The crowd in the cloud: three challenges for measuring public opinion online

Josh Cowls, Oxford Internet Institute

Introduction

‘Public opinion’, ‘national mood’, ‘vox populi’: such phrases and many more are used in societies around the world to refer to the aggregated opinions and attitudes of the individual members of a society towards political ideas, policies and actors. In democratic systems, these attitudes and opinions solidify into votes for particular parties and politicians at an election, but at other times and even in less democratic places, the idea that there are opinions held by the public occupy the thoughts of every political leader. Courting public opinion can be the difference between election and eviction, or between the survival of a regime and revolution.

Thus political leaders, as well as political scientists, have long held an interest in understanding more about public opinion. The way in which public opinion has been conceptualised and – more recently – rigorously measured, has changed over time. In particular, as the first section of this paper argues, these changes in conceptualisation have frequently coincided with the evolving availability of different information and communication technologies (ICTs), which mediate the vast space in which most modern nation states are constituted.

The first section of this paper, charting evolving perspectives on public opinion, lays the groundwork for the analysis which follows. The subsequent section introduces the Internet as the latest ICT with the potential to alter how public opinion is conceptualised and measured, in a similar way to other earlier technologies. From there, the paper turns to empirically investigate the extent to which new sources of public opinion data, collected from the Internet, can result in valid findings about public opinion. The paper uses insights from interviews conducted with researchers at the forefront of this area of research, to look in depth at three challenges to validity – in relation to the reliability of the data; the representativeness of those ‘sampled’; and the replicability of this form of research. The paper concludes with further thoughts about the prospects of online public opinion research, as a form of ‘big data’, making valid contributions to our understanding of public opinion in the future.

Evolving perspectives on public opinion

Offering a secure, uncontroversial definition of public opinion is a challenge that has been undertaken many times, and no fresh attempt is made here. The very diversity of definitions – Childs (1965) identified fifty such efforts fifty years ago, and there have been many more since – gives this author pause. Instead, it is the tensions that have arisen between different conceptualisations of public opinion which are especially relevant to this paper, and in particular, the changes which can be observed over time. For the core contention of this section is that the notion of public opinion has been buffeted throughout history by the winds of technological advance. As the range of information and communication technologies (ICTs) in society has expanded – particularly, but not exclusively, in the twentieth century – both the ability of members of the public to express their opinions, and the tools with which to measure this opinion, have advanced.

To see how this phenomenon has developed over time, for a suitable baseline we might look to classical antiquity. The idea of direct democracy often associated with ancient Greece is of course a flawed one: the franchise extended only to free men, excluding women and slaves. Yet an idealised version of the ancient Greek city-state offers a lens through which the later development of public opinion can be viewed. For our purposes, the crucial aspect of the polity is its size: Plato argued for the size of a polity to be fixed at 5040 citizens (Plato, 1970), while Aristotle distinguished between a city, manageable in (literal) view of government and magistrates, and the unwieldy multitude of a nation (Aristotle, 1997). Central to this conception of the small, self-contained city-state was the right for every citizen not just to speak – as is common in modern liberal democracies – but to *be heard* by all others in attendance. To be sure, this right was more evident in principle than in practice, yet, as John Durham Peters notes, "a political assembly could never be experienced as a spectacle that passed indifferently before the audience's eyes, since every Athenian citizen had the theoretical option of contributing to the public discourse" (1995, p.15). The ancient Greek assembly was thus an inherently participatory affair.

The characteristics of the ancient Greek city-state – its small size, the physical co-presence of its inhabitants, and their right to speak and be heard – offer a useful point of comparison for the shifting notion of public opinion over time. The first and foremost challenge to this ancient conception of public opinion was the increasing geographic size and demographic diversity of the state. Platonic and Aristotelian notions of small, self-contained communities gave way to larger states; this precluded the physical assembly of populations (regardless of whether or not such assemblies would have been empowered to make decisions). Thus as the size of political units exceeded the practical limits of physical co-presence, what was missing were technologies that allowed these distances to be overcome.

This changed with the development of Gutenberg's printing press and the subsequent emergence of newspapers. Nineteenth century political thinkers Alexis de Tocqueville and John Stuart Mill both argued that the newspaper represented a means for public opinion to be constituted amongst physically dispersed populations (Tocqueville, 2004; Mill, 1946). But far from being a perfect substitute for physical co-presence, the mediating role that the newspaper played necessitated a different conceptualisation of public opinion. In this sense newspapers are central to Benedict Anderson's notion of 'imagined communities' – how the development of a national consciousness can arise in large modern states. Anderson describes the reading of newspapers as an "extraordinary mass ceremony ... performed in silent privacy [yet] replicated simultaneously by thousands of others in whose existence [one] is confident" (Anderson, 2006). While the privacy in which newspapers are read stands in stark contrast to the public spectacle of the ancient Greek assembly, the simultaneity of this ceremony can be seen as emulating it.

