

# 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

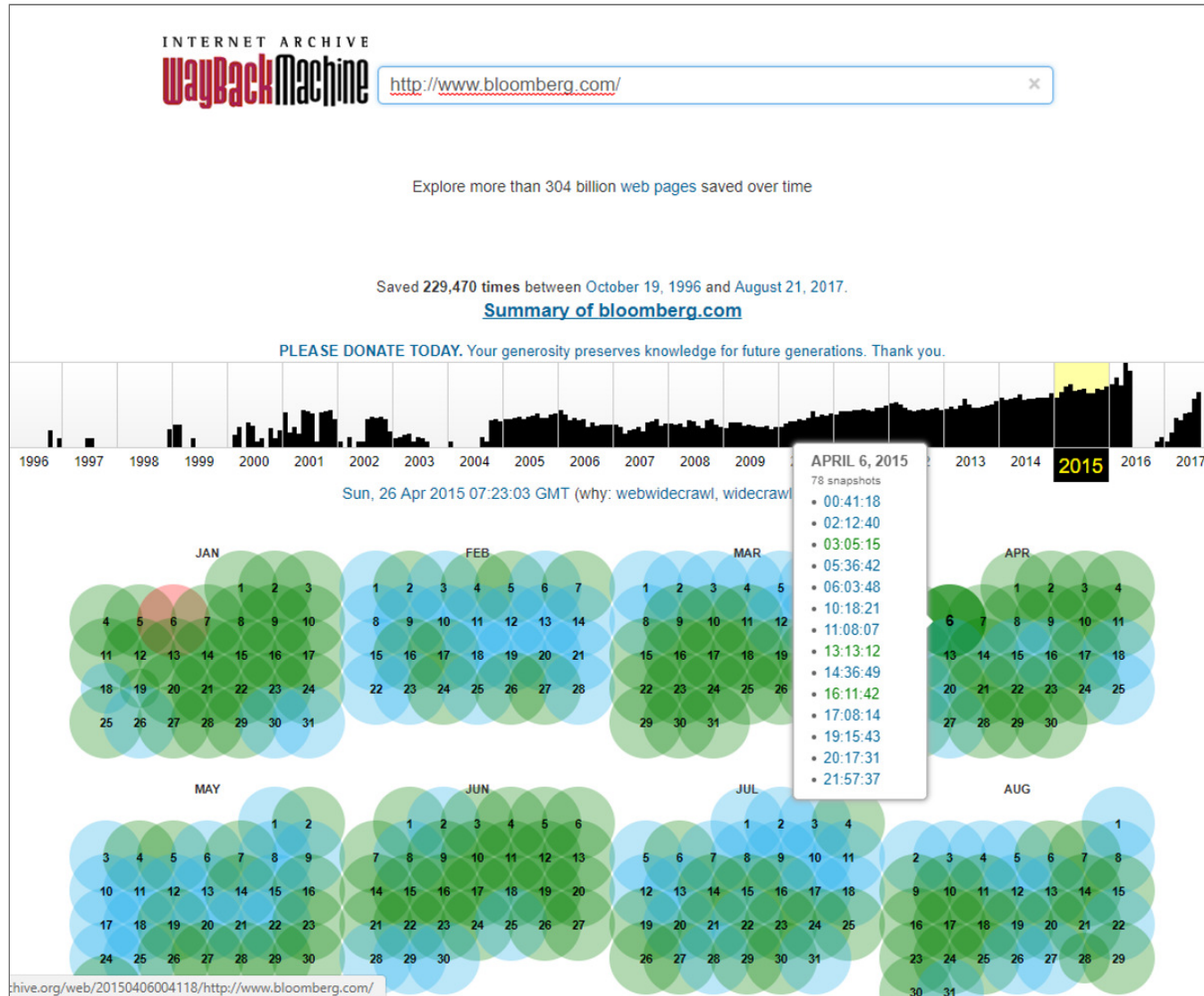


**Data**

## 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



# Web data: The Internet Archive

The screenshot shows a web browser window displaying the Bloomberg Business website as it appeared in April 2015. The browser's address bar shows the URL: <https://web.archive.org/web/20150406101821/http://www.bloomberg.com/>. The Internet Archive Wayback Machine interface is visible at the top, indicating 229,470 captures between October 1998 and August 2017. The Bloomberg website header includes the logo and navigation links for News, Markets, Insights, Video, and Search. A 'LIVE' indicator is present in the top right corner.

