Online data sources: Linking old and new, big and small

I'm Susan Banducci from the Exeter Qstep Center at the University of Exeter and I am Iulia Cioroianu from the Qstep Center at the University of Exeter.

Today we'll be talking about online data sources and how you use those to answer questions about information exposure. What can social media data tell us?

There are a lot of interesting social science questions that social media data can answer for us if we're interested in questions about candidate statements or candidate positions if we're interested in events data or even perhaps if we're interested in public opinion social media data might be able to help us answer those questions. One of the questions that where social media data have been used is to look at ideological polarization and using Twitter data some have found that the population is highly segregated in terms of its partisan structure and there's little connectivity between different ends of the ideological spectrum but what if we look at other sources of information about online use of data and information. If we look at other sources of data do we still see the same picture of ideological polarization and it turns out that when we look at other sources of data we see much less ideological polarization so for example when we look at web browsing histories we see less ideological polarization.

So we can look at social media data and those are platforms where people are sharing user-generated content but also often sharing links to other news stories or other information we can also look at online browsing histories where users will select to be exposed to certain news and information sources so for example they may directly click on their favourite news source or they may get to news and information sources through a news aggregator and they also may get to news and information sources by using a search engine. So they may click or enter topics of interest and that will return news and information stories to them. So that's another way of looking at how people are exposed to information online.

A third way that we can examine news and information exposure is to ask people in surveys what they choose to see online and through social media platforms so it's these three sources of information that we will be looking at in this talk. And what we will cover in terms of these sources of information is how we extract the data from these sources how we process the raw data and then finally how we analyze these data and our interest today will be in how we compare across these sources of data and what those sources of data tell us whether they're the same or different about news and information exposure. So this is the flow chart of what we've been looking at today the blue boxes represent the sources of data surveys clickstream data and twitter data the green boxes are how we process the data before we analyze the data and the orange circles represent the outputs of the analysis and the processing of the data. So this flowchart gives you an idea of what we will be covering. And I will start by talking about survey data and web browsing or clickstream data. In the example that we use we will be relying on survey data that was collected around the EU referendum in the UK in June of 2016 and we collected survey data at three points in time during the campaign leading up to the vote.

We had surveyed over a thousand respondents about their opinions about Brexit and about their use of news and information during the campaign. We also found of those survey respondents of those who consented to participate in this study that they had installed a web browsing app that collected all the sources and information that they saw online so it is a collection and history of their online browsing and it captured all sorts of information about

not only their news and information exposure but where they else they visited on the web and that is a matter that we have to sort out all that extra data when we are cleaning the data. One of the issues that we want to address when we look at the data are how representative these data are both in terms of the UK population as well as how representative the data are of what people are seeing online. In terms of the sample of respondents that we have to the survey data they reflect what we generally know about online samples. They tend to be younger they also tend to be more politically interested and we see those sorts of characteristics in the sample of data that we have.

The first thing that we need to do after we collect the clickstream data is clean it and it does require extensive cleaning because what we get is not the exact URL that that the user has clicked on and that we are interested in but a string of URLs that go through the server and which are happening in the background out of those URLs we need to filter the one of interest. So what are these are their URLs for example you can have a page with a map there are URLs that correspond to that map that are loading in the background there are also other URLs that corresponds to the widget at the bottom or URLs that correspond to the two other articles on the website that that are displayed as links on the side however what we care about is the URL at the very top which corresponds to the page that the user is actually reading at the moment which has the text and the title and the information that they are extracting from that page. So out of all these other links that may be on a similar domain we want to identify the one that corresponds to the article itself.

So how do we do that? First we start with a list of UK news domains we get that list from Amazon Alexa which is an Amazon service that ranks the most popular domains among users in a certain country. The advantage of Alexa is also that it provides topic specific domains so we do get a list of the 500 most popular news domains among users in the UK. We start with that list then we have expert coders go through the list and eliminate those that do not include any political information. We end up with a list of 460 domains news domains that we are working with. The next step is to identify news articles on those domains as opposed to what we saw before adds widgets images and videos to do that we take advantage of the fact that most newspaper websites have a very clearly structured database in the background where they store articles but also videos and photos and all these other links that we saw previously in the data. And so we can write regular expressions patterns that match the position of articles in the database.

