



## Big data and Wikipedia research: social science knowledge across disciplinary divides

Ralph Schroeder & Linnet Taylor

**To cite this article:** Ralph Schroeder & Linnet Taylor (2015) Big data and Wikipedia research: social science knowledge across disciplinary divides, *Information, Communication & Society*, 18:9, 1039-1056, DOI: [10.1080/1369118X.2015.1008538](https://doi.org/10.1080/1369118X.2015.1008538)

**To link to this article:** <http://dx.doi.org/10.1080/1369118X.2015.1008538>



Published online: 24 Feb 2015.



Submit your article to this journal [↗](#)



Article views: 1010



View related articles [↗](#)



View Crossmark data [↗](#)



Citing articles: 4 View citing articles [↗](#)

## Big data and Wikipedia research: social science knowledge across disciplinary divides

Ralph Schroeder<sup>a\*</sup> and Linnet Taylor<sup>b</sup>

<sup>a</sup>*Oxford Internet Institute, University of Oxford, 1 St. Giles, Oxford OX1 3JS, UK;* <sup>b</sup>*Faculty of Social and Behavioural Sciences, University of Amsterdam, Nieuwe Achtergracht 166, 1018 WV, Amsterdam, Netherlands*

(Received 18 December 2013; accepted 5 January 2015)

This paper examines research about Wikipedia that has been undertaken using big data approaches. The aim is to gauge the coherence as against the disparateness of studies from different disciplines, how these studies relate to each other, and to research about Wikipedia and new social media in general. The paper is partly based on interviews with big data researchers, and discusses a number of themes and implications of Wikipedia research, including about the workings of online collaboration, the way that contributions mirror (or not) aspects of real-world geographies, and how contributions can be used to predict offline social and economic trends. Among the findings is that in some areas of research, studies build on and extend each other's results. However, most of the studies stay within disciplinary silos and could be better integrated with other research on Wikipedia and with research about new media. Wikipedia is among few sources in big data research where the data are openly available, unlike many studies where data are proprietary. Thus, it has lent itself to a burgeoning and promising body of research. The paper concludes that in order to fulfil this promise, this research must pay more attention to theories and research from other disciplines, and also go beyond questions based narrowly on the availability of data and towards a more powerful analytical grasp of the phenomenon being investigated.

**Keywords:** Wikipedia; big data; interdisciplinarity; new media

### Introduction

In this paper, we examine how different disciplines analyse Wikipedia. We focus not only on social science disciplines, but also examine research that is carried out outside of social science disciplines but related to social science questions. One reason for choosing Wikipedia is that research about Wikipedia for addressing social science questions has rapidly taken off in recent years, much of it using big data or computational approaches. The core question of the paper is: Does big data research about Wikipedia constitute a coherent body of work which advances our understanding of new media, or does this body of work consist of disparate findings that are contained within disciplinary silos? We shall argue that indeed, studies of Wikipedia are a burgeoning research area with contributions from many disciplines, but that these contributions could be even more powerful if they were more explicit about the object they are investigating and its social significance. The paper reviews a number of studies about Wikipedia, asking in

---

\*Corresponding author. Email: [ralph.schroeder@oii.ox.ac.uk](mailto:ralph.schroeder@oii.ox.ac.uk)

each case about the strengths and limitations of distinctive disciplinary approaches. Thus, we shall review the contributions and claims of these studies, and how these fit into the disciplines in which they are conducted – or go beyond them.

Wikipedia is consistently ranked as one of the top websites visited globally. Currently, it is the sixth most visited website (<http://www.alexa.com/topsites>, last visited 27.10.2014). This makes Wikipedia unique since it can be counted among new social media (van Dijk, 2013, pp. 132–153) which yield valuable insights into, among other topics, the production of user-generated content. We focus specifically on ‘big data’ research, which has also rapidly taken off in recent years, particularly in research on new social media (Golder & Macy, 2014; Schroeder, 2014a). But again, one reason Wikipedia is unique in big data research is that the data it provides are open. This makes it different from big data sources such as Google searches since it is not clear how the data are arrived at, and so studies using Google (which is most highly visited website world-wide) face the problem, among others, that they cannot be replicated (see Lazer, Kennedy, King, & Vespignani, 2014). To take another example, Twitter (which is number 8 among top websites) also has issues of validity if the researchers do not have access to the full data set, which can lead to biases in data analysis (see González-Bailón, Wang, Rivero, Borge-Holthoefer, & Moreno, 2014) or having to pay for access to the full data set, which is beyond the means of most researchers (see Puschmann & Burgess, 2013 for the conditions under which the data can be obtained). Wikipedia is thus a more reliable, transparent and free source of big data that can be built upon.