The main content area features a market overview table with the following data:

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

The main headline reads: **Bloomberg Business**. Below it, a large image shows a man in a suit looking at a financial display. The headline text is: **Surging S&P 500 Beats Wage Growth by Most in Five Decades**. To the right, a smaller headline reads: **Modi Faces Opposition to Environmental Law Changes** with the sub-headline **It's hard to go green**. The left sidebar contains a section titled 'The Brief' with the sub-headline 'Happening Now' and a sub-headline **'BAD REPORT'** U.S. Stock Futures Drop on Jobs Data as Gold, Oil Rise. Below this is another sub-headline **MONEY, MONEY** Euro Rises Fourth Day on Greece as Aussie Approaches Kiwi Parity.

At the bottom of the page, there are several small images: a factory interior, a hand holding a stack of Euro banknotes, and a man in a suit. The browser status bar at the bottom left shows 'Waiting for web.archive.org...'.

## Web data: The Internet Archive

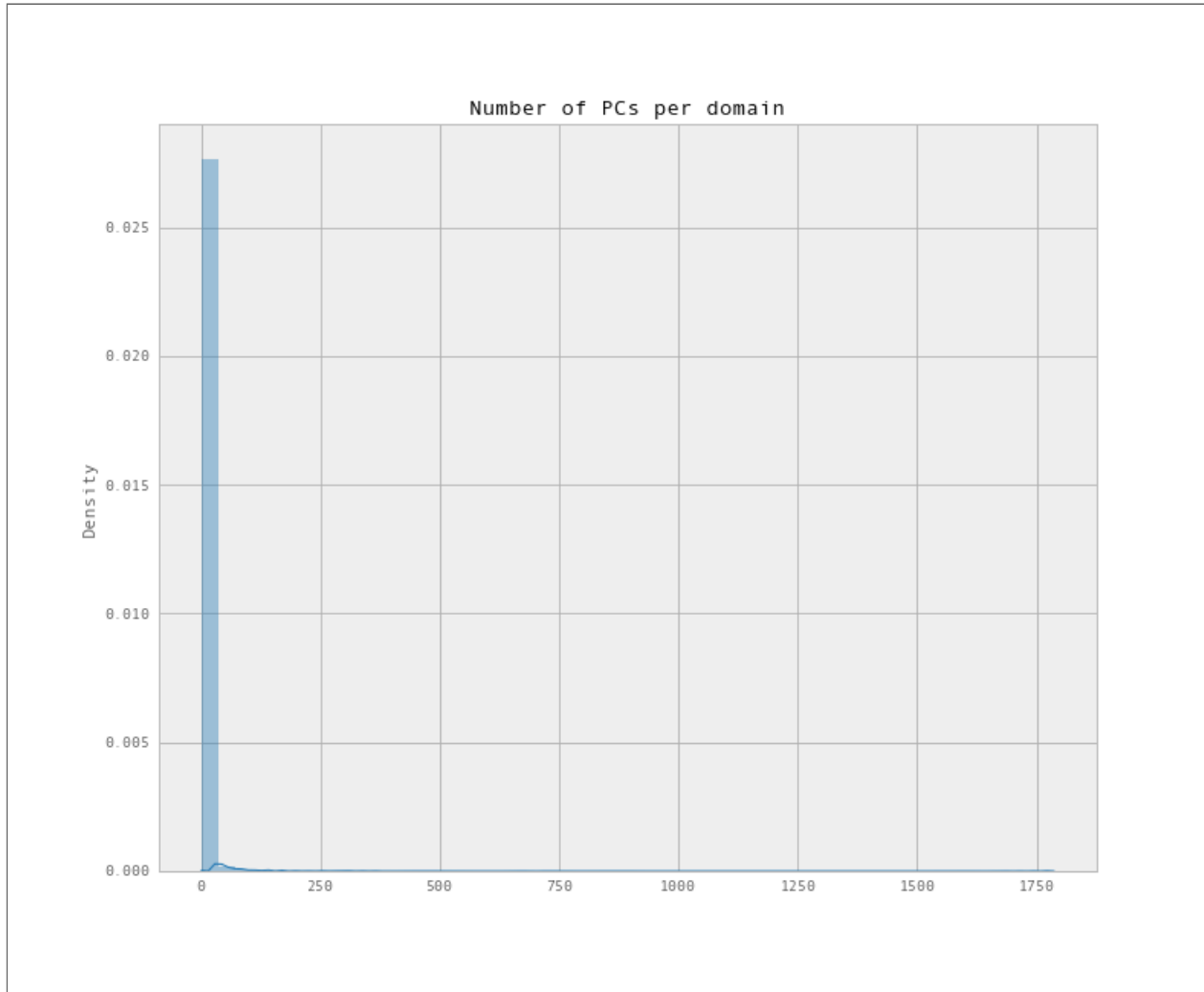
- JISC UK Web Domain Dataset: all archived webpages from the .uk domain 1996-2012
- Curated by the British Library
- Geindex: a subset of the .uk archived webpages which contain a UK postcode
- circa 0.5 billion URLs with valid UK postcodes

