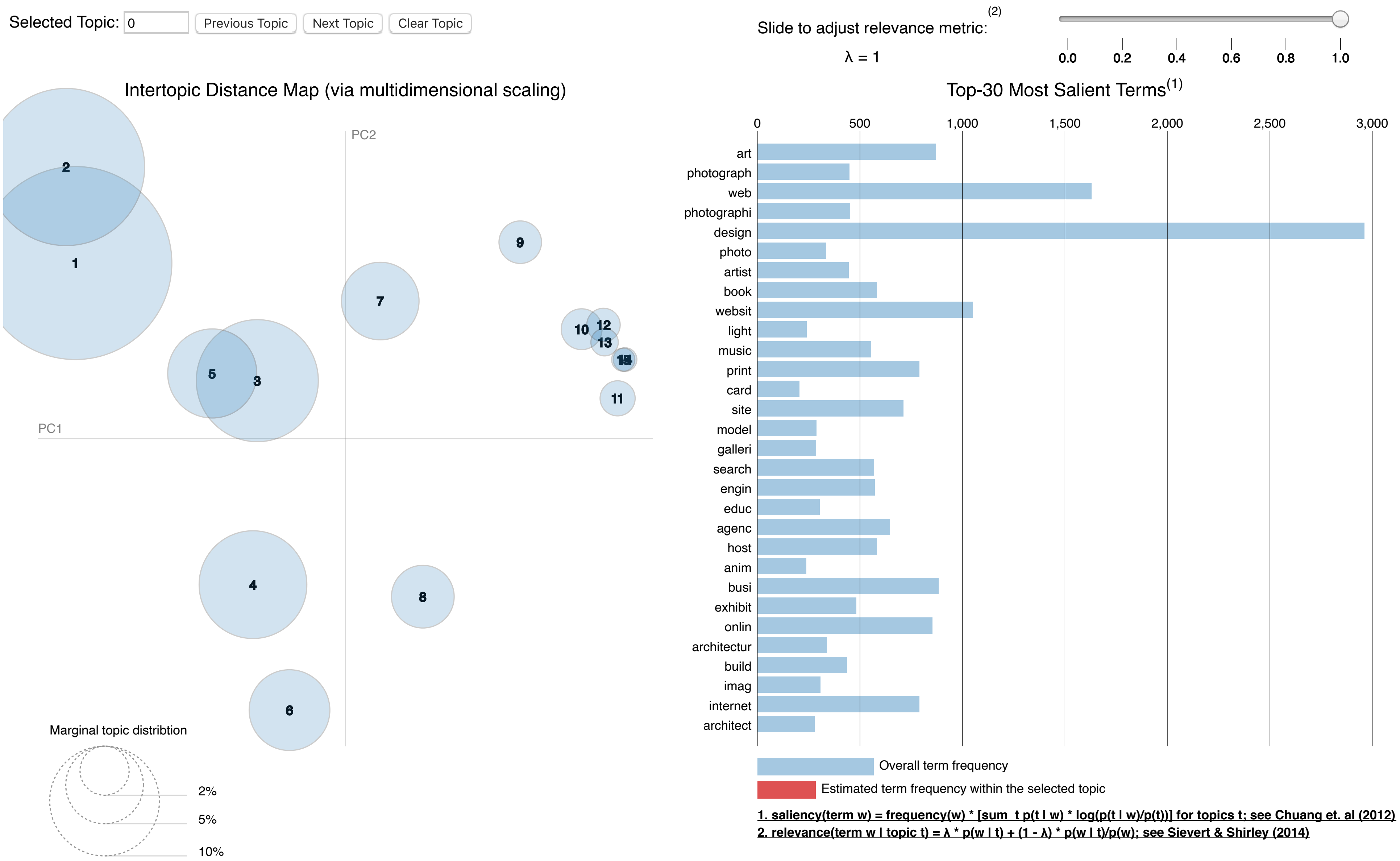


Distribution of topic 1

Distribution of topic 2

Distribution of topic 3

Distribution of topic 4


Selected Topic: 0

Slide to adjust relevance metric:[2]

λ = 1

Intertopic Distance Map (via multidimensional scaling)

Top-30 Most Salient Terms[1]

▬ Overall term frequency

🟥 Estimated term frequency within the selected topic

1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)

2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Topics and keywords

- **Digital media**: design, brand, art, graphic, digit, print
  - Digital content creation; internet services and advertisement
  - publishing and performance arts; visual arts; photography services
  - interior design; garden services; home appliances
- **Visitor and leisure economy**: hotel, club, shop, holiday, car, travel
- **Financial and business service activities**: job, manang, properti, servic, busi, invest, account
- **Health and education**: cours, train, health, care, learn, test, treatment
- Topics: bundles of economic activities

Traditional data

- Administrative data: UK registrar of companies
- SIC codes
- Plotting frequencies of SIC within Shoreditch
- Firms active 2000-2012

Traditional data

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

SIC Codes	Count	Description	Share
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**

Conclusions

- Modelling clusters and their dynamics *is not* a trivial problem
- Hard-to-solve empirical challenges
- Powerful and flexible approach
  - empirical challenges
  - implement key theoretical concepts (within-cluster co-location patterns, local distinctiveness, related / unrelated variety of activity, and cluster evolution)
- More informative than next-best analysis using open administrative data
- Detect unknown or emerging cluster formations