NCRM National Centre for Research Methods

Online Data Sources: Linking Old and New, Big and Small

Susan Banducci and Iulia Cioroianu

Exeter Q-Step Centre

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL

UNIVERSITY OF Southampton

MANCHESTER 1824 The University of Manchester

THE UNIVERSITY OF EDINBURGH

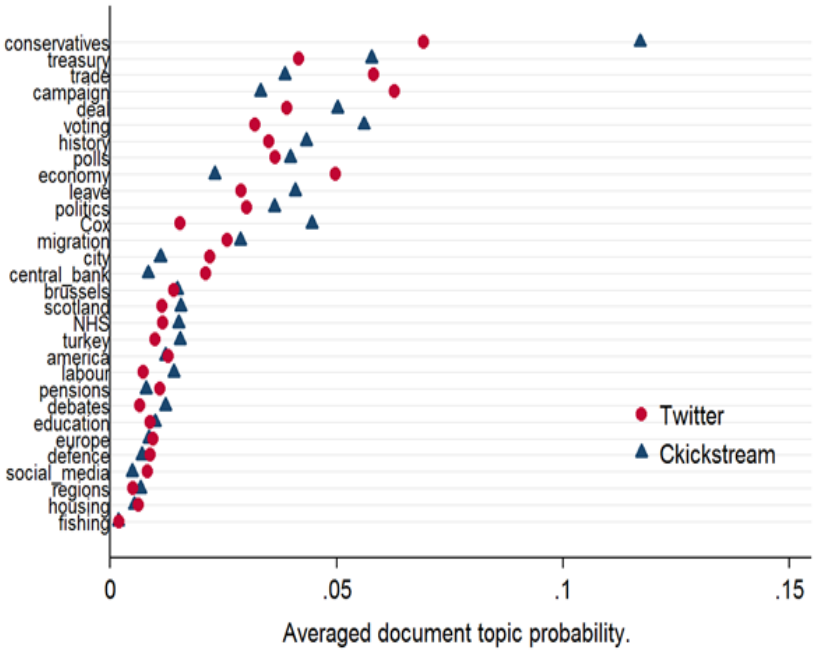
NCRM National Centre for Research Methods

What can social media data tell us?	• Many questions of interest to social scientists: candidate messages, events data, public opinion…
e.g. ideological polarisation online (Twitter)	• highly segregated partisan structure … limited connectivity..." (Conover et al. 2011)
But what happens when we look at other data sources?	• "ideological segregation … is low in absolute terms" (Gentzkow & Shapiro 2011)

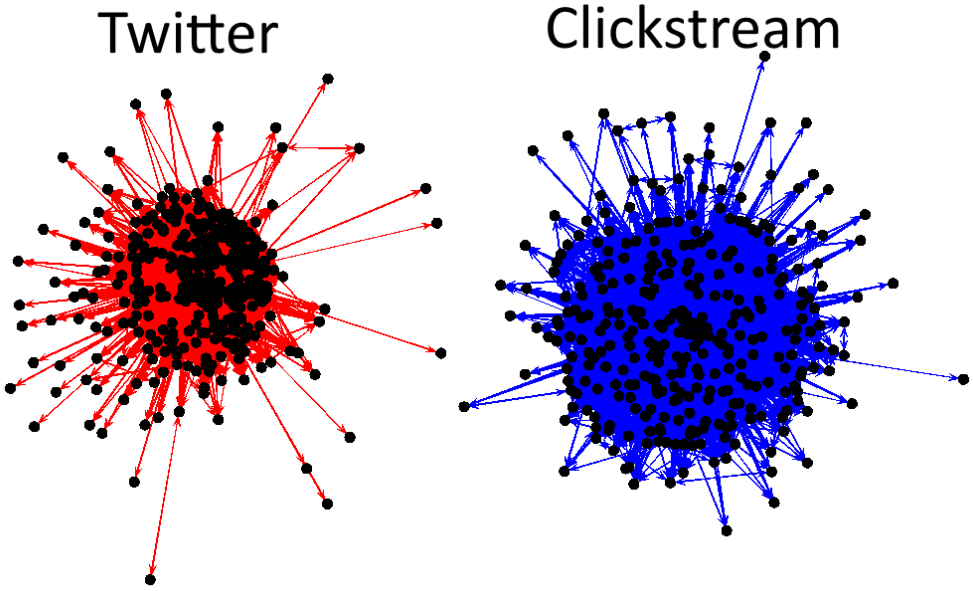
NCRM National Centre for Research Methods