Big data is often defined in terms of the three V’s: high volume, high variety and high velocity (Gartner, 2011). However, Wikipedia does not fully meet the criterion of ‘velocity’ since the data, although constantly updating, for the purposes of research are downloaded as a single ‘dump’ comprising major portions or sometimes the entire history of the platform up to a given moment. This makes Wikipedia, for researchers at least, more static than other popular web platforms such as, say, Twitter, which produces a lot of data quickly and can be studied in real time. As we shall see, some Wikipedia studies also do not use a variety of data, but instead use data from one or few dimensions. Therefore, we use a somewhat different definition: research that marks a step change in terms of its scale and scope in advancing knowledge in relation to a given object or phenomenon (Schroeder, 2014b). Wikipedia, an object with millions of contributors, contributions and interactions between them, clearly meets this criterion (which equates to the first ‘V’, volume), but also has the feature, like other big data studies of new social media (Golder & Macy, 2014), that they often use the data set about the whole of the object, that is of the whole of Wikipedia during a certain period, or how the whole of English-speaking Wikipedia covers certain topics.

Wikipedia provides data about collaboration and user-generated content that marks a step change in terms of scale and scope from data that is available about large-scale collaborative efforts or about knowledge or information that is produced via mediated social relations. Note that what is distinctive here is the scale of the data that is readily available for computational manipulation: as we shall see, this property of the object has advantages, but also poses challenges about how to situate the results of the research. In any event, Wikipedia is an example of a web-based phenomenon that makes available for research a large self-contained whole ‘universe’ of activity (Twitter and Facebook provide other examples, but access to the whole of the Facebook data set requires special circumstances, such as being part of the team within the company), and this is a typical feature of big data research – again, as we shall see, with certain advantages and limitations.

This paper will begin by reviewing previous work which has assessed Wikipedia research and work related to how the social sciences relate to each other and to other disciplines. We also discuss our method and compare it to other approaches. Then, we will review a number of

different studies of Wikipedia, describing in each case briefly how the studies can be located in disciplinary terms, the main findings, and some of the challenges and advantages of the research. We have organized these not under disciplinary headings, but under the main questions or aims of the study, beginning with collaboration, moving on to studies related to geography and conflict, and finally research aimed at predicting social trends. While we will discuss some of the limitations of these studies in each case, we leave until the conclusion a more sustained analysis of how the studies interrelate (or fail to do so), and how they contribute more generally to social science knowledge about new media.

### **Background, previous research and method**

Wikipedia, again, is among the top 10 most visited websites ([www.alexa.com](http://www.alexa.com)) and contains more than 4.5 million articles (<http://en.wikipedia.org/wiki/Wikipedia:Statistics>). Research about Wikipedia and using Wikipedia has grown rapidly since Wikipedia started in 2001. There are now reviews and lists of research topics, including research in all disciplines (Okoli, Mehdi, Mesgari, Nielsen, & Lanamäki, 2012, see also [http://en.wikipedia.org/wiki/Academic\\_studies\\_about\\_Wikipedia](http://en.wikipedia.org/wiki/Academic_studies_about_Wikipedia)). But it is difficult to quantify how much scholarship about Wikipedia is carried out by which discipline. The number of articles about Wikipedia depends on which source is used and can differ widely: Okoli et al. based on Park (2011) found 1746 articles in Web of Science and Scopus, but Wikipedia's own list has 607 items. Similarly, Bar-Ilan and Aharony (2014) found quite different numbers of publications, depending on the publications database used. However, they carried out a review of publications related to Wikipedia by extracting all articles with Wikipedia in the title of the article, the abstract and keywords in the bibliographic database most commonly used in bibliometric studies (Elsevier's Scopus), obtaining 2968 relevant publication records.

Bar-Ilan and Aharony (2014) found that there are almost an equal number of publications 'about' Wikipedia (1431) as there were 'using' Wikipedia (1537), but there were more publications with a 'technological approach' (1856) compared to a 'social approach' (1112). It should be noted here that what they call the 'social approach' is broader than what we examine here (social science) since it includes, for example, visualizations of Wikipedia which may or may not fall into social science. Bar-Ilan and Aharony have one further finding that is relevant to report here, which is that after a steep take-off between 2005/2006 and 2010, the number of publications has slowed down and plateaued between 2010 and 2012, and this applies to all the categories of publications they examined ('about' versus 'using', 'technological' versus 'social'). This finding suggests that Wikipedia research has established itself as a sizeable (there were more than 500 publications per year in all the four categories between 2010 and 2012) but no longer rapidly expanding field of research (if we assume that these trends will continue). Despite being able to analyse the overall number of publications about Wikipedia (though with varying results, depending on the database), it is not possible to quantify the studies that fall within our definition of big data approaches because they are often not labelled as 'big data' studies.