Yet this "ceremony" is by no means perfectly equivalent to the more direct, instrumental public opinion of the ancient Greek city-state. Gabriel Tarde, anticipating Katz and Lazarfeld's 'two-step flow' theory of communication (Katz & Lazarsfeld, 1970), saw public opinion as arising in the first instance from the "conversation" between individual and newspaper, before being reinforced in the real-world coffee house or salon (Tarde, 1901). Peters similarly reflects that newspapers, as with other 'one-to-many' ICTs such as the television and radio which followed in the twentieth century, are more suited to "mak[ing] something visible before the people through the media rather than involving them as active participants", such that "it is far easier with current technical and institutional arrangements to constitute a society of spectators than actors" (Peters, 1995, pp.15/27). Thus, the press, television and radio have affected how public opinion is both created and conceptualised in society.

Another ICT, the telephone, has also exerted an influence on the conceptualisation of public opinion. Opinion polling emerged in the early part of the twentieth century as an attempt to quantify public opinion in a population at large through the confluence of the telephone as a convenient means of conducting opinion surveys, and inferential statistical methods which allow responses to be generalised to the population as a whole. This technique transformed understanding of public opinion once again, further shifting attention away from social discussion and dialogue towards discrete, isolated clusters of opinion in the population. This shift has created new tensions, leading some to question whether the modern concept of public opinion is really 'public' or 'opinion' at all. Critics have noted the contradiction between the claim of public opinion and the private context in which they are conducted and the fact that, through interviews and surveys, public opinion is generated rather than observed (Peters, 1995). In addition, the sociologist Pierre Bourdieu in his (1979) article 'Public Opinion does not exist' argues that three assumptions of contemporary opinion polling – that everyone has an opinion, that everyone's opinion is of the same value, and that everyone agrees on the question – are flawed.

This section has demonstrated that the notion of public opinion has shifted over the course of history, and particularly in response to the emergence of various ICTs. The remainder of the paper examines how the large-scale adoption of a more recent ICT – the Internet – is affecting the conceptualisation of public opinion once more.

The Internet and social media: another shift in understanding?

The emergence of the Internet and the wide array of popular information and communication platforms that it supports – including search engines, news services and particularly social media sites – raises the prospect of new, rich sources of data on public opinion. Yet the nature of the data itself, and the tools and techniques used to access, capture and analyse it, have the potential to transform how public opinion is conceptualised once again – just as the press, television, radio and telephone affected earlier understanding.

Because of the huge range of information and communication activities possible on the Internet, public opinion research using digital research and methods can take myriad forms. Some of this research closely resembles traditional methods, such as online surveys and interviews conducted by polling companies like ICM and YouGov, where – just as with telephone surveys – a weighted representative sample of respondents is constructed and results reported with a standard error. No doubt such an approach offers new opportunities for traditional opinion polling, particularly at a time when response rates for telephone surveys are in decline (Pew Research, 2012).

Yet discussion of the potential of the Internet as a tool to supplement existing sample-survey research is outside the scope of this paper. Instead, this paper focusses on sources of data, and techniques for utilising them, that are fundamentally at odds with the existing sample-survey approach – such as data emerging from social media sites and other transactional activity online. These are often characterised as being sources of “big data”: that is, of an unprecedented scale and scope (Taylor, Schroeder, & Meyer, 2014) or volume, velocity and variety (Beyer & Laney, 2012) in relation to the analysis of a particular phenomenon. In light of the previous section, it is easy to see how the streams of data emerging from the Internet are unprecedented in the analysis of public opinion: the corpora of social media content now available to researchers are orders of magnitude more voluminous than in traditional surveys; data are recorded at a previously unimaginable velocity; and there is an immensely wide variety of different types of content available for analysis. Yet amongst all the hype surrounding the volume, velocity and variety of big data, the question of validity is sometimes overlooked (boyd & Crawford, 2011). This paper thus presents an overview of the various challenges to validity raised by the use of big data sources to derive public opinion.

Data

The interviews used in this paper were conducted as part of a larger project investigating the emergence of big data research in the social sciences, which has so far held several workshops, conducted 125 interviews, and produced a number of studies of various issues surrounding this new type of research (Schroeder & Cowls, 2014; Taylor et al., 2014). Interviews were conducted with practitioners at the leading edge of big data research. Interview participants were sampled purposively rather than systematically: given the nascent nature of big data approaches, interviewing pioneers or ‘early adopters’ within this exploratory framework was best suited to the aims of the project. Nonetheless, in scoping out prospective participants, diverse set of perspectives and experiences were deliberately sought. Most interviewees belong to an academic institution, whilst a minority of interviewees belong to commercial, governmental or other not-for-profit groups. A semi-structured interview approach was utilised, posing both technical questions (in regards to data sources, tools, methods and so on) and pursuing more philosophical, reflexive issues explicitly. The responses given – which were offered in response to questions about the use of large-scale data in general rather than public opinion data specifically – are embedded in the analysis and commentary that follow.

Big Data for Public Opinion: three challenges to validity

As previously argued, the capture of big data to measure public opinion represents a significant shift from the traditional mode of survey research. Surveys are conducted actively: generally, respondents are identified and questions specified in advance. They are also conducted explicitly: respondents are usually fully informed as to the nature and purpose of the research prior to participating. This approach is largely "an artefact of a period of information scarcity" (Mayer-Schönberger & Cukier, 2013, p.12), a result of the finite resources researchers can draw upon: surveys are expensive and time-consuming, and these costs increase in proportion to the number of respondents questioned. Yet this is also a consequence of the analytical approach taken to surveys. Statistically speaking, beyond a certain point the inclusion of additional observations will not increase the accuracy of estimating the statistical significance of a phenomenon. Thus for a combination of practical and theoretical reasons, survey methods typically involve between one and two thousand respondents at most.