Twitter vs Online Browsing: Comparing Networks & Topics on Information Exposure

Topic models



Network analysis



EXPONet

Computational methods for studying social phenomena: Comparing data sources

NCRM National Centre for Research Methods

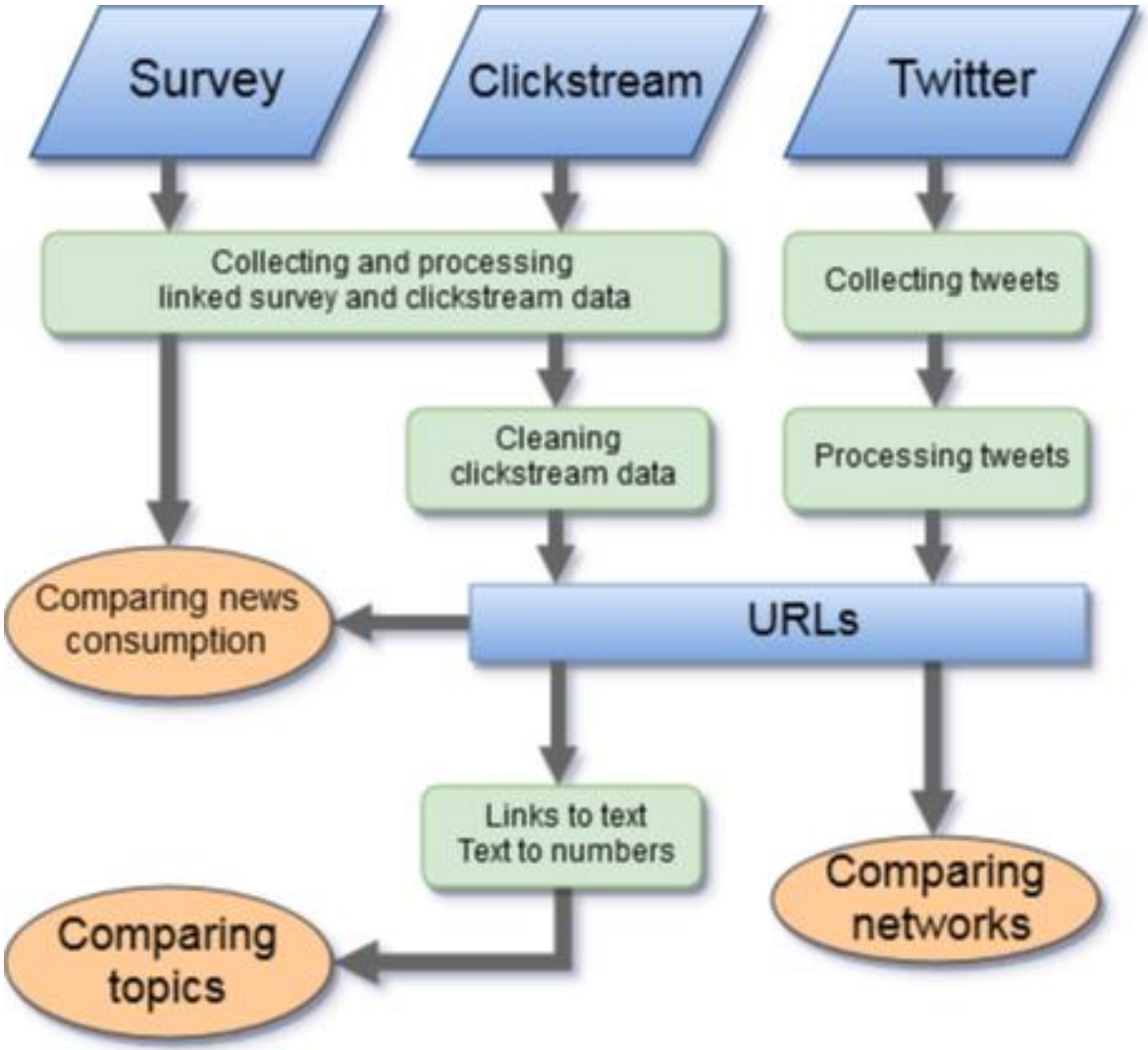
What we cover

- Measuring information exposure with multiple data sources

What we don't cover

- Anything in depth, this is an overview
- Ethical considerations

- Theory – social science, methodological or statistical, this is an overview of data and skills


Survey
Clickstream
Twitter
Collecting and processing linked survey and clickstream data
Collecting tweets
Cleaning clickstream data
Processing tweets
Comparing news consumption
URLs
Links to text
Text to numbers
Comparing networks
Comparing topics

NCRM National Centre for Research Methods

Survey (ICM)	• 3 waves: February, April, June 2016 • 1154 respondents
Browsing History (Clickstream)	• 673 respondents with clickstream/browsing history • 2 weeks per wave
Representativeness (BES f2f)	• Underrepresents retired respondents (15% to 30%), younger (avg age < 4 yrs), similar on women/men, regionally representative.

Web browsing data

Where do we get data on what people read online?

NCRM National Centre for Research Methods

Individual level

- Our data: ICM Reflected Life panel.
  - February 17 to June 23 2016
  - 3310 total users, 959 inat least one of the survey panels
- Other methods
  - Bing browser extension
  - Other apps (CosComm)

Cleaning web history data

https://mvt.api.bbc.com/buckets?activate=false

http://www.bbc.co.uk/news/components?alternativeJsLoading=true&batch%5Bmost-popular%5D%5Bid%5D=comp-most-popular&batch%5Bmost-popular%5D%5Bopts%5D%5BassetId%5D=36616028&batch%5Bmost-popular%5D%5Bopts%5D%5Bloading_strategy%5D=include_content&batch%5Bmost-popular%5D%5Bopts%5D%5Bposition_info%5D%5BinstanceNo%5D=1&batch%5Bmost-popular%5D%5Bopts%5D%5Bposition_info%5D%5BpositionInRegion%5D=4&batch%5Bmost-popular%5D%5Bopts%5D%5Bposition_info%5D%5BlastInRegion%5D=true&batch%5Bmost-popular%5D%5Bopts%5D%5Bposition_info%5D%5BlastOnPage%5D=true&batch%5Bmost-popular%5D%5Bopts%5D%5Bposition_info%5D%5Bcolumn%5D=secondary_column

http://www.bbc.co.uk/news/uk-politics-36616028

http://www.bbc.co.uk/news/special/2016/newsspec_14367/content/iframe/english/uk-parties-leave.html?v=0.5.55&initialWidth=592&childId=responsive-iframe-86457556&parentUrl=http%3A%2F%2Fwww.bbc.co.uk%2Fnews%2Fuk-politics-36616028

http://www.bbc.co.uk/news/special/2016/newsspec_14367/content/iframe/english/uk-parties-remain.html?v=0.5.55&initialWidth=592&childId=responsive-iframe-68662470&parentUrl=http%3A%2F%2Fwww.bbc.co.uk%2Fnews%2Fuk-politics-36616028

http://www.bbc.co.uk/news/special/2016/newsspec_14367/content/iframe/english/uk-turnout.html?v=0.5.55&initialWidth=592&childId=responsive-iframe-3191489&parentUrl=http%3A%2F%2Fwww.bbc.co.uk%2Fnews%2Fuk-politics-36616028

http://www.bbc.co.uk/news/special/2016/newsspec_14367/content/iframe/english/uk-winners.html?v=0.5.55&initialWidth=592&childId=responsive-iframe-36415436&parentUrl=http%3A%2F%2Fwww.bbc.co.uk%2Fnews%2Fuk-politics-36616028

http://www.bbc.co.uk/news/special/2016/newsspec_14368/content/iframe/english/index.html?v=0.2.35&initialWidth=592&childId=responsive-iframe-72190634&parentUrl=http%3A%2F%2Fwww.bbc.co.uk%2Fnews%2Fuk-politics-36616028

https://ssl.bbc.co.uk/idcta/init?policy=notification&buttonColour=white&ptrt=http%3A%2F%2Fwww.bbc.co.uk%2Fnews%2Fuk-politics-36616028

http://www.bbc.co.uk/news/components?alternativeJsLoading=true&batch%5Bfrom-other-news-sites%5D%5Bid%5D=comp-from-other-news-sites&batch%5Bfrom-other-news-sites%5D%5Bopts%5D%5BassetId%5D=36616028&batch%5Bfrom-other-news-sites%5D%5Bopts%5D%5Bconditions%5D%5B%5D=is_local_page&batch%5Bfrom-other-news-sites%5D%5Bopts%5D%5Bloading_strategy%5D=include_content&batch%5Bfrom-other-news-sites%5D%5Bopts%5D%5Basset_id%5D=uk-politics-36616028&batch%5Bfrom-other-news-sites%5D%5Bopts%5D%5Bposition_info%5D%5BinstanceNo%5D=1&batch%5Bfrom-other-news-sites%5D%5Bopts%5D%5Bposition_info%5D%5BpositionInRegion%5D=7&batch%5Bfrom-other-news-sites%5D%5Bopts%5D%5Bposition_info%5D%5BlastInRegion%5D=true&batch%5Bfrom-other-news-sites%5D%5Bopts%5D%5Bposition_info%5D%5BlastOnPage%5D=false&batch%5Bfrom-other-news-sites%5D%5Bopts%5D%5Bposition_info%5D%5Bcolumn%5D=primary_column&batch%5Bfrom-other-news-sites%5D%5Btemplate%5D=%2Fcomponent%2Ffrom-other-news-sites