What we did instead to capture the main studies that fall within the definition used here is to include only studies that Wikipedia as an object for social science research and that use big data approaches. Wikipedia is clearly proving to be popular object of research, analysed from a range of perspectives (for an overview, see Reagle, 2010), but we sought out the subset of 'big data' studies by systematically examining the annual conferences about Wikipedia research and open collaboration such as WikiSym (<http://www.wikisym.org/>) and Wikimania (<http://wikimania2014.wikimedia.org/wiki/Wikimania>). We also singled out research that falls within big data from the reviews of research just mentioned. This research is undertaken in many disciplines, often without awareness of related work in other fields, something that we became aware

of when asked our interviewees about this. In this respect, Wikipedia is similar to other new media, such as Twitter or Facebook or search behaviour on Google, which also have several disciplines tackling a variety of topics and which also are experiencing rapid growth. We therefore sought as broad a range of disciplinary perspectives of big data Wikipedia research as we could find, using the various reviews and conferences as well as asking our interviewees about related research (which can also be found in the references of their papers).

As already mentioned, Wikipedia does not pose the same kinds of constraints about privacy and replicability of findings that often present challenges to research on other new media, which may mean, for example, that big data research in particular is primarily carried out by researchers with privileged access to data. At the same time, Wikipedia raises questions that are different from other new media: for example, in so far as it is not a commercial service or social network, how does collaboration on Wikipedia compare with other forms of online or offline collaboration? Collaboration has thus been one of the main topics of research about Wikipedia (Reagle, 2010).

In this paper, we examine various social science topics related to Wikipedia. Our key question is how coherent or otherwise this research is: Do different social science disciplines (and disciplines outside the social sciences) work towards a common goal across disciplines, or are they isolated within their own domain? Do they build on previous work (also in the social sciences generally, not just about Wikipedia), or pursue new directions without drawing on related work? One argument that has been made by Whitley (2000) about how social science disciplines are integrated compared to natural sciences relates to ‘mutual dependence’ (or the necessity to build on previous work), which Whitley argues is low in the social sciences. Another argument is that disciplines are often typically protective of their ‘turf’ or ‘territory’ (Becher & Trowler, 2001), and this holds for natural and social sciences as well as for humanities, though to varying degrees. Others have argued against the idea that social science is not integrated: Rule (1997), for example, has argued in the face of scepticism that certain areas in the social sciences are cumulative and that it is in fact possible to identify areas of social science advances in relation to certain topics and approaches. Finally, there have been a number of studies which have examined interdisciplinarity (for an overview, see Klein, 1996), which analyse how different disciplines work together in terms of collaboration and how they are published in different venues (e.g. conference proceedings versus journals) for different audiences.

This research is part of a larger project which examines social science big data research funded by the Sloan Foundation, which has so far interviewed more than 100 researchers (more than a dozen of whom have published research about Wikipedia) and held a number of workshops about this topic. The interviews took between half an hour and one hour and were transcribed and coded for content. We used a mixture of structured and unstructured questions. (The interviewees for this paper are listed separately after the reference list and all quotes from interviews in the text are indicated as such.) Our project used techniques including snowball sampling and contacting experts for interview selection. In this paper, we selected studies and interviewees that are clearly representative of a wide range of disciplinary approaches, including interdisciplinary work. The aim is to illustrate the distinctiveness of both particular social science disciplines (e.g. economics or geography) and the variety of topics and methods. Such a selection cannot be exhaustive, but this is not necessary in view of the fact that the aim is to show how disciplines and topics converge on common findings – or how they fail to engage with each other.