20080509162138/[http://uk.eurogate.co.uk/contact\\_us](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.* multiple 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

The screenshot shows the homepage of Local.co.uk, a directory website. At the top, there is a navigation bar with a search box containing 'http://local.co.uk/' and a 'Go' button. The date is displayed as 'FEB 24 SEP DEC' with navigation arrows. Social media icons for Facebook and Twitter are visible in the top right corner. Below the navigation bar is a large blue banner with the 'LOCAL.CO.UK' logo and a map of the United Kingdom. A horizontal menu lists various categories: HOME, JOBS, PROPERTY, SOLICITORS, PLUMBERS, CARS, TOYS, GIFTS, DATING, INSURANCE, RESTAURANTS, HOME & GARDEN, TRAVEL, HEALTH, BUILDERS, COMPUTERS, CLOTHES, MOBILE PHONES, LOANS, BROADBAND, FLORISTS, BOOKS. The main content area is divided into several sections: 'SEARCH LOCAL.CO.UK' with a search box and a 'SEARCH NOW' button; 'ADD TO FAVOURITES ADVERTISE ABOUT LOCAL.CO.UK'; 'TOP TEN SEARCHES' listing popular searches like Car Insurance, Dating Agencies, and Cars; 'COMPARE PRICES' listing products like Digital Cameras and MP3 Players; 'ALPHABETIC SEARCH' with a grid of letters; and a central grid of category links including Jobs, Property, Solicitors, Plumber, Toys, Dating, Insurance, Home & Garden, Health, Computers, Mobile Phones, Loans, Florists, and Books. A 'SPONSORED LINKS' section on the right features advertisements for Directory Enquiries, Classified Ads, Compare Prices, Cheap Flights, Mortgage Brokers, and Dictionary.co.uk. The footer contains the text 'Local.co.uk | Valueclick Europe Ltd' and links for 'Privacy Policy' and 'Terms of Use'.

# Website with a unique postcode in Shoreditch

The screenshot shows the website for Geeksnerds, a company based in London and Birmingham. The page features a navigation menu with links for SEO, Web Design, IT Support, Data Recovery, and Downloads. The main content area is divided into several sections: a header with the company name and contact information, a section for network services, a detailed description of the company's services, and three main service categories: Search Engine Optimization (SEO), Web Design and Development, and IT Support. Each category includes a brief description and a 'Read more' link. On the right side, there are additional links for 'Optimize & Design Your Site Today', 'Networking & Data Recovery Solutions', 'Enquiry Form', 'Pricing Details', and 'Contact Us'.

INTERNET ARCHIVE WaybackMachine http://www.geeksnerds.co.uk/ 138 captures 13 Oct 2007 - 30 Dec 2018

GO OCT MAY JUN 01 2009 2011 2012

**GEEKSNERDS** CALL US TODAY 020-7374-4696 (UK)

SEO WEB DESIGN IT SUPPORT DATA RECOVERY DOWNLOADS

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

network services

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...  
[Read more](#)