Cleaning web history data

https://mvt.api.bbc.com/buckets?activate=false

http://www.bbc.co.uk/news/components?alternativeJsLoading=true&batch%5Bmost-popular%5D%5Bid%5D=comp-most-popular&batch%5Bmost-popular%5D%5Bopts%5D%5BassetId%5D=36616028&batch%5Bmost-popular%5D%5Bopts%5D%5Bloading_strategy%5D=include_content&batch%5Bmost-popular%5D%5Bopts%5D%5Bposition_info%5D%5BinstanceNo%5D=1&batch%5Bmost-popular%5D%5Bopts%5D%5Bposition_info%5D%5BpositionInRegion%5D=4&batch%5Bmost-popular%5D%5Bopts%5D%5Bposition_info%5D%5BlastInRegion%5D=true&batch%5Bmost-popular%5D%5Bopts%5D%5Bposition_info%5D%5BlastOnPage%5D=true&batch%5Bmost-popular%5D%5Bopts%5D%5Bposition_info%5D%5Bcolumn%5D=secondary_column

http://www.bbc.co.uk/news/uk-politics-36616028

http://www.bbc.co.uk/news/special/[illegible]/newsspec_14367/content/iframe/english/uk-parties-leave.html?v=0.5.55&initialWidth=592&childId=responsive-iframe-86457556&parentUrl=http%3A%2F%[illegible]ww.bbc.co.uk%2Fnews%2Fuk-politics-36616028

http://www.bbc.co.uk/news/special/[illegible]/newsspec_14367/content/iframe/english/uk-parties-remain.html?v=0.5.55&initialWidth=592&childId=responsive-iframe-68662470&parentUrl=http%3A%2F%[illegible]ww.bbc.co.uk%2Fnews%2Fuk-politics-36616028

http://www.bbc.co.uk/news/uk-politics-36616028

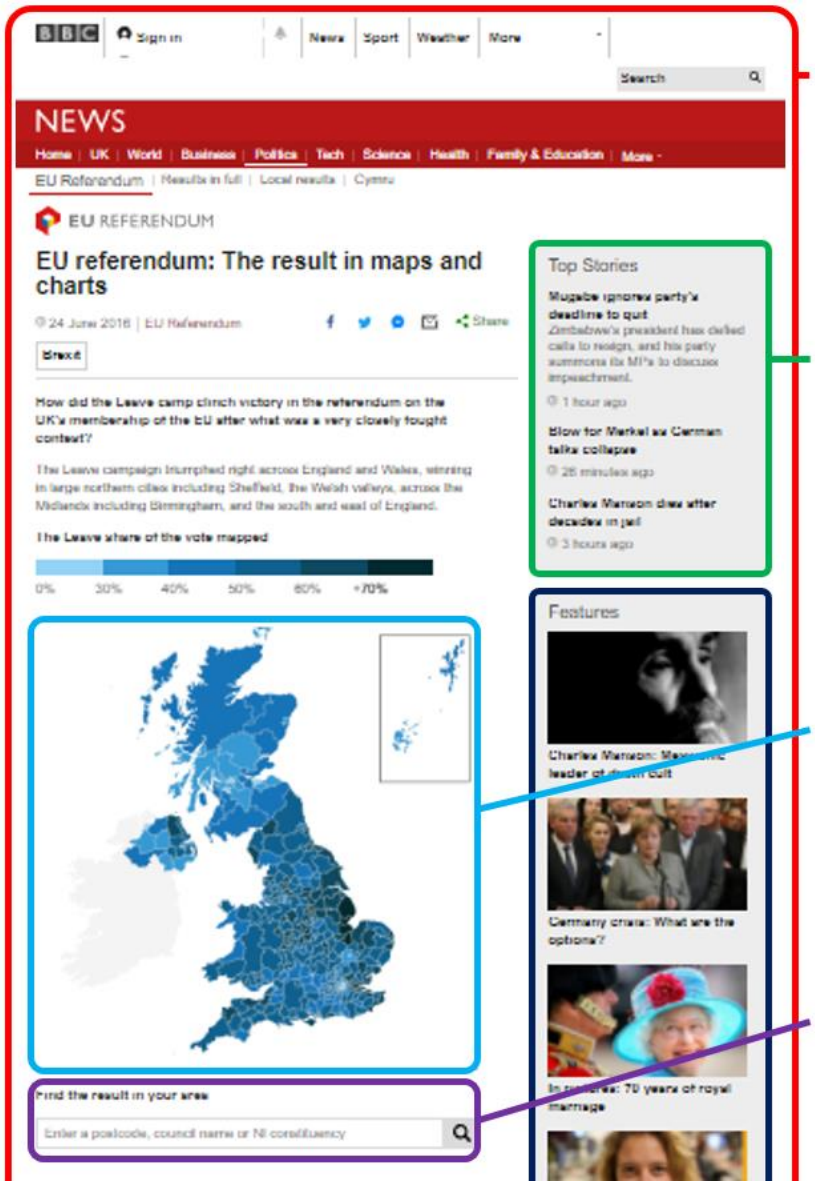
http://www.bbc.co.uk/news/special/2016/newsspec_14367/content/iframe/english/uk-turnout.html?v=0.5.55&initialWidth=592&childId=responsive-iframe-[illegible]

http://www.bbc.co.uk/news/special/2016/newsspec_14367/content/iframe/english/uk-winners.html?v=0.5.55&initialWidth=592&childId=responsive-iframe-36415436&parentUrl=http%3A%2F%2Fwww.bbc.co.uk%2Fnews%2Fuk-politics-36616028

http://www.bbc.co.uk/news/special/2016/newsspec_14368/content/iframe/english/index.html?v=0.2.35&initialWidth=592&childId=responsive-iframe-72190634&parentUrl=http%3A%2F%2Fwww.bbc.co.uk%2Fnews%2Fuk-politics-36616028

https://ssl.bbc.co.uk/idcta/init?policy=notification&buttonColour=white&ptrt=http%3A%2F%2Fwww.bbc.co.uk%2Fnews%2Fuk-politics-36616028

http://www.bbc.co.uk/news/components?alternativeJsLoading=true&batch%5Bfrom-other-news-sites%5D%5Bid%5D=comp-from-other-news-sites&batch%5Bfrom-other-news-sites%5D%5Bopts%5D%5BassetId%5D=36616028&batch%5Bfrom-other-news-sites%5D%5Bopts%5D%5Bconditions%5D%5B%5D=is_local_page&batch%5Bfrom-other-news-sites%5D%5Bopts%5D%5Bloading_strategy%5D=include_content&batch%5Bfrom-other-news-sites%5D%5Bopts%5D%5Basset_id%5D=uk-politics-36616028&batch%5Bfrom-other-news-sites%5D%5Bopts%5D%5Bposition_info%5D%5BinstanceNo%5D=1&batch%5Bfrom-other-news-sites%5D%5Bopts%5D%5Bposition_info%5D%5BpositionInRegion%5D=7&batch%5Bfrom-other-news-sites%5D%5Bopts%5D%5Bposition_info%5D%5BlastInRegion%5D=true&batch%5Bfrom-other-news-sites%5D%5Bopts%5D%5Bposition_info%5D%5BlastOnPage%5D=false&batch%5Bfrom-other-news-sites%5D%5Bopts%5D%5Bposition_info%5D%5Bcolumn%5D=primary_column&batch%5Bfrom-other-news-sites%5D%5Btemplate%5D=%2Fcomponent%2Ffrom-other-news-sites



http://www.bbc.co.uk/news/uk-politics-36616028

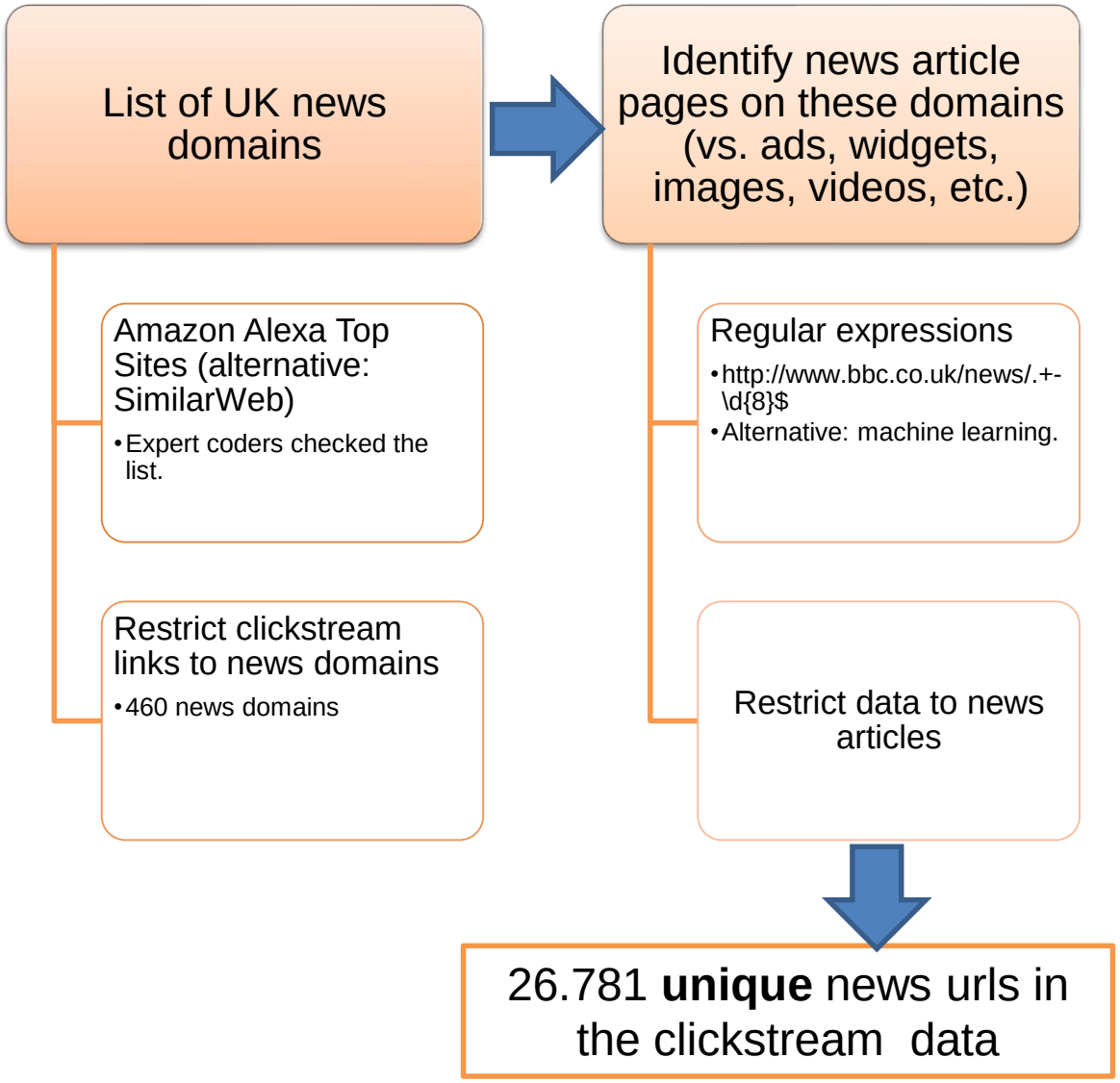
http://www.bbc.co.uk/news/components?alternativeJsLoading=true&batch%5Bmost-popular%5D%5Bid%5D=comp-most-popular&batch%5Bmost-popular%5D%5Bopts%5D%5BassetId%5D=36616028&batch%5Bmost-popular%5D%5Bopts%5D%5Bloading_strategy%5D=include_content&batch%5Bmost-popular%5D%5Bopts%5D%5Bposition_info%5D%5BinstanceNo%5D=1&batch%5Bmost-popular%5D%5Bopts%5D%5Bposition_info%5D%5BpositionInRegion%5D=4&batch%5Bmost-popular%5D%5Bopts%5D%5Bposition_info%5D%5BlastInRegion%5D=true&batch%5Bmost-popular%5D%5Bopts%5D%5Bposition_info%5D%5BlastOnPage%5D=true&batch%5Bmost-popular%5D%5Bopts%5D%5Bposition_info%5D%5Bcolumn%5D=secondary_column

http://www.bbc.co.uk/news/special/2016/newsspec_14367/content/iframe/english/uk-parties-leave.html?v=0.5.55&initialWidth=592&childId=responsive-iframe-86457556&parentUrl=http%3A%2F%2Fwww.bbc.co.uk%2Fnews%2Fuk-politics-36616028

http://www.bbc.co.uk/news/special/2016/newsspec_14368/content/iframe/english/index.html?v=0.2.35&initialWidth=591&childId=responsive-iframe-72190634&parentUrl=http%3A%2F%2Fwww.bbc.co.uk%2Fnews%2Fuk-politics-36616028

Cleaning clickstream data

How do we identify news page urls?



Collecting Twitter data

The method you choose depends on:

- Your programming skills or willingness to develop these skills
- The characteristics of the data you want to collect
- Your budget

NCRM National Centre for Research Methods

Twitter APIs

- Some programming skills required
- Many available packages in Python (tweepy) and R (twitteR)
- Free scripts and tutorials: **ExpoNET tools**
- Flexible but constraints on data collected.

Commercial or free software

- Chorus, NodeXL, Voson, etc.
- Easy access
- Some include data analysis options
- Less flexible
- Same constraints as the API

Purchasing data from Twitter

- Convenient but very expensive

Processing social media data


```
{
  "created_at": "Thu Apr 06 15:24:15 +0000 2016",
  "id": 850006145121695744,
  "id_str": "850006145121695744",
  "text": "Support our campaign! nhttps://t.co/XweGngmxlP",
  "user": {
    "id": 2244994945,
    "id_str": "2244994945",
    "screen_name": "Campaigner",
    "followers_count": 47684,
    "friends_count": 1524,
    "favourites_count": 251,
    "statuses_count": 3121
  },
  "entities": {
    "urls": [{
      "indices": [
        32,
        52
      ],
      "url": "http://t.co/IOwBrTZR",
      "display_url": "youtube.com/watch?v=oHg5SJ…",
      "expanded_url": "http://www.campaign.com"
    }]
  }
}
```


NCRM National Centre for Research Methods

Working with .json files.

- Hierarchical data.
- More efficient.
- Scales up.

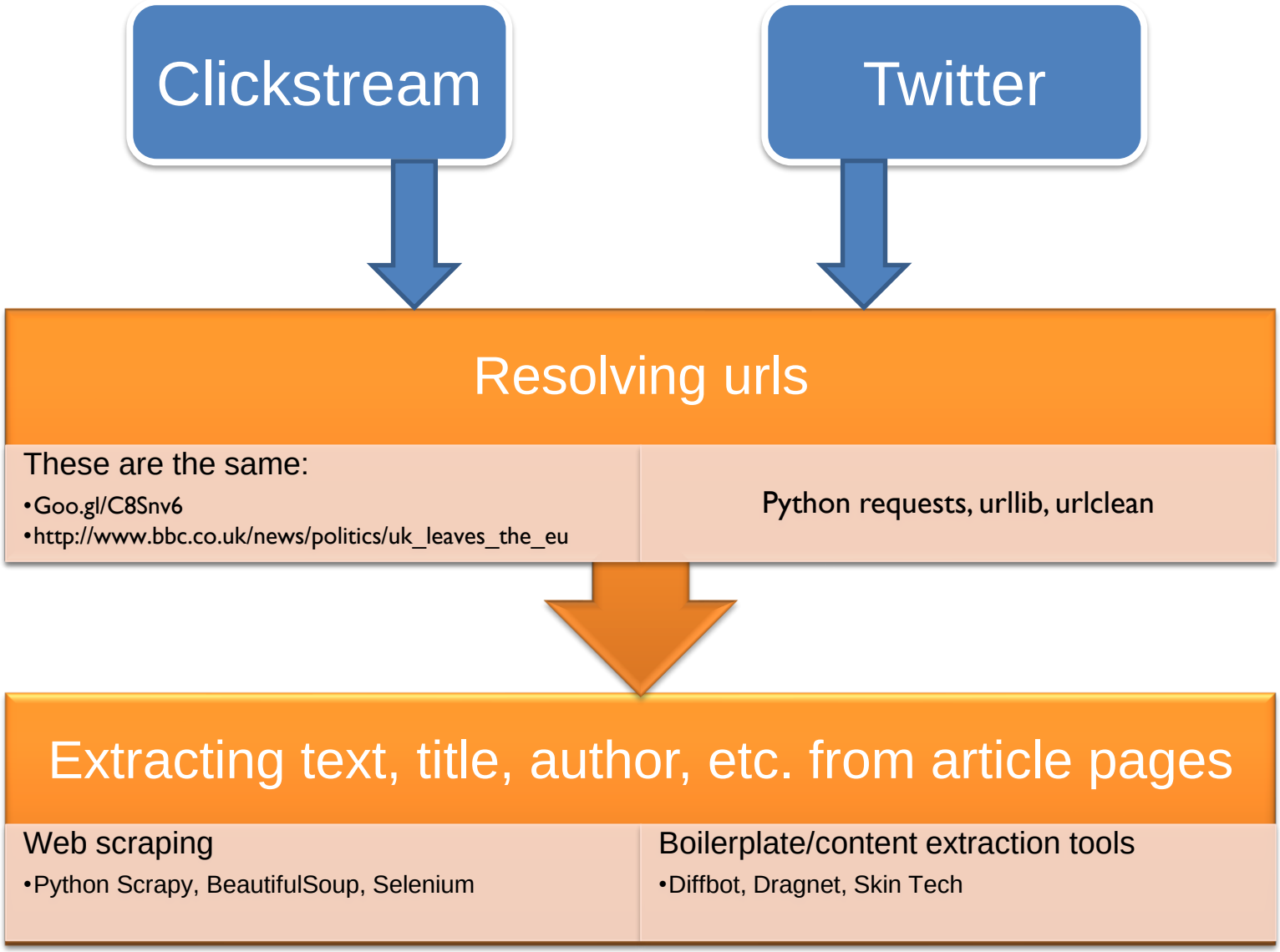
Data storage

- Relational databases
  - MySQL
- Non-relational databases
  - MongoDB

Extracting relevant information

- Text
- User information
- Lists of friends and followers
- **Links shared**

1.6 million tweets with links to **news domains**


Clickstream
Twitter
Resolving urls
These are the same:
•Goo.gl/C8Snv6
•http://www.bbc.co.uk/news/politics/uk_leaves_the_eu
Python requests, urllib, urlclean
Extracting text, title, author, etc. from article pages
Web scraping
•Python Scrapy, BeautifulSoup, Selenium
Boilerplate/content extraction tools
•Diffbot, Dragnet, Skin Tech

Extracting information from text

From words to numbers.

NCRM National Centre for Research Methods

Text processing

- Standard text cleaning and NLP methods
  - Lower case, removing punctuation and stopwords, stemming, tokenization
  - Part of speech tagging, n-grams

Counting keywords

Measuring distances/similarities

Topic extraction

Topic models

Uncovering hidden thematic structures in documents.

NCRM National Centre for Research Methods

Latent Dirichlet Allocation (LDA)

- Documents are a mixture of topics.
- Topics generate words based on their probability distribution.
- Algorithm:
  - Determine number of words in document.
  - Determine mixture of topics in document.
  - Based on the topics' multinomial distribution, assign words to documents.

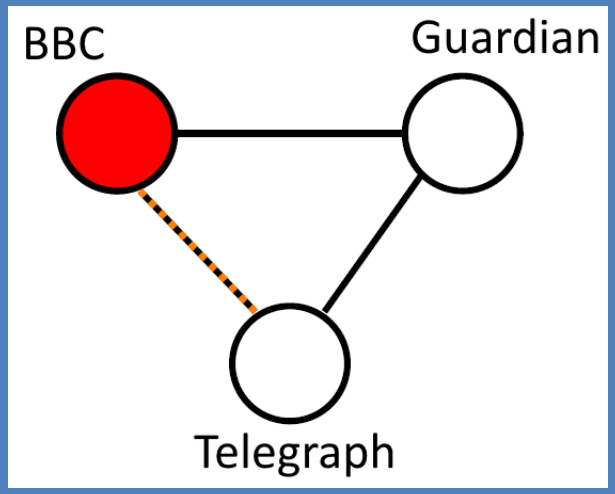
Mallet, Python Gensim, R Quanteda.

NCRM National Centre for Research Methods

Topics in our corpus

label	keys
conservatives	cameron,eu,campaign,johnson,leave,minister,prime,boris,referendum,david,vote,remain,tory,brexit,leader,former,gove,secretary,britain,cor
treasury	eu,leave,vote,brexit,osborne,economy,economic,uk,leaving,treasury,chancellor,britain,campaign,government,nhs,george,remain,public,budget
trade	eu,uk,trade,brexit,market,britain,european,countries,single,leave,free,europe,economic,access,leaving,british,world,economy,membership,r
campaign	britain,european,eu,union,british,leave,vote,referendum,brexit,remain,campaign,london,minister,cameron,bloc,united,june,country,prime,eu
deal	eu,government,european,deal,uk,referendum,new,parliament,law,member,states,union,may,court,cameron,state,bill,minister,agreement,two,leg
voting	vote,referendum,polling,people,result,results,june,uk,remain,voters,leave,voting,day,electoral,votes,eu,local,wales,election,station,clo
history	europe,european,britain,british,war,years,world,history,common,union,economic,brussels,one,continent,project,first,peace,century,since,c
polls	remain,leave,poll,polls,vote,cent,voters,eu,campaign,referendum,lead,brexit,last,support,point,camp,showed,survey,two,points,people,ahea
economy	uk,growth,brexit,year,economy,uncertainty,referendum,term,economic,may,investment,cent,likely,impact,vote,months,last,since,global,marke
leave	eu,britain,leave,europe,vote,european,remain,people,country,uk,us,union,stay,control,world,jobs,want,brussels,back,trade,better,british,
politics	political,party,even,one,may,many,politics,right,campaign,left,brexit,much,public,might,far,politicians,yet,anti,power,election,case,der
Cox	cox,jo,mp,police,death,labour,man,murder,west,attack,two,first,family,people,old,year,yorkshire,shot,parliament,children,campaigning,kil
migration	eu,immigration,uk,ireland,migration,people,citizens,british,work,migrants,irish,northern,living,workers,border,britain,movement,year,sys
city	business,london,financial,companies,banks,city,chief,businesses,company,investment,firms,executive,industry,group,jobs,bank,staff,servic
central_bank	bank,rates,rate,central,fed,interest,banks,financial,policy,global,markets,economy,monetary,england,economic,japan,inflation,soros,feder
brussels	eu,britain,brussels,president,leaders,european,brexit,cameron,juncker,british,summit,tusk,commission,talks,may,french,merkel,jean,le,jun
scotland	scotland,scottish,independence,referendum,party,uk,vote,snp,sturgeon,england,labour,first,second,nicola,eu,majority,leader,election,scot
NHS	health,nhs,public,new,government,service,energy,funding,environment,local,care,work,services,environmental,information,use,doctors,chang
turkey	turkey,eu,refugees,migrants,join,free,security,country,turkish,borders,crisis,people,migration,europe,control,joining,border,intelligenc
america	trump,obama,president,us,donald,united,states,barack,britain,clinton,presidential,republican,america,american,trade,world,british,visit,
labour	labour,corbyn,party,jeremy,leader,mps,leadership,shadow,june,people,report,mp,members,today,eagle,anti,launch,israel,supporters,stand,fo
pension_funds	money,investors,funds,fund,pension,cash,buy,investment,market,year,vote,markets,may,companies,asset,income,stock,years,bonds,buying,asse
debates	bbc,debate,audience,news,referendum,ms,live,davidson,tv,questions,question,programme,radio,june,ruth,sides,editor,coverage,pen,night,sta
education	young,university,people,women,students,research,professor,school,education,eu,student,science,study,debate,politics,year,generation,men,
europe	germany,europe,german,france,euro,european,countries,spain,greece,eurozone,italy,crisis,french,debt,spanish,country,sweden,italian,greel
defence	security,russia,nato,military,eu,defence,putin,russian,foreign,former,army,china,us,world,europe,european,international,war,policy,ukra
social_media	twitter,facebook,com,news,video,tech,google,click,http,share,youtube,technology,media,users,app,watch,content,online,https,us,social,ema
regions	north,remain,east,south,john,david,james,leave,director,jones,conservative,andrew,green,sir,paul,davies,peter,mark,philip,chris,officer,
housing	prices,house,property,london,housing,market,price,homes,mortgage,buyers,average,home,new,demand,first,value,buy,cent,houses,year,rates,c
fishing	fishing,farage,fish,fisheries,nigel,industry,fishermen,june,ukip,river,day,guy,v,policy,stocks,quotas,sea,common,water,llc,fisherman,may

Building networks



NCRM National Centre for Research Methods

What is a network?

- Mathematical construct of nodes and edges.
- Information (attributes) on the nodes as well as the edges.

In our case:

- News domains are nodes
- Edges are formed between two domains if a user shares/reads an article on both domains.

Why model it like this?

- Explicitly model interdependencies between observations.
- Patterns of information exposure
  - Detect communities/echo chambers

Data linking

Multiple levels.

NCRM National Centre for Research Methods

Survey and clickstream data

- User level

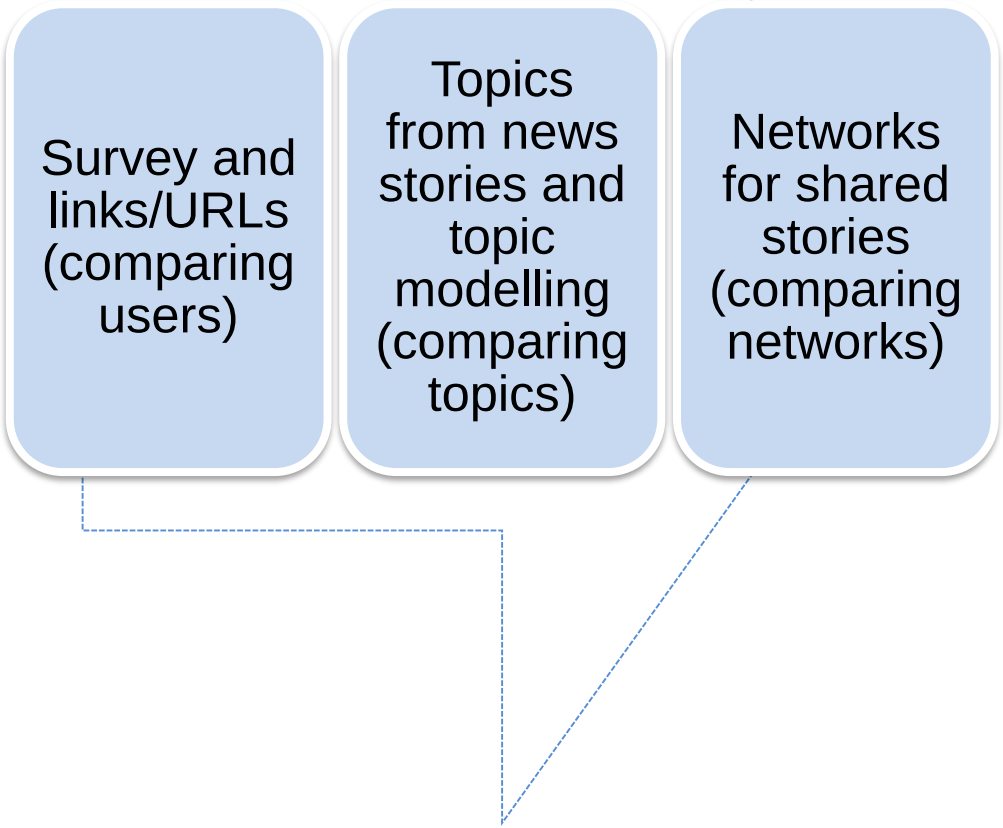
Clickstream and social media data

- Domain level
- Article level

Survey, clickstream and social media data

- Domain level
- Article level

Following slides show results from:



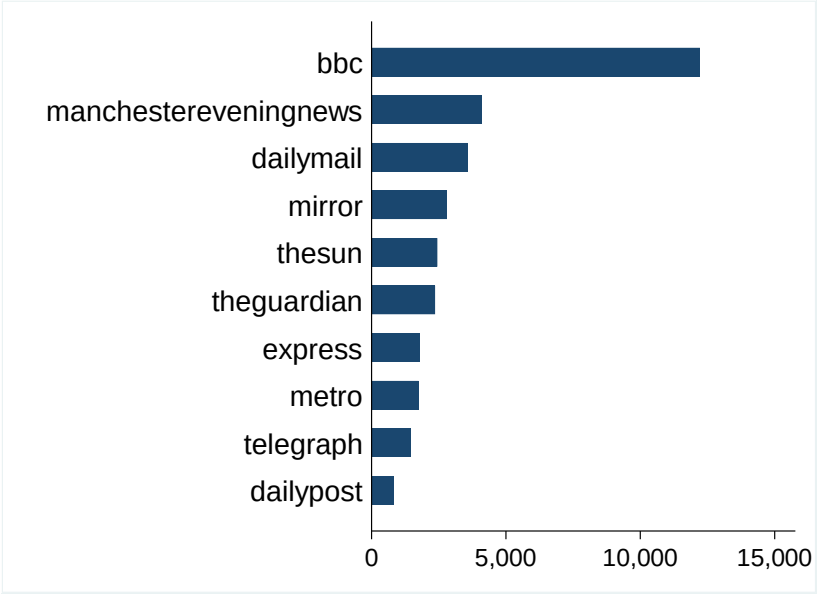
Comparing Users: Clickstream Data After Processing

56,289 data points

	username	time	domain
10.	22313	2016-02-19 05:15:38	www.dailymail.co.uk

url
http://www.dailymail.co.uk/news/article-3453825/Simple-brainteasers-bamboozle-adults-Test-15-noodle-scratching-conundrums.html

Create Quantities of Interest: Total URLs visited, text analysis of URLs, Left/Right URLs,



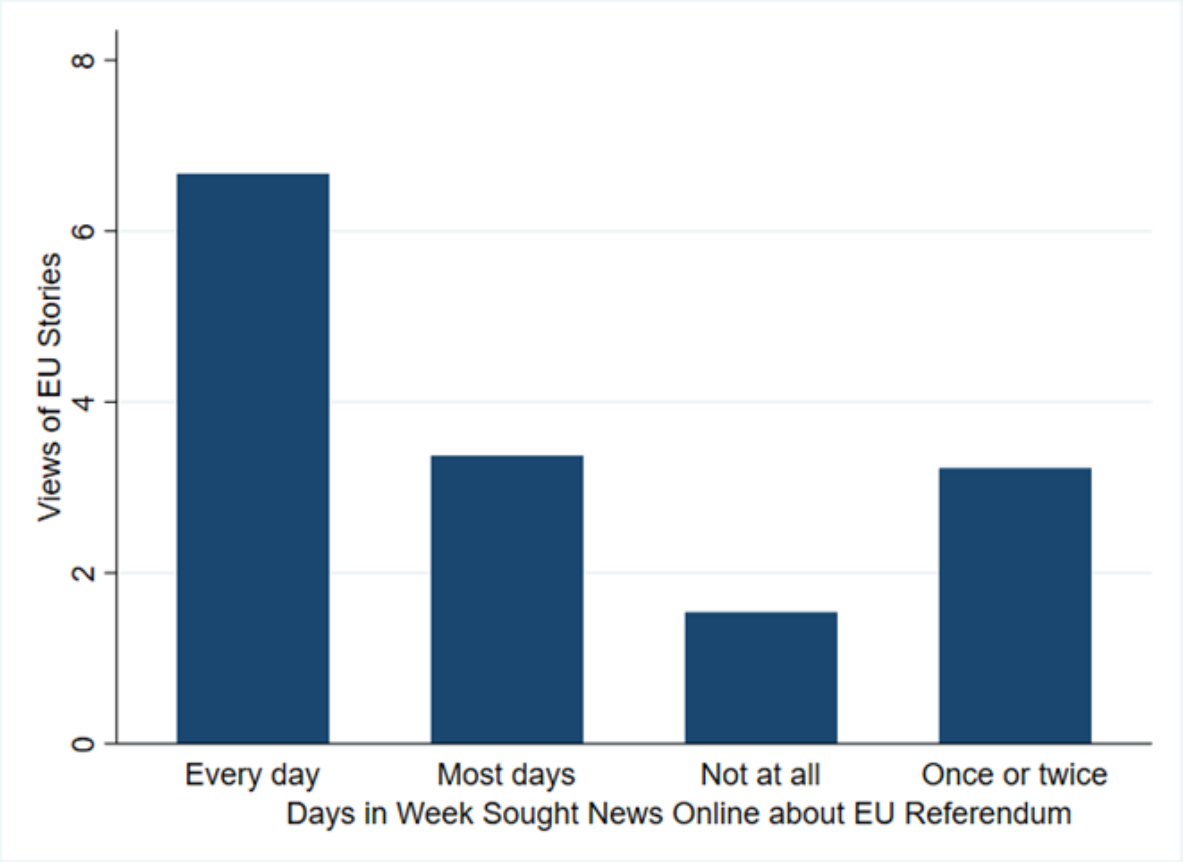
Comparing Users: Reported Use & URL visits

Clickstream processed URLs

- Quantity of interest – count of visited news URLs per user about EU

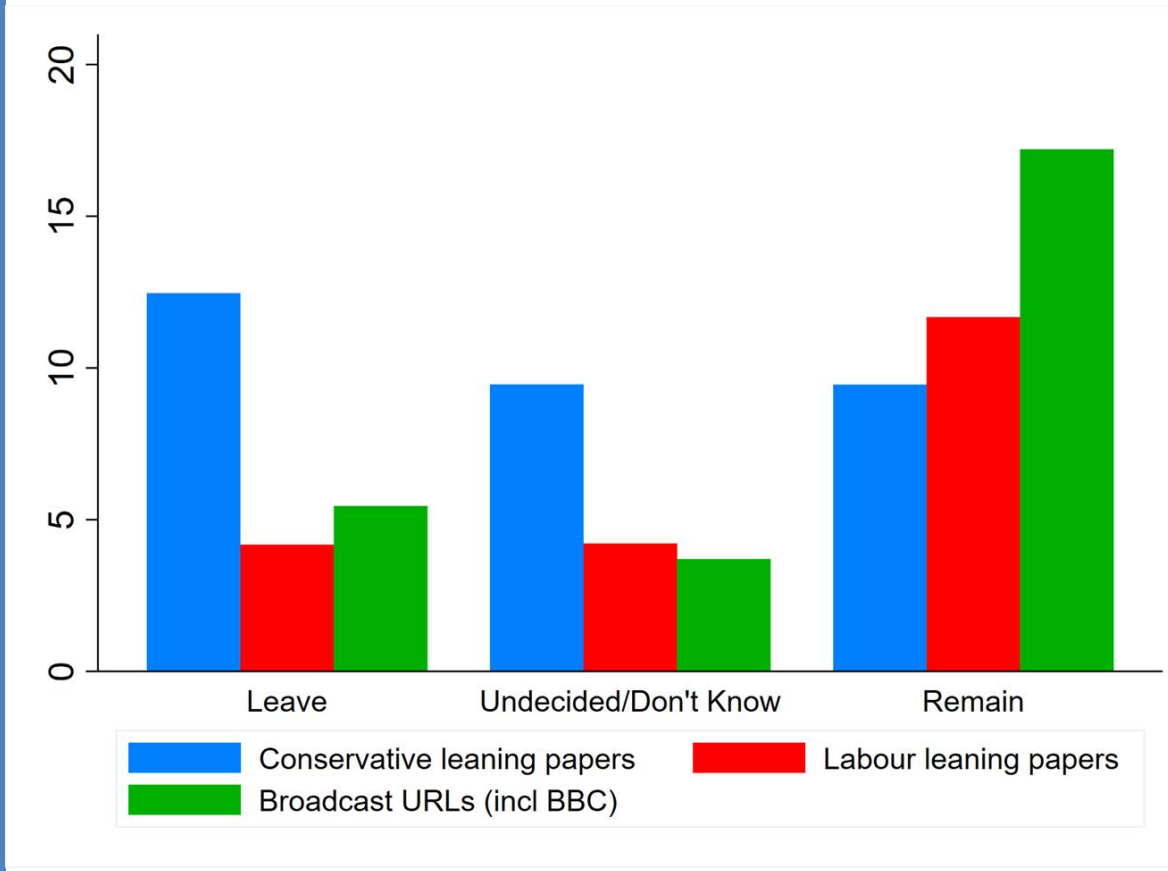
Survey Data

- Reported online news exposure
- Over the **past 7 days**, how often have you done the following with respect to the EU Referendum…?



Comparing users: Brexit preference and news exposure.

Bars indicate mean number of views per user during fieldwork period (n = 513).

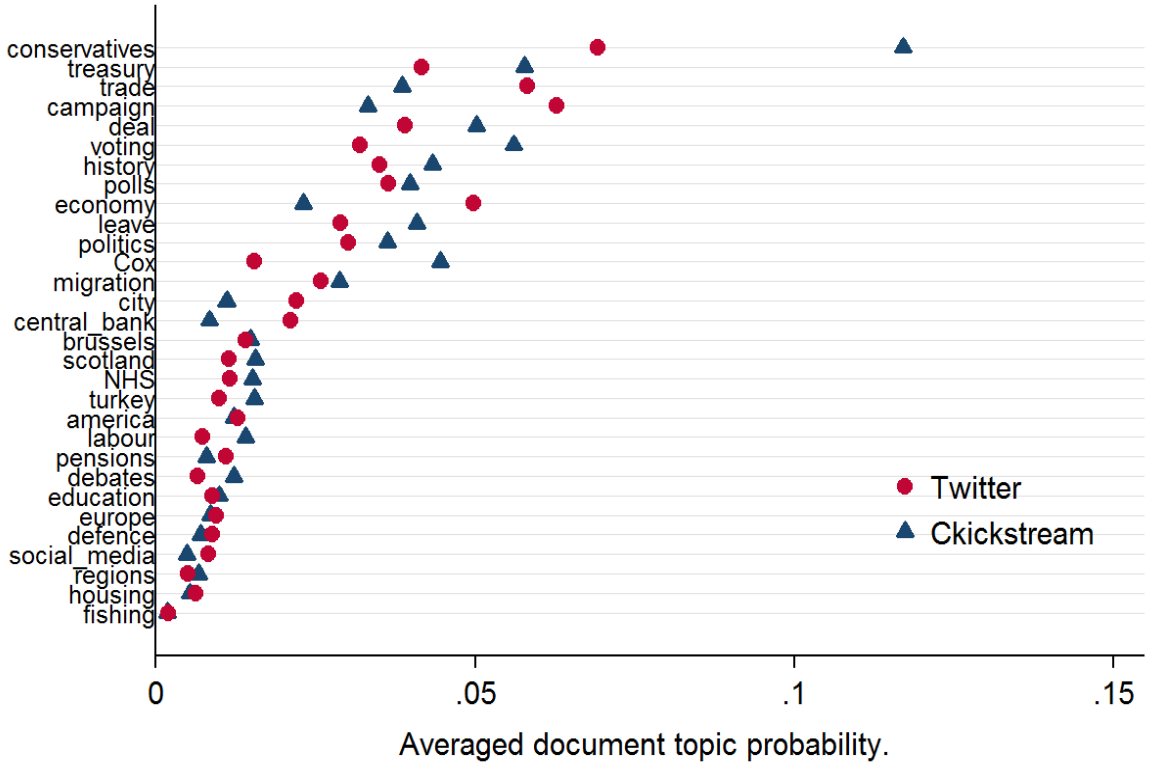


Comparing Content

What do we learn about information exposure by comparing content/topics from news stories shared on Twitter to news stories viewed online?

NCRM National Centre for Research Methods

Content

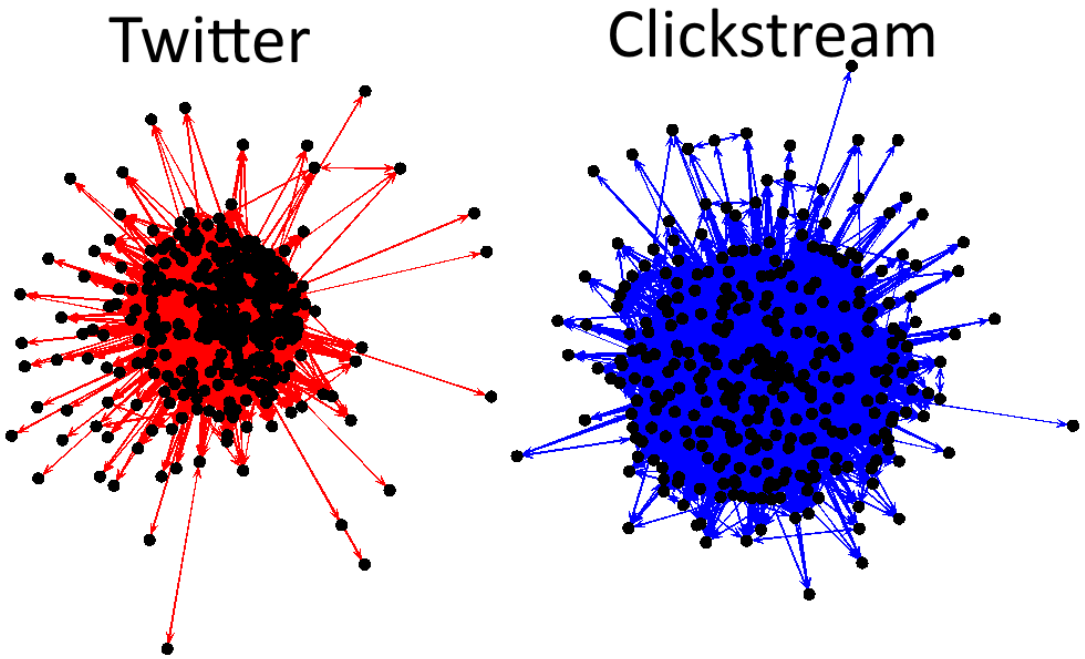


Comparing networks

What do we learn about information exposure by comparing news exposure from browsing histories to social media data?

NCRM National Centre for Research Methods

Networks



Methodological challenges in using online data.

NCRM National Centre for Research Methods

Can we conclude that social media leads to segregation in news exposure based on 40million Tweets?

Challenges: Representativeness, Measurement error & causal inference

Multiple sources of linked data as a possible solution.

Additional resources

- https://mediaeffectsresearch.wordpress.com/
- Annotated bibliography, notes from presentation
- Jupyter notebooks