### **Promoting more contributors and enhancing collaboration**

One particular issue that has been of concern to Wikipedia is that the number of contributors, which had previously experienced rapid growth, has begun to taper off or plateau. This is related to a second issue of concern, which concerns the diversity of contributors, which over-

represents men and over-represents certain languages. In economics, one way to think about contributions is in terms of ‘group size and incentives to contribute’ (Zhang & Zhu, 2011). The two authors of this paper have backgrounds both in economics and in computer science (Zhang Interview, 2013). They studied Chinese-language Wikipedia, which has been blocked and unblocked on the mainland of China for certain periods, including being selectively blocked in certain geographic areas. This blockage – or censorship – allowed the two authors to gather data from a ‘natural experiment’ where the ‘experiment’ that took place was the blocking and unblocking of Wikipedia. This provided different conditions under which contributors could join (outside of mainland China during these periods) or be prevented from joining (on the mainland). The reason they have a quasi-experiment here is that contributors outside the mainland were not blocked, so they have a comparable ‘control group’. Their hypothesis was that when groups of contributors grow in number, contributions drop. The reason for this hypothesis rests on a well-established assumption in economics: Wikipedia is an example of what economists call a ‘public good’; essentially, something provided for free to many users – such as (non-toll) roads or parks. This premise can be combined with the so-called free-rider problem (Zhang & Zhu, 2011): put briefly, if I know that others will contribute anyway to something that I can benefit from, why should I bother (e.g. donating blood)?

The free-rider problem raises a host of issues in economics about altruistic behaviour or peoples’ willingness to contribute to public goods, a topic that is typically studied in a laboratory setting. Usually, this research takes the form of providing a small number of student participants with artificially constructed scenarios whereby they are offered choices about how they contribute in the light of the other participants’ choices, often iteratively. As Zhang and Zhu (2011) point out (and this is a feature of many big data studies), studying this phenomenon instead on a large scale and in the ‘natural’ setting of a real-world task has a number of advantages over small-N laboratory studies.

What did they find? The authors discover that when a large proportion of contributors are blocked from participating, the contributions of those who are not blocked also decrease. They also find that ‘the more contributors value social benefits’ – social benefits are the ‘warm glow’ or ‘moral satisfaction’ or ‘joy-of-giving’ that people feel when they are part of a common effort in large groups, which can override the utility that a person needs in contributing to a public good – ‘the greater their reduction in their contribution after the block’ (Zhang & Zhu, 2011, pp. 1601, 1602). In other words, when contributors see that fewer others are contributing, then they are more likely to stop as well. This finding contrasts with the behaviour found among contributors who were not blocked. The study makes an important contribution to research on public goods and free-riding because the largest previous experiments were based on groups ranging from 40 to 100 subjects, whereas the Wikipedia study assesses effects within a population of 21,496 contributors – a significant change in scale from previous research.

Several other features of the study are noteworthy. First, it was published in the *American Economic Review*, the top journal in the field. This needs to be highlighted because economics journals are comparatively exclusive in publishing only articles that fall within the purview of economics as a discipline. Second, the method used here is regression analysis, a standard method for experiments – but again, platforms such as Wikipedia are unique in providing readily accessible data about all transactions from (in this case) a ‘natural’ experiment which readily provides data to perform regression analysis. Third, the approach here to understanding the ‘incentives’ to contribute is clearly based on rational economic motivations: other studies of Wikipedia, as we shall see, use quite different understandings about why people contribute to a collaborative effort.

It can be mentioned that the paper has two principal audiences: economists who are interested in testing the validity of existing findings about free-riding and public goods; and those, such as the developers or organizers of Wikipedia (and perhaps contributors) or similar web-based

platforms, who may be able to use the insights from the study to enhance contributions or otherwise improve Wikipedia or similar tools. However, while the paper has interesting findings for economic understandings of collaboration (or contributions to shared efforts) which could also inform the design of collaborative platforms, it is also useful to note a major limitation for a broader understanding of Wikipedia or for the role of social media in China, which is that the study relates to two settings (Chinese-language Wikipedia inside and outside mainland China). However, within China, there is a rival online encyclopaedia, Baidu Baike, developed by an internet company that is close to the government, which has many more contributors and Chinese-language users and content (Liao, 2009). If Zhang and Zhu would put their study into the context of comparing the two dominant online encyclopaedias in China, they might shed light not only on the economics of collaboration, but also about, for example, competition between new social media in the Chinese-speaking world.

The comparison between Chinese Wikipedia and Baidu Baike raises many questions which are beyond our scope here: suffice it to say that while Chinese-language Wikipedia is an opportune object of research because it offered the experimental condition of blockage and non-blockage, and thus for the specific question that was answered from an economics perspective, these findings apply to the rather unique circumstances of Chinese Wikipedia in China. It is not clear whether the findings would apply to Wikipedia contributors in other languages, or contributors in other platforms for aggregating user content, or outside of conditions where there was an alternation between being blocked and not being blocked. There is also a strength to this study, which is that most research about Wikipedia is limited to the English-speaking version. In any event, this study made use of experimental conditions to advance knowledge in economics about