With big data research, the situation changes substantially. On a practical level, recent technological developments in the storage and analysis of huge amounts of data mean that the costs of handling additional observations in a data set are negligible compared with the proportional increase in resources required for every new survey respondent. But big data moves the goalposts on a theoretical level too. Statistical techniques designed "to confirm the richest finding using the smallest amount of data" (Mayer-Schönberger & Cukier, 2013, p.20), are less useful in an environment in which both data itself and our technical capacity to deal with it have increased markedly.

This shift to a data-abundant environment raises a number of novel challenges. Many of these centre around the quality of the data collected and the efficacy of techniques applied to it. Indeed, "messiness" is a central characteristic noted for big data (Mayer-Schönberger & Cukier, 2013). It is therefore important to consider more closely the challenges involved with establishing validity in the use of this data, which affect the utility of the findings that can be drawn from this new area of research. These challenges can be grouped arranged under three headings: reliability, representativeness, and replicability. Drawing on the interview data described above, each of these is addressed in turn.

Reliability

Issues around reliability are central to the process of gauging public opinion through online sources. In contrast to survey-based methods, where opinion is collected actively and explicitly, data online is typically collected passively and tacitly. While this offers some potentially very insightful data, it also poses new challenges to ensuring the reliability of data collected. This sub-section looks at two of these challenges, relating to sentiment and identity, in detail.

First, collecting public opinion online raises new questions around sentiment, and the question of what people really 'mean' by what they say online. Online communication takes myriad forms, and is presumably affected by the technical complexities of a particular platform and the social and political context of a speaker. For example, Facebook and Twitter – two of the world's most popular social media sites – differ in various ways: Twitter limits tweets sent to 140 characters, but these messages are public by default; Facebook has no limits on post length but posts are by default visible only to a user's friends. These seemingly small technical and social differences raise larger questions about the extent to which online expressions reflect genuine sentiment on the part of the speaker. Alex Pentland, a professor at MIT and chair of the World Economic Forum's Data Driven Development Council, reflected on these limits:

> **Alex Pentland**: I tend not to look at things like Twitter and Facebook because those are your public persona, they're the things that you express in this very culturally constrained socially intentional way; this is the thing you say to make your friends respond or to appear good … It depends also on all the cultural context and your individual context.

Brandon Stewart, a PhD student at the Department of Government at Harvard, offered similar reflections on the ambiguities of meaning inherent to social media data:

> **Brandon Stewart:** social media is really, really fascinating, and the reason is because it is clearly, it falls into this category of there's something there but we don't know what it is. So you can measure public opinion on Twitter and clearly that's indicative of something, but we don't quite know what it's representative of. In fact we're fairly sure it's not representative of much of anything.

Mike Thelwall, Professor of Information Science at the University of Wolverhampton, who develops sentiment analysis tools, reflected on this difficulty:

> **Mike Thelwall**: really the big problem that we haven't cracked is that if someone tweets a sentiment it's not necessarily what they're feeling, it can be for a variety of reasons, so it doesn't really reflect directly what they feel necessarily. But using Twitter is a particular communication strategy, so ... it's a stretch to say that if someone tweets, "I'm happy" that they're actually happy, to give a simple example.

Researchers recommended various methodological techniques for coming to grips with such challenges. Danie Stockmann, an assistant professor at the University of Leiden, discussed the fact that dissent on Chinese social media is usually expressed obliquely; the Tiananmen Square protests of June 4th, 1989, were long referred to as 35th May in an effort to evade censorship (Nordin & Richaud, 2014).

> **Danie Stockmann**: [this] is a problem that is I think more that you run into more frequently online ... my approach usually is to start with qualitative research and really learn the language and that the authors of those sources use. And then once you, once you have a good understanding of the language which you learned from talking to them but also reading the texts obviously, and talking to them about the meaning of the texts then you come up with a, with a, or in my case I developed dictionaries to measure certain concepts.

Similarly, Mike Thelwall recommended qualitative methods to overcome issues of reliability:

> **Mike Thelwall**: I think it's a problem that would need a lot of qualitative research to investigate, so interviews, questionnaires, maybe even experiments to find out what the real relationship is between what people tweet and how they actually feel. I think it's a big complex issue that is not a computational issue, I think.

Thus, accurately gauging sentiment expressed online – especially in sensitive political contexts – requires qualitative checks on the data collected.

The second question surrounding reliability concerns not the content of speech but the identity of speakers. Again, this issue varies based on the platform – on some websites, users are encouraged or required to provide accurate information about themselves, while on others anonymity is taken for granted. Thus on more open platforms in particular, the potential for automated messages to influence the flow of public discourse should not be underestimated (Ratkiewicz & Conover, 2011). Nick Anstead, a lecturer in media and communications at the London School of Economics, compared spam messages to genuine political engagement online:

> **Nick Anstead**: things like spam, and astroturf and bots [are] arguably more problematic ... I mean, we might say it's a sign of financial resources for example, but, you know, it's not maybe a sign of real, or virtual-less social organisation. You know, generally, we might assume that political mobilisation [online] is a good thing, SPAM bots are a bad thing. But that's a normative judgement.