**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...  
[Read more](#)

**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...

OPTIMIZE & DESIGN YOUR SITE TODAY

NETWORKING & DATA RECOVERY SOLUTIONS

ENQUIRY FORM

PRICING DETAILS

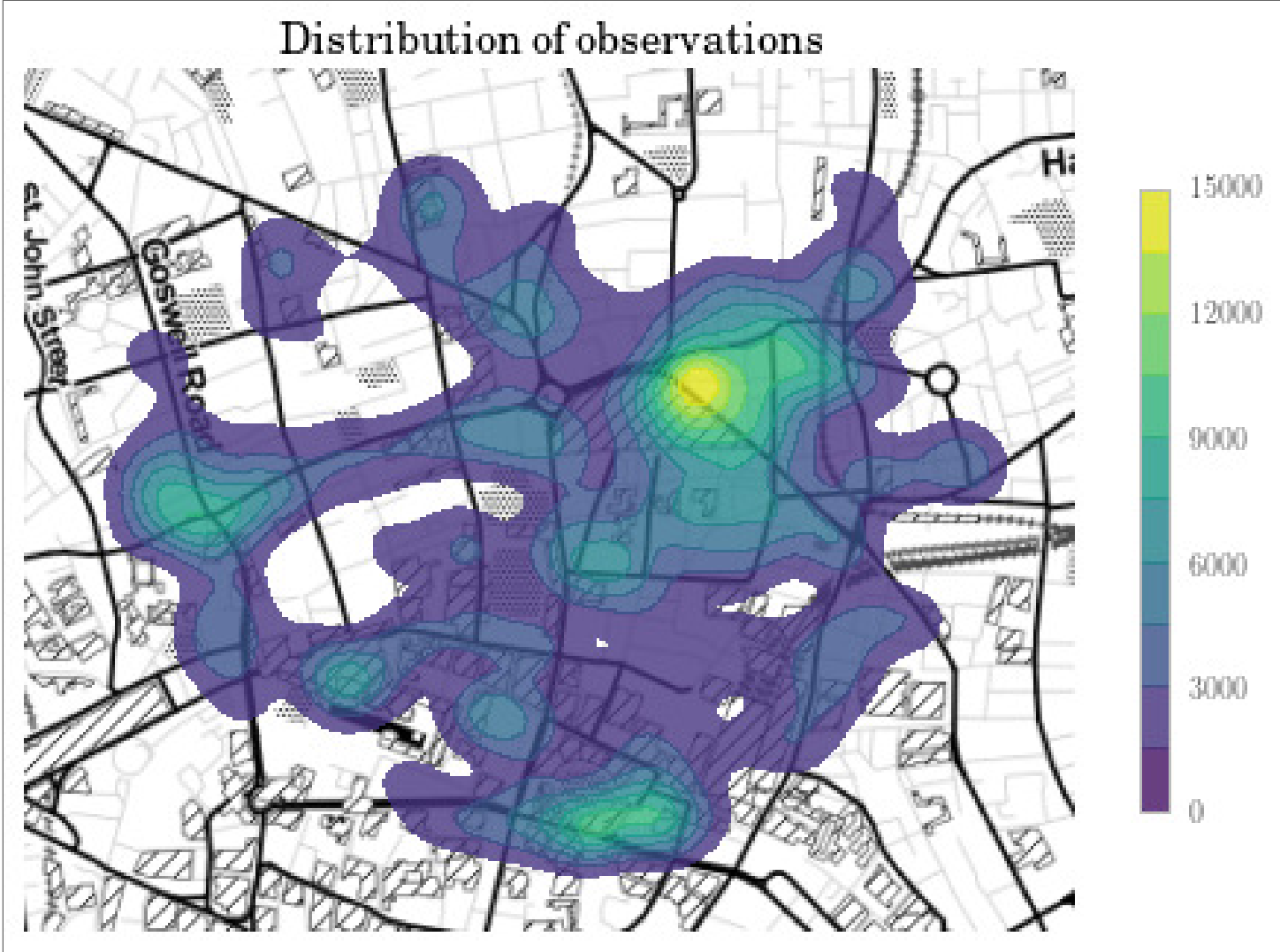
Contact Us

## Data cleaning

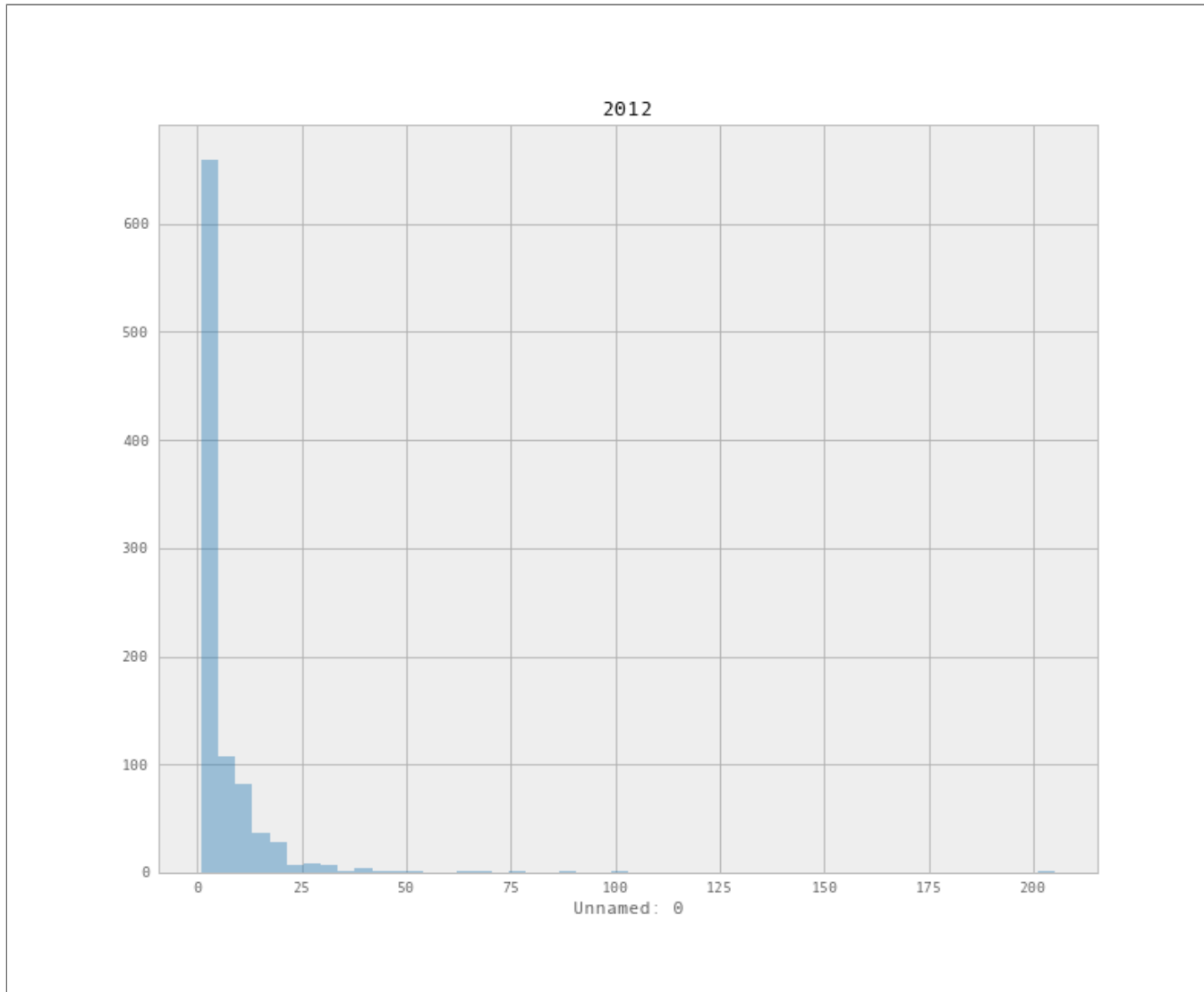
- Current analysis: website with a *unique* postcode in Shoreditch
- 71% of all the archived, commercial, geolocated websites for 2010