So for example on the BBC website the location where they store articles it is at www.bbc.co.uk/news/the title of the article followed by an 8 digit number. So this is what we care about we care about the news on that website and capturing those articles that correspond to news. And so we can write a regular expression that matches that exact pattern and we can then do that for all of the 460 domains that we've identified with Alexa and filter our clickstream data to only include articles on those domains and when we do that we end up with a list of about 26 thousand unique news URLs in our clickstream data. We have our survey data and we have our clickstream data the next type of data that we are going to work with is the Twitter data.

So how do we collect Twitter data there are multiple methods and choosing one will depend on your programming skills or your willingness to develop a new programming skills it also depends on the characteristics of the data that you are trying to collect and it also depends on your budget. The most popular method for collecting Twitter data among researchers is by using the Twitter API. This method does require some programming skills but there are many packages in Python and in R that are available for you. There are also free scripts and tutorials available online and the ExpoNet team at University of Exeter has made available a number of notebooks and tutorials on our GitHub page and on our on our blog on how to collect and download Twitter data. The advantage of using the APIs is is that they are flexible

however there are constraints on the data to be collected such as the total number of tweets that you can collect the number of tweets that you can collect at a single moment in time and how far back in time you can go. Another method for collecting Twitter data is web scraping. This one does require slightly more advanced programming skills however it is more flexible and it allows you to get slightly more data so for example go further back in time than you would by going through the APIs however you still have to abide by the Twitter's rules of service.
Another method for data collection is to use commercial or free software and there are a number of software packages available some of them are free some of them you have to pay for Chorus is one that is developed for research purposes NodeXL, Voson are others these packages provide easy access to the data. Some even include some data analysis options however the disadvantage is that they are less flexible than the methods above and you still have to follow the same constraints as you would if you're going through the API or if you're scraping the web for it.
And finally the last method would be to purchase data from Twitter itself. This one is very convenient you get all the data that you can that you want you can even go back in time to historical data however it is very expensive. So now that you collected your Twitter data the next step that you have to take is to process that data. If you went through the Twitter API most likely the type of data that you are going to get back is stored as JSON files. This is a specific file format that includes a string of dictionaries JSON data JSON files are very good for using for working with hierarchical data such as the one that we are getting from Twitter. They are more efficient and it's easy to scale up. The way we store those types of files you can either store them in a relational database such as MySQL database or more efficiently you can store them in a non-relational database which is MongoDB.
After you store them what you want is to be able to extract the relevant information the information that is relevant to you such as the text of the tweet or the user information or maybe you want a list of friends and followers or all the other the other elements that are included in Twitter data. What we care about and what we are extracting from our data are the links that users are sharing in their tweets. And when we do that we end up from a database of 74 million tweets that were about Brexit and that we collected in the weeks before the Brexit referendum we end up with about 1.6 million tweets that have links to the news domains that we've identified in Alexa. So this is the dataset that we are working with on Twitter data. We have our clickstream data and we have that the links from the tweets so we have links from clickstream data and we have links from tweets the next thing that we have to do is resolve those URLs those links. What does that mean? So for example you can have an article in the data that is on the BBC website the title is UK leaves the EU. The same article however can show up in the data is a shortened link. We want to make sure that in the end we end up with a single article that ideally is the expanded version of that so we have to resolve the URLs and expand all of them and get to the expanded version.
There are multiple ways of doing that we use Python and a number of packages in Python to resolve the URLs and expand them and get to a single link in our data. Once we get to a single link that corresponds to the article that we are interested in we need to extract from that article the text the title the author and other relevant information. We can do that by using web scraping so we can visit that page and we can web scrape the content from that page into a database that has all these fields that we care about. However this is a lengthy process and since most websites don't have an identical structure we would have to do that for all the 460 domains that we've identified in our data. So we outsource that and we rely on a tool that does that better than we could ever do it for all the 460 domains and to do that we use a Boilerplate or content extraction method which is called Diffbot there are others out there other tools that extract content from web pages such as Dragnet or Skin tech we use Diffbot

because it's one of the most popular ones. Now that we have the content from web pages the content from the articles that we are interested in the next thing that we have to do is turn that into text and turn that text into numbers that can be used for quantitative analysis.