Anstead emphasises the importance of judgements made to deal with identity when designing research. Filip Noubel, director of the Internews China program, discussed how governments are able to tackle this issue directly through policy:

> **Filip Noubel**: the Chinese government has seen that how popular social media are, you know, we have big, big platforms with as I said several hundred million users combining also we have to remember not just text but also images, video and now sound on Weixin. It makes it quite difficult and the volume produced every minute in China is astonishing. So the Chinese government came up with this idea based on what they thought was a good example in South Korea which is called you know, the real vision and policy where anyone who goes online can use a pseudonym, but the first time you register on any platform you actually have to give your real name and some form of ID number, right, so that in case there's a problem the police or whoever is in charge can actually track that person.

Absent policy-making power, for academic researchers the question of identity is one which must be considered at the point of analysis and interpretation. Sometimes, however, total anonymity can actually be beneficial to research. Seth Stephens-Davidowitz, a data scientist at Google who recently obtained a PhD in Economics from Harvard, uses search query data to investigate socially sensitive issues and attitudes (Stephens-Davidowitz, 2012, 2013). He explained how the anonymity of the data is what allows such insight in this mode of analysis, in contrast to conventional methods:

> **Seth Stephens-Davidowitz**: Anything of a sensitive nature, whether it's racism, child abuse, sexual behaviours of various sorts, drugs, all these topics, the traditional data sources are very, very limited ... you can't just ask a Gallop survey, 'Are you racists?' for a thousand people on the phone ... And this is an area where I surmised and I think I'm right on this that Google would be really good at – people [are] for the most part alone online, you need information, you have an incentive to tell the truth, if you get off on racist jokes and you want to see them then you've got to type it into Google so you really have a clear incentive to tell the truth and to tell what you really want. And it's anonymous and everything; all the data is anonymous data, so people are very honest and open.

Stephens-Davidowitz's work thus involves an innovative use of data from an online platform to elicit attitudes in society at large, with more precision than traditional survey methods. It is important to note, however, that Stephens-Davidowitz's success is dependent on these opinions being expressed in private. While revealing, in no way do they form deliberate expressions of opinion in a public forum. As noted, the size and makeup of a user's audience on a particular platform is an important factor in what they say, but Google is used with the assumption that *no one* is listening, and it would thus be inaccurate to describe this as public opinion. However, it is worth restating the fact that most conventional research into public opinion is also conducted in private settings, via the telephone or in-person. Thus, while Stephens-Davidowitz's work on Google data may push too far in the direction of a private context, traditional polling research is arguably closer to this form of data collection than the public expressions of opinion collected online.

In this section, the reliability of sentiment expressed online and the identity of those expressing it have both been questioned. The recommendations offered by researchers for combatting these challenges centre on establishing a more sensitive understanding of the context in which public opinion data is generated and collected, including the use of qualitative methods, domain-specific knowledge, and reflexivity on the part of researchers. The following section turns to look at a second cluster of challenges related to deriving public opinion from the Internet.

Representativeness

The traditional model of opinion polling, as described earlier in this paper, rests on the assumption that the sample of respondents are representative of the population about which claims are being made, in regard to a number of variables which usually include age, race and gender. When public opinion data is collected on the Internet passively and tacitly, this situation changes drastically. This section identifies various ways in which data collected online threatens this notion of representativeness.

First, Internet penetration has not yet reached 100% in any country (World Bank, n.d.), and the persistence of a 'digital divide', based on socioeconomic factors (Dutton and Blank, 2013), would suggest a systematic skew amongst Internet users towards the better-off. The situation is murkier still at the level of individual social networks, membership of which can be more uneven still (Pew Research, n.d.). Amber Boydstun, Assistant Professor of Political Science at the University of California, described these limitations in terms of Twitter:

> **Amber Boydstun**: we can get very stodgy about the types of data that we use and, for example, anyone who does a Twitter study has to really work hard, I've noticed, to justify why we should care about Twitter because Twitter is not representative of the United States population or the world population, right. And it's not.

Bente Kalsnes, a PhD student at the University of Oslo, similarly reflected on how the relative membership of social network sites affects her choices of which networks to study:

> **Bente Kalsnes**: In Norway, Twitter is used – 20% of Norwegians have a Twitter profile. I don't have the data of how often they use it but, basically, Twitter is fairly small in the population. When you look at Facebook, I think something like 75% of the population has a Facebook profile, so it's huge. So, in that sense, it's much more interesting to look at Facebook.

Scott Hale, a PhD student and research assistant at the Oxford Internet Institute, weighed these questions of validity against the extent to which data from different platforms could be accessed by researchers:

> **Scott Hale**: one big challenge for researchers doing these sorts of studies is often you choose one platform, because the data is there, you have access to it, you have the connections, whatever it is – in that one platform, you may find something very interesting, but it's not clear whether that generalises to the broader set of social media platforms that are similar.

Yet even if the *membership* of a particular social network or other online platform were representative of a population at large, the distribution of *activity* amongst users might not be – and it is activity in the form of expression which public opinion researchers look to measure. Jamie Bartlett, director of the Centre for the Analysis of Social Media at the think tank Demos, reflected on this:

> **Jamie Bartlett**: We don't think Twitter is particularly good at delivering on [representativeness]. You could do all the analysis of biography data to try to create a representative sample, and it will only ever get you so far because you always have a self-selection bias in who is producing data.