- Sensitivity: repeat the analysis including websites with up to 9 postcodes, at least one within Shoreditch
- 95% of all the archived, commercial, geolocated websites 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




# Digital squatting

HOME OUR SERVICES FAQ'S BLOG LOGIN

 CAPITAL OFFICE  
A TEAM WITH OVER 120 YEARS EXPERIENCE

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

**We are here!**

Chat now



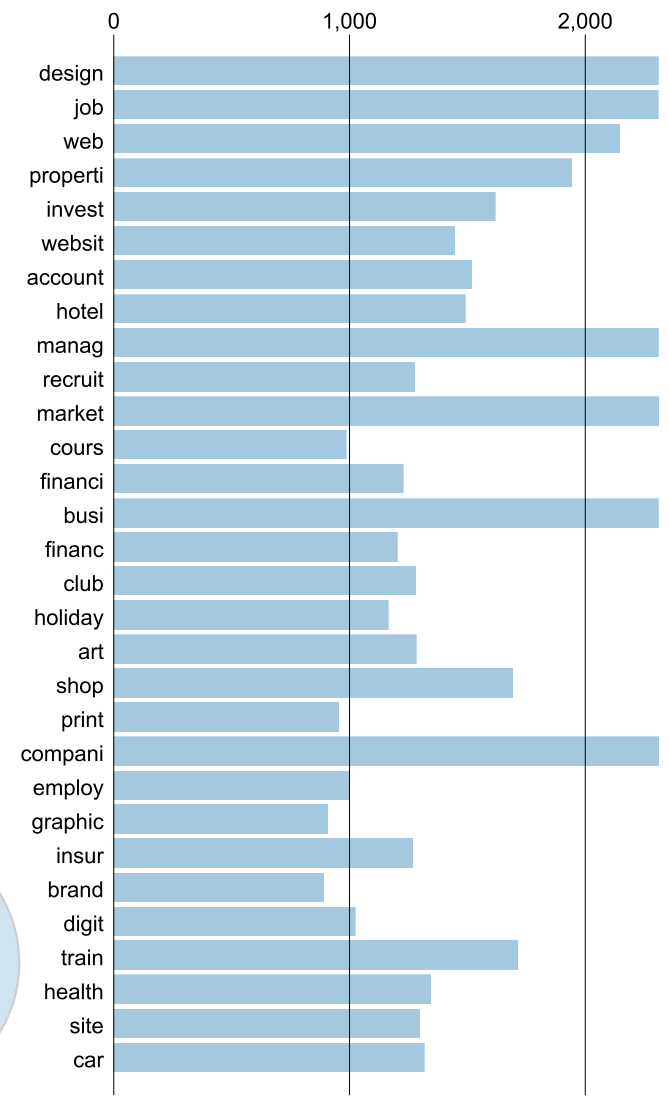
Selected Topic:

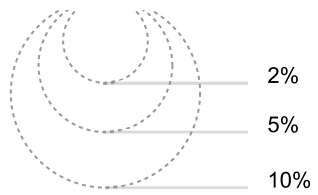
Slide to adjust relevance metric:  (2)  
 $\lambda = 1$  0.0

Intertopic Distance Map (via multidimensional scaling)



Top-30 Most Sa

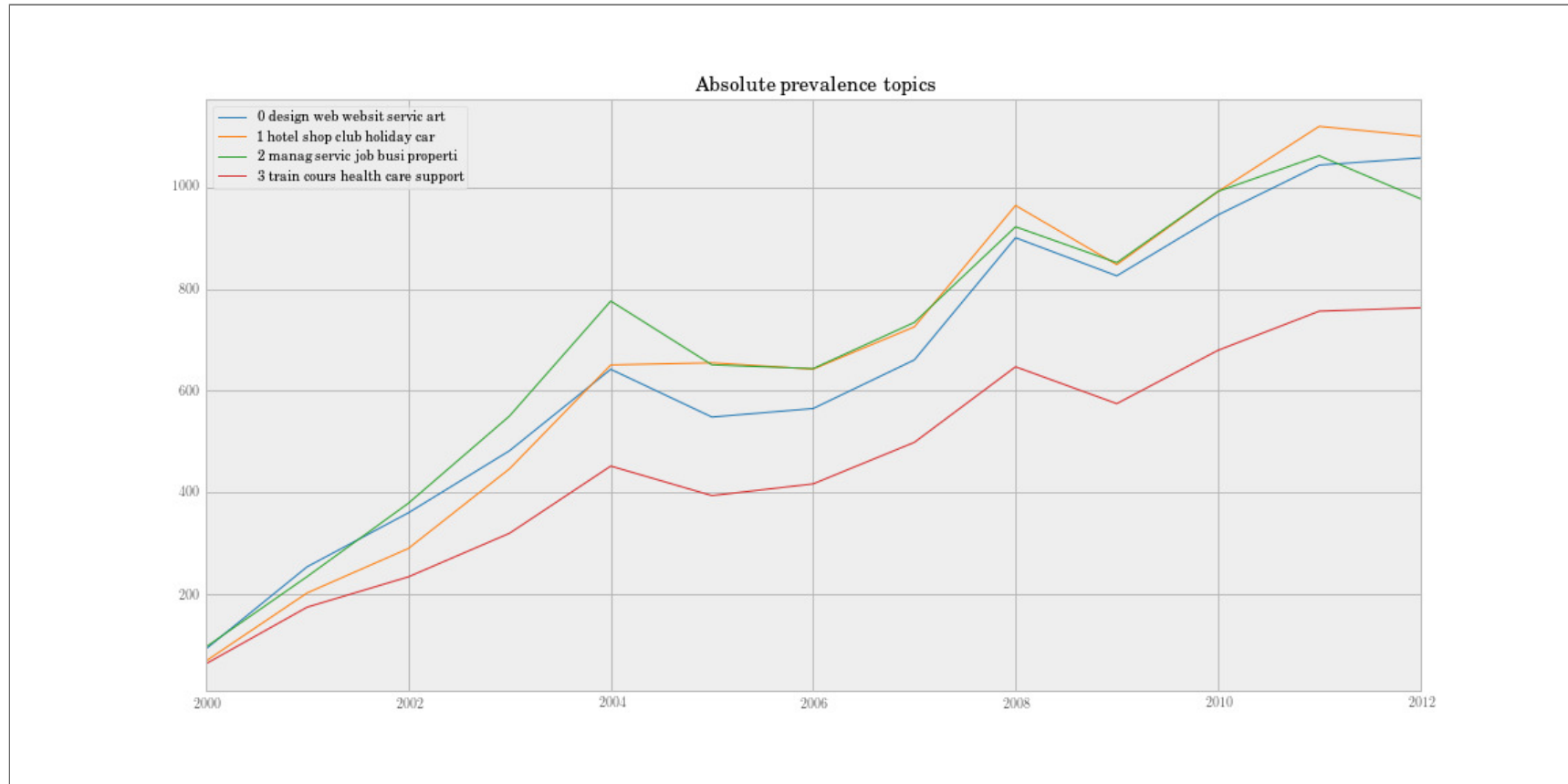