So how do we do that? We use standard text cleaning and natural language processing methods to do that such as turning the entire text into lowercase removing punctuation removing stop words stemming and turning the text into tokens into words. We also use part of speech tagging and engrams which are combinations of words. Now with this process text you can do a number of things you can for example count key words in the text keywords that are of interest to you. You can measure distances or similarities let's say you care about the similarity between the articles that are shared by The Telegraph and the articles that are shared by The Guardian and you want to see whether they are more similar to each other versus the articles that are shared that are the Daily Mail webpage.
Another method that you can apply to the text is topic extraction. The most popular method for topic extraction is called Latent Dirichlet Allocation or LDA. This is the one that is most often used by researchers. What this method does is uncover hidden thematic structures in your documents the assumptions it makes is that first documents are a mixture of topics the topics generate words based on their probability distribution and what the algorithm does is determine the number of words in a document it then determines the mixture of documents in the topics in the document and then based on the topics multinomial distribution it assigns words to documents.
Now there are multiple ways of performing topic analysis and specifically LDA on your text files you can use Mallet which is a self-standing tool you can also use Python Gensin which is a Python library for doing that or R Quanteda which is an R Quanteda for extracting topics but also for processing text more generally. We use Python Gensin in our data since most of our other work is in Python as well. So what you get out of LDA is a number of topics depending on how many you asked at the beginning along with a list of words and their corresponding probability of being assigned to a specific topic. So for example we have a topic number one in the list that has the words young university people women student what you don't get out of LDA is the label for that topic so what you have to do is read through the list of words and try to figure out what is this topic about and then put the labels in.

Another example for the second the third topic in our list immigration the words are EU immigration UK Ireland migration people by reading the words and the probabilities that they are assigned to the specific topic you can figure out that the topic is about immigration. What we do then is we compute topic probabilities for each document and then for each topic we average those probabilities across documents across the articles that we have in our data set the articles that are extracted from the clickstream and the twitter data so we end up with two numbers an average probability of topic let's say the economy in the clickstream data and an average probability of the topic of the topic the economy in the Twitter data. And these are the numbers that we are working with in the next part of the analysis.

Our analysis has another part that looks at networks so we look at the content behind the articles but we also care about the networks. And why would we care about networks why do we want to model our data like that that's because we believe there are inter dependencies in the data and we want to be able to explicitly model those inter dependencies between observations. We are interested in a patterns of information exposure in the data we also care about detecting communities or echo chambers in the data so what is a network. A network is a mathematical construct of nodes and edges information can be stored on the nodes as well as the edges in our case the nodes are the domains for example the node is one node is the

BBC and another node could be the Guardian and The Telegraph so these are news domains are the nodes in our network. What are the edges well the edges are formed when a user either shares on Twitter or reads in the clickstream data articles on both domains if they read something on the BBC website if they read something on the Telegraph they may draw an edge between those two nodes. Now we have these three types of data the survey data a clickstream data and the Twitter data how we link those three types of data. We want to link it because we believe that there is more information in the link data than in analysing these three sources separately so first we link the server click stream data now in the social sciences most of the time when we link data we do it at the individual level so we link data across multiple sources on the same individuals. However if we want to link the click stream and social media data we can also link it at the domain level or at the article level and that's because that's the level that we are interested in when we are studying information exposure.

We care about the information that people receive from those domains or from those articles so we also link the click stream and the social media data at the domain level in the article level and finally we link across the survey click stream and Twitter data again both at the domain level but also at the article level. And in our analysis we are now going to talk about linking data at the domain level. So the different sources of data have been extracted they've been processed and they can be linked across various levels across the individual level and across the news domain level and the story level.