This thorny issue of selection bias extends to research conducted on other online platforms. Alex Leavitt, a PhD student at the Annenberg School for Communication & Journalism at the University of Southern California, discussed how the analysis of what he calls 'trace data' on Reddit precludes analysis of other types of use:

> **Alex Leavitt**: So, for example, one thing that Reddit allows is that I can easily do any sort of study by working from the trace data of people who participate on the site. Except that in reality many people on Reddit only use the site by viewing what other people have posted. To really contextualize how Reddit is used, I will never be able to study them by only looking at trace data. So I am basically missing about 50% of what Reddit is.

Even accounting for differences in Internet access in general, relative usage of various social networks sites, and the different levels at which and ways in which users contribute to these sites – all factors which can reduce the representativeness of the data collected – additional biases can be introduced by researchers themselves. The most obvious quality of social media data is that there is a lot of it: billions of pieces of content are created every day on social network sites around the world. In most cases (though not all: c.f. Hale, Yasseri, & Cowls, 2014) researchers investigating a particular phenomenon need to look at only the relevant parts of a larger dataset. This is especially crucial in the case of public opinion, given that large swathes of conversation on social media have little or nothing to do with political issues. Thus, researchers must find ways to filter large datasets to create a more relevant corpus of content to analyse.

Different platforms offer different facilities for filtering. A particularly salient example is that of hashtags on Twitter, which are keywords added to tweets by the speaker to enhance organisation and categorisation, as the inbuilt hashtag facility allows other users to see all tweets using a particular hashtag in real time. This has become particularly useful as a means for organising protest (González-Bailón &

Wang, 2012, Dubois, 2013) and wider discussion among and between communities of interest (Bruns & Burgess, 2011, Conover, Ratkiewicz, & Francisco, 2011). Just as hashtags have proved helpful to citizens organising collective action and debate, they have also been utilised by researchers. Yet this has raised questions about what hashtags signify, and whether the use of them as a filtering device creates systematic inaccuracies in datasets. Carl Miller, Research Director at the Centre for the Analysis of Social Media, discussed the challenges of sampling Twitter conversation around a televised debate:

> **Carl Miller**: Normally when you sample, you basically give the Twitter API a series of keywords it goes for. Things like [live, televised] debates are great because they have hashtags nowadays, so those 62,000 tweets were at least nominally all relevant to the question ... but [it] was a BBC-created hashtag that we put out just for that debate. No one was using it ten minutes before the debate. Actually, they announced it about three minutes in which caused us to start scurrying around manically trying to get the collection up and going.

Practical challenges notwithstanding, a hashtag created and promoted by a national broadcaster for use during a debate could result in a fairly comprehensive and representative collection of tweets. Yet the assumption that all hashtags are a representative repository of conversation is more dangerous. Axel Bruns, Professor at the Queensland University of Technology, reflected on the uncertainties of using a single hashtag for analysis of wider political conversation:

> **Axel Bruns**: we've done some analysis of a hashtag called 'auspol', which is ongoing discussion of politics in Australia, the lead user in that hashtag, the most active user, over the course of less than a year tweeted 30,000 times … that's more of the problems with the hashtags stuff, we have wonderful case studies but we don't know what they sit in essentially, what the framework is, if that's 1% or 10% or 100% of the current conversation in Australia or whatever.

Filtering in this way could be a much less robust approach when the hashtag in question is created for ostensibly partisan purposes. Nick Anstead offered such an example:

> **Nick Anstead**: During the last election campaign, the Labour Party launched a Tweeting campaign called #WeLoveTheNHS, yeah. When we spoke, sort of, opinion pollsters, they say this as a real problem, 'cause they saw it as a distortion.
> Interviewer: The Labour Party's use of the hashtag?
> Anstead: Yeah, right, because basically it was being organised. It was activists doing it.

Anstead went on to discuss the wider political issues involved with the use and misuse of search terms to filter social media datasets:

> **Nick Anstead**: we tend to search for key terms, yeah. We tend to say, this is the issue we're interested in, but there's a problem here, right, because you're already saying if people don't talk about, or think about this issue in the way we talk about it, or think about it, they won't appear in our filter ... Okay, classic situation right, you know, when we are looking at politics do we look at references to the three party leaders, or do we look at, you know, the kind of traditional issues that would have been raised by feminists for example. And [do] our search terms, and our search terminology pick up on that? It might not.

The evidence presented thus far has suggested myriad threats to the validity of public opinion research conducted online, in view of the lack of representativeness of the Internet itself, particular online platforms, and the filters created by researchers. Yet the interviewees also offered thoughtful responses to the idea that a lack of representativeness impedes valid research in this area. Jonathan Bright, a research fellow at the Oxford Internet Institute, pushed back against the notion that online platforms were not representative enough.

> **Jonathan Bright:** I think you can make a plausible case, by saying look they are not a random sample of the population, but no one ever takes a random sample of the population, even in opinion polling. But these people are sufficiently representative of how people get and pass on information, that these conclusions can be generalised beyond Twitter and Facebook. And I think that is, for me, that is valid, but people are usually pretty careful about this, when you want to say something like this is the opinion of the nation … [with] Google, you can make a really plausible case, this is the way over 50% of the population try and find out information about something … I mean Google has this astonishing kind of coverage.