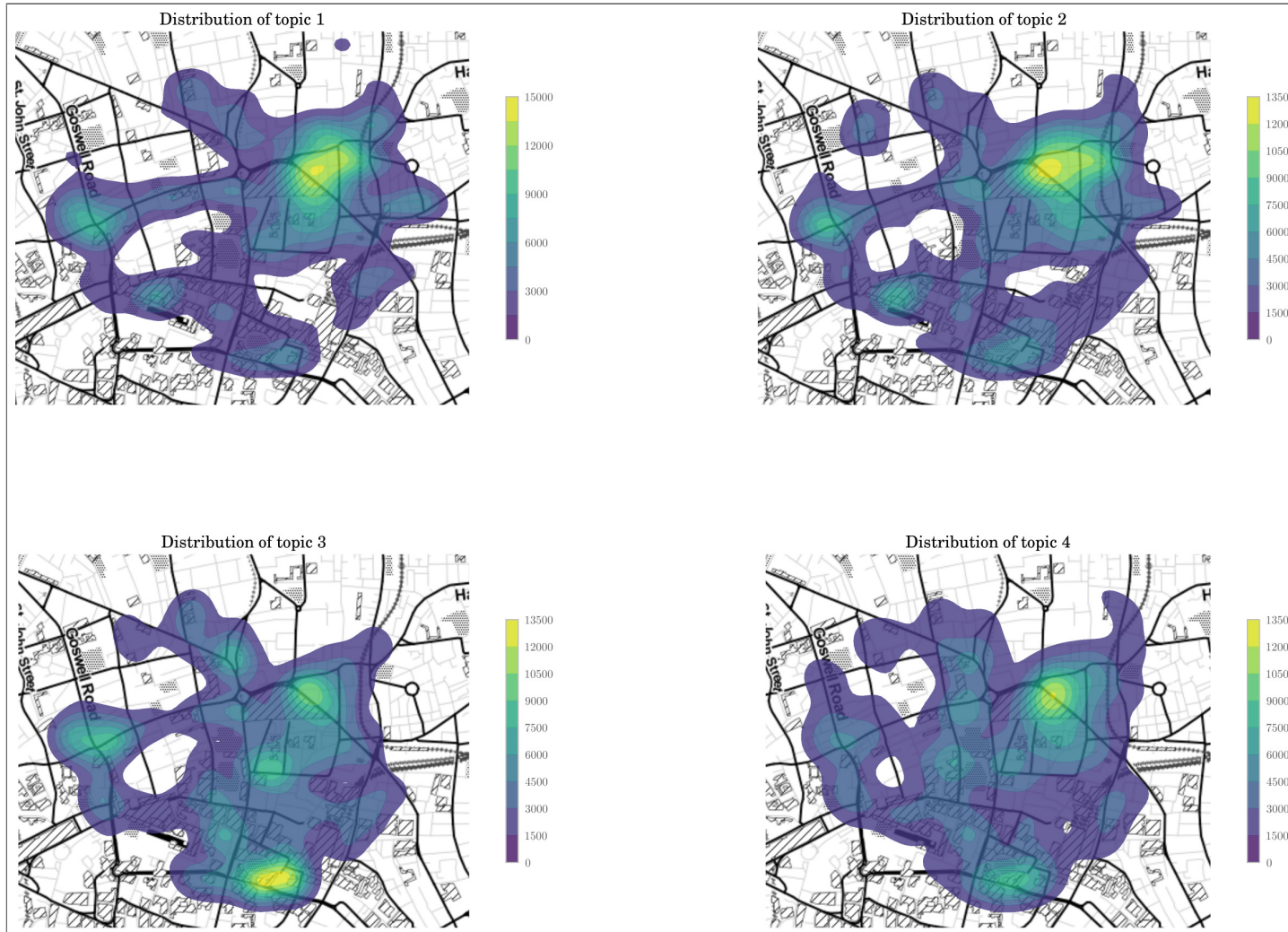
Overall term frequency  
Estimated term frequency within the selected

- saliency(term w) = frequency(w) \* [sum\_t p(t | w) \* p(w | t)]**
- relevance(term w | topic t) = λ \* p(w | t) + (1 - λ) \* p**