Now we can move on to the analysis and the results. So we will be showing results from the surveys linked to the URLs from topics in the news both on Twitter and from the clickstream data and we'll be looking at the networks of sharing of stories across Twitter and across the clickstream data.

So we'll first start with comparing users and news consumption from the surveys and the clickstream data where the URLs have been extracted. What we're able to see in the data once it once it's extracted and clean is we have a line of data that's represented by a respondent by the URL that they clicked on by the time they clicked on it and then that is linked to the survey data and their responses. From the clickstream data that's linked to the survey responses we can create various quantities of interests we can collect and aggregate the total news and information URLs that they've clicked on across the weeks of data that we've collected we can analyze the new story titles if we wish to we can also designate the news and information URLs that they visited as either left leaning or right leaning URLs and then we can proceed on with our analysis.

One of the first steps in the analysis that we've done is to look at the most visited URLs in the clickstream data and it turns out that the BBC is the most visited site in the news and information URL now if you look on the Twitter data it also turns out that the BBC is the most shared news domain or the news stories are most likely to come from the BBC that are shared on Twitter data. So that is a consistency across both the Twitter and the clickstream data.

One of the next things that we can do is to take our survey responses and compare what people report about their news consumption online to what they're actually doing online so that's a matter of linking at the user level and what we did in the survey was ask how often in the past week did they go online to read about the EU referendum and we had four categories of responses they could go on every day most days some a couple of days or not at all. When we compare that to where they actually went online and how many times they clicked on stories about the EU it's fairly consistent those who went online every day to read about the EU referendum were more likely to click on stories about the EU so again that's a consistency that we find in our data. We can also as I said create and classify the URLs into left leaning or right leaning URLs or URLs that are more likely to be consistent with the Conservatives so

something like The Telegraph online and those URLs that are more consistent with or more likely to be labour leaning something like The Guardian so if we classify all the news and information URLs like that we can also see how those who expressed a preference for remain or for leave or didn't know what their online viewing habits were like and what we find is that amongst those who are remainders or expressed a preference for remain their news viewing online was balanced or more balanced across conservative labour leaning as well as broadcast outlets. However when we look at those who expressed a preference for leave we see that they're more likely to be clicking on stories from conservative-leaning URLs and less likely to be clicking on the BBC. Amongst those who are undecided we see that like those who expressed preference for leave were more likely to click on conservative-leaning newspapers. So again that shows us the sort of different news viewing habits based on their preference for the EU referendum and it does show differences in terms of their news consumption behaviour online using the clickstream data.
So if we move next on to comparing the topics from the news stories that we extracted from the click stream and from the Twitter data this allows us to look more in depth at the actual stories that people were sharing on Twitter and clicking on Twitter and then going to in their web browsing activity. So we have the Twitter stories and we have the stories that we've extracted from the clickstream data. When we look at those what we are trying to do is to build topics out of those news stories and from all of those stories we are trying to extract different topics and we extracted up to 50 topics from the corpus of text and we are able then to compare these topics across the Twitter news stories and the clickstream news stories. One final consideration after viewing the results is to ask ourselves how certain we are that we can make determinations and draw inferences about online news consumption from the data that we are examining. And in order to do so we still have to address considerations about the representativeness of the sample that we're using whether or not we're making appropriate measures with the data that we're using and whether or not we can establish that what people are seeing online is causing them to have particular attitudes or to behave in particular ways.

So even though we're using these new sources of data to look at old questions about information exposure we still have to ask ourselves these fundamental questions that are central to any sort of social science research. What we've shown today is that by using multiple sources of data we hope to provide a possible solution to looking at these fundamental questions of representativeness and error and causal inference.

If we're interested in further exploring the data and the tools that we have used today we have two sources of information for you one is a website where we have further readings and links to data sources and then we also have a set of Jupiter notebooks that are available in our GitHub repository.