Nick Anstead offered a historical perspective, suggesting the incorporation of earlier notions of public opinion into modern understanding.

> **Nick Anstead**: one alternative solution is to actually think about the, sort of, theoretical paradigm you're operating within. So, you know, not to accept the Big Data 'all we are about is the correlation' approach – but also not to get, sort of, straight-jacketed into the 'public opinion is only ever generated by representative sample of opinion poll' model. So, you know, can we start to think of public opinion, for example, as being conversational? Can we start to think of it as being more dynamic? And what you actually find is, if you want to think about theory on this, is if you go back further in time. If you go back to the 19th century, you find public opinion [in] coffee houses – what James Bryce calls 'organs of public opinion' in his book *The American Commonwealth.* Maybe we should think of social media as an organ of public opinion. Now at this point, some of the methodological problems don't exactly melt away, they become at least things we can reconcile, and we can think about this as being a useful tool.

Anstead here proposes a significant reorientation in how public opinion is conceptualised, drawing attention to the historical antecedents discussed earlier in this paper. Jamie Bartlett similarly suggested a shift in our understanding of public opinion to accommodate these new streams of digital data:

> **Jamie Bartlett**: you've got a pretty big problem in trying to convince someone this is as good as a poll. But when you get them away from that and talk about is being a particular group of people … a response that a certain number of people have to an event – and that can be a very significant and important one – then I think people will begin to see the value in it, because they see that it's quick, it's often actually very detailed, very large volume, very rich, very important constituency of people that are using it often at times, and gives you completely new ways of understanding groups that traditional polls don't. It is a challenge because polls are what people are used to and that's what people consider the gold standard because they are talking about a sample that represents a bigger population and people always seem to want to know what the whole population means. I think it has been a struggle, not a struggle, but part of the challenge for us has been to find new ways of talking about this and describing this in a way that people will understand.

Filip Noubel, speaking in terms of the Internet in China, volunteered the concept of 'crowd-sourcing' to signify the more dynamic understanding of public opinion similar to that which Anstead and Bartlett advanced:

> **Filip Noubel**: I would say that what is probably, probably unique to China because of the size of the population and also to what extent social media is vibrant here is that we call this more maybe sort of, you know, national crowd sourcing or sort of you know, crowd sourced national enquiries and we have a lot of sort of big news related to natural disasters because China's into earthquakes. We had maybe remember from the media a train accident, a high speed train accident last year which initially the local government was not so sure on how to cover it and try to also not release the data of the number of victims but then literally minutes and then hours, the whole country knew what had happened and the collection of the data and the transmission of the data was made actually mostly by citizens.

In responding to arguments about the lack of representativeness of public opinion expressed online, interviewees offered important philosophical considerations, highlighting how the historical antecedents of contemporary public opinion research may offer ways of understanding the concept more appropriate to the data collected. These responses serve to defend the use of digital sources of data in modern public opinion research. The next section turns to the third and final challenge to validity.

Replicability

The last challenge to online public opinion research which emerged from the interview data is that of replicability. As has been noted, a standard set of measures, such as sample size, a margin of error and demographic weighting, have become a staple part of any opinion poll. For the reasons already discussed in regard to the reliability and representativeness of data, such metrics do not cohere with opinion data collected tacitly passively from online platforms. Instead, the extent to which this research can be scrutinised and replicated is limited in numerous ways. First, where there is restricted access to data, other researchers are unable to replicate research – a fundamental requirement for empirical research, which Brandon Stewart reflected upon:

> **Brandon Stewart**: I'm always very sceptical of data sources that are essentially where, you know, things fundamentally can't be verified by external researchers, and that isn't a matter of thinking that there's anything nefarious on the part of the original authors as much as it is that there's almost nothing that couldn't be usefully expanded or extended or developed by other people because nobody can do everything.

In particular, multiple interviewees drew attention to a 'black box' situation in which data is collected *for*, rather than necessarily *by*, researchers. Nick Anstead described the validity issues which arise when academics obtain data from third party companies:

> **Nick Anstead**: what do you actually get from working with these companies? Do you get raw data sets that you go and do stuff with yourself? More commonly, I would suggest, what you probably get is access to, sort of, a black box tool. … You know, you put in certain search terms and it churns certain data out, and that's sort of problematic as well, because then the question: where does that data come from and how was it processed, and how was it created?

Axel Bruns described the limitations that researchers face in publishing raw data alongside findings, even in the case where data was obtained directly from the platform:

> **Axel Bruns**: So there's a real problem with that, sort of, black box in the middle, we have a wonderful description of what we're going to find and the theory and how we gathered the data, and we have a bit of description of what we did with the data, and then there's the findings and, oh look, they tell us something that we've always suspected was the case, or whatever. But there's a disconnect there unfortunately and it's very difficult to get around that, unless we can publish the data alongside with, alongside the article, which we can't obviously under Twitter's rules to begin with.