# Topics over time



# Topics over space



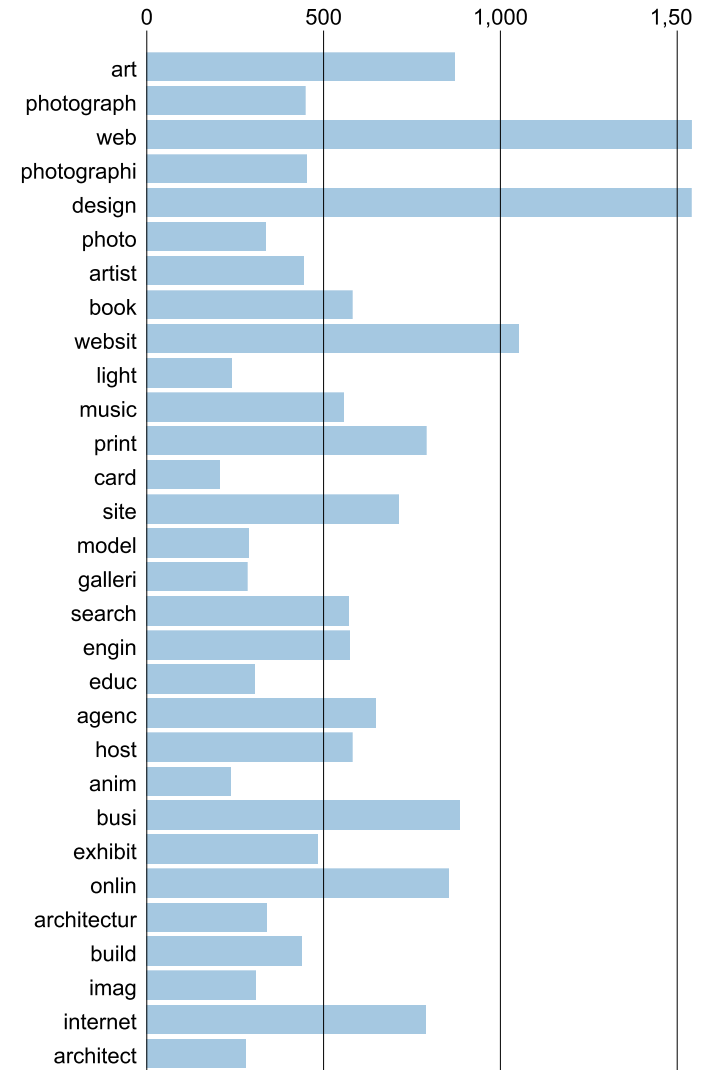
Selected Topic:

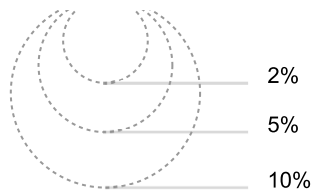
Slide to adjust relevance metric:  (2)  
 $\lambda = 1$  0.0

Intertopic Distance Map (via multidimensional scaling)



Top-30 Most Sa





Overall term frequency  
Estimated term frequency within the selected

- saliency(term w) = frequency(w) \* [sum\_t p(t | w) \* p(w | t)]**
- relevance(term w | topic t) = λ \* p(w | t) + (1 - λ) \* p**

## 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 <b>other</b> than financial management	0.201
64999	517	Financial intermediation <b>not elsewhere classified</b>	0.092
74909	387	<b>Other</b> professional, scientific and technical activities <b>n.e.c.</b>	0.069
68209	371	<b>Other</b> letting and operating of own or leased real estate	0.066
62012	326	Business and domestic software development	0.058
78109	185	<b>Other</b> activities of employment placement agencies	0.033
64209	171	Activities of <b>other</b> holding companies <b>n.e.c.</b>	0.030
56101	157	Licensed restaurants	0.028
59111	154	Motion picture production activities	0.027

SIC Codes	Count	Description	Share
69201	130	Accounting and auditing activities	0.023
71111	123	Architectural activities	0.022
43999	86	<b>Other</b> specialised construction activities <b>n.e.c.</b>	0.015
64205	85	Activities of financial services holding companies	0.015
93199	73	<b>Other</b> 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; <b>other</b> legal activities <b>n.e.c.</b>	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

SIC Codes	Count	Description	Share
59113	63	Television programme production activities	0.011
71129	61	<b>Other</b> 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 <b>not elsewhere classified</b>	0.007
18129	40	Printing <b>n.e.c.</b>	0.007
<b>Total</b>			<b>0.849</b>

## 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