Sandra Gonzalez-Bailon, an assistant professor at the Annenberg School for Communication, echoed these concerns:

> **Sandra Gonzalez-Bailon**: in research in general, you always have to take account the data is not perfect and that if you know that there are some potential sources of noise you have to control for … Now the difference with big data is that often, because of the way it is collected, the biases are outside of your control. And with the Twitter APIs that's particularly the case because have black boxes … You get a sense of how they operate, so you know that the API is the main communication channel with the Twitter servers where the full stream of information is and that APIs give you access to that, but you don't know exactly what kind of access you get.

Bente Kalsnes has investigated this phenomenon further herself, by comparing data downloaded with the use of a bespoke tool with that obtained from a commercial organisation:

> **Bente Kalsnes**: I wanted to look at the data we have, and I compared it with a commercial tool, in order to get a sense do we get 80% of the population, do we get 50%, do we get 100%? We don't get – when I was comparing, I never got exact numbers between the two tools and sometimes I got more data on [one tool] than on the other tool, and other times the opposite. So there's no clear pattern. But, basically, I have a better understanding of the access to the Twitter data we get … so I know it's not completely off. And that was also important for me since Norway has a much smaller Twitter population compared to the UK or Australia or the US. So if you lose 3000 Tweets on a hashtag, that would be crucial. It was important for me to know that this is fairly good.

Other researchers have also looked into this question, examining the variation between data captured in different ways. Undoubtedly, such efforts help to establish more knowledge about the scope of data collected, improving researchers' confidence in the data upon which conclusions about public opinion are drawn. Yet even if data were accessible or entirely comprehensible, the shortage of skills to handle this data limits the replicability of this sort of research, as Jonathan Bright argued:

> **Jonathan Bright**: I think the problem is the skills gap. … the place that hired me first was the Oxford Internet Institute, which is a particular institution that doesn't focus on political science. I don't think lots of political science departments are really keyed into teaching this area. They might want to hire computer scientists to do some of it, as you said, but they … they don't see that their political scientists should have this sort of skill in their tool kit, they don't see it in the same way as even quantitative statistics, it is not really there, and I think being able to manipulate data is much more important than knowing how to run a series of statistical tests, which may or may not be useful.

Gary King, Professor of Government at Harvard University, repeated these concerns, but suggested that students were gradually gaining skills more appropriate to more computational modes of research.

> **Gary King**: So twenty years ago we were trying to take our social scientists and get them trained in statistics enough so that after they were trained in our department, they'd be able to go on to learn statistics in more advanced levels in other departments, like in the statistics department, let's say. And I think we've basically accomplished that, not that they have the right sufficient level of expertise but they've greatly increased their expertise. In fact, students now quite regularly, will end up with a PhD in political science and a Masters in statistics. So, what about now, now actually it's computational issues, so twenty years ago they did sort of baby statistics, and they learned how to make them more sophisticated, now they do sophisticated statistics with baby computer science. … So you see students now, yeah they take their statistics, they figure that out, they of course try to get a social science PhD but they're also now taking computer science courses.

Thus it is only through the combination of greater access to data, and more advanced skills among more researchers, that both the promise and limitations of data can be fully understood. This, in turn, will improve the replicability of research conducted in this area, ensuring that both researchers and reviewers have the ability to scrutinise data and in some cases falsify findings.

Concluding remarks

This paper began with an overview of the development of public opinion, from its philosophical conception in the ancient and early modern eras through its empirical applications in the twentieth century. This outline stressed the significance of various ICTs in influencing the evolving understanding of public opinion, in both societal and academic contexts. The emergence of the Internet, and the myriad applications and uses of it, thus augurs another reconceptualisation of public opinion, and this paper proceeded to investigate what form this would take.

Drawing on interviews with researchers at the cutting edge of online public opinion and related political communication research, this paper identified and investigated three challenges to the validity of this research. None of these challenges are new to public opinion research. But the use of the Internet in research fundamentally reconfigures each of these issues, requiring new standards and norms which venture onto new epistemological and methodological terrain.

Establishing reliable data for public opinion research is a different proposition when that data is being captured passively and tacitly. Even when researchers apply filters to capture only fractions of the masses of content which emerge on the Internet every day, they will not find a neatly ordered and organisable dataset, but rather data that is “messy” by its nature (Mayer-Schönberger & Cukier, 2013). In addition, data collected in this fashion also tends to suffer from a lack of representativeness of the wider population (the public at large) to whom researchers would like to generate findings. Finally, the ability to replicate findings – a crucial foundation of empirical research – is constrained at a time when both access to data and the skills required to manipulate it are limited to a small minority of practitioners.

As well as highlighting many of the practical steps that can be taken to combat these threats to the validity of this emerging area of research, this paper has also offered some theoretical engagement with evolving perspectives on public opinion. In particular, this paper showed how many researchers propose a reconceptualization of public opinion, inverting the twentieth century notion built around statistical certainty provided like concepts of margins of error. No doubt these strict statistical studies will still be vital to our understanding of the ebbs and flows of the public mood – particularly as elections approach and quantitative certainty is required (see Gayo-Avello, 2013 for an overview of largely undistinguished attempts to predict elections using social media.)

Yet, if it can meet the core challenges to validity which have been outlined, research into public opinion as expressed online offers something fundamentally new and different: an opportunity to gain insight into public opinion with more speed, flexibility and often granularity. In this sense, we can see how this new research front stands up better to the criticism levelled by Pierre Bourdieu (1979) against conventional public opinion research. Collecting public opinion data online does not require that everyone has an opinion; is sensitive to the varying ‘value’ of different people’s opinions (conceived in terms of network centrality, for example); and implicitly accepts that not everyone agrees on the question, since no questions are asked, but rather opinions expressed without researcher provocation. While this paper has served to ring alarm bells over questions of validity surrounding this new area of research, it also seeks to issue a clarion call as to its exciting potential.

References

Anderson, B. (2006). Imagined Communities: Reflections on the Origin and Spread of Nationalism (Vol. 17, p. 240). London: Verso.

Aristotle, & Reeve, C. D. C. (1997). Politics (p. 745). Indianapolis: Hackett Publishing.

Beyer, M., & Laney, D. (2012). The importance of 'big data': a definition. Stamford, CT: Gartner.

Bourdieu, P. (1979). Public opinion does not exist. In Mattelart, A. and Siegelaub, S. (eds.) Communication and Class Struggle, Vol 1, pp.124-130.

boyd, danah, & Crawford, K. (2011). Six Provocations for Big Data. SSRN Electronic Journal. doi:10.2139/ssrn.1926431

Bruns, A., & Burgess, J. E. (2011). The use of Twitter hashtags in the formation of ad hoc publics. In 6th European Consortium for Political Research General Conference, 25 - 27 August 2011, University of Iceland, Reykjavik.

Childs, H. (1965). Public opinion: Nature, formation, and role. New York: D. van Nostrand.

Conover, M.D., Gonçalves, B., Ratkiewicz, J., Francisco, M., Flammini, A., and F. Menczer (2011). Political polarization on twitter. In Proceedings of the

5th International Conference on Weblogs and Social Media.

Dutton, W.H. and Blank, G., with Groselj, D. (2013) Cultures of the Internet: The Internet in Britain. Oxford Internet Survey 2013. Oxford Internet Institute, University of Oxford.

Dubois, E. (2013). Telling Vic Everything: Digital Contention and the Traditional

Media. Annual meeting of the American Sociological Association. New York, USA.

Gayo-Avello, D. (2013). A meta-analysis of state-of-the-art electoral prediction from Twitter data. Social Science Computer Review, 0894439313493979.

González-Bailón, S., Wang, N., Rivero, A., Borge-Holthoefer, J., & Moreno, Y. (2012). Assessing the bias in communication networks sampled from twitter.arXiv preprint arXiv:1212.1684.

Hale, S., Yasseri, T., Cowls, J., Meyer, E. T., Schroeder, R., & Margetts, H. Z. (2014). Mapping the UK Webspace: Fifteen Years of British Universities on the Web. Proceedings of WebSci 2014.

Katz, E., & Lazarsfeld, P. (1970). Personal Influence, The part played by people in the flow of mass communications. New Jersey: Transaction Publishers.

Mayer-Schönberger, V., & Cukier, K. (2013). Big data: A revolution that will transform how we live, work, and think. New York: Houghton Mifflin Harcourt.

Mill, J.S. (1946). On liberty; and, Considerations on representative government. London: Blackwell.

Nordin, A., & Richaud, L. (2014). Subverting official language and discourse in China? Type river crab for harmony. China Information, 28(1), 47-67.

Peters, J. (1995). Historical tensions in the concept of public opinion. In Glasser, T. and Salmon, C., Public Opinion and the Communication of Consent. New York: Guildford Press

Pew Research (n.d.) Social Networking Fact Sheet. Retrieved from <http://www.pewinternet.org/fact-sheets/social-networking-fact-sheet/>

Pew Research (May, 2012) Assessing the Representativeness of Public Opinion Surveys. Retrieved from <http://www.people-press.org/2012/05/15/assessing-the-representativeness-of-public-opinion-surveys/>

Plato. (1970). The Laws. Harmondsworth: Penguin.

Ratkiewicz, J., Conover, M., Meiss, M., Gonçalves, B., Patil, S., Flammini, A., & Menczer, F. (2011, March). Truthy: mapping the spread of astroturf in microblog streams. In Proceedings of the 20th international conference companion on World wide web (pp. 249-252). ACM.

Schroeder, R., & Cowls, J. (2014). Big Data, Ethics, and the Social Implications of Knowledge Production. Presented at Data Ethics Workshop, KDD@Bloomberg, 24th August 2014. Retrieved from <http://dataethics.github.io/papers/BigDataEthicsandtheSocialImplicationsofKnowledgeProduction.pdf>

Stephens-Davidowitz, S. (2013). The cost of racial animus on a black presidential candidate: Using google search data to find what surveys miss.SSRN Journal 2012: 1, 55.

Stephens-Davidowitz, S. (2013). How googling unmasks child abuse. The New York Times.

Tarde, G. de. (1901). L'opinion et la foule. Paris: Alcan.

Taylor, L., Schroeder, R., & Meyer, E. (2014). Emerging practices and perspectives on Big Data analysis in economics: Bigger and better or more of the same? Big Data & Society, 1(2). doi:10.1177/2053951714536877

Tocqueville, A. de. (2004). Democracy in America (p. 941). New York: Library of America.

World Bank (n.d.) Internet users per 100 people [dataset]. Retrieved from <http://data.worldbank.org/indicator/IT.NET.USER.P2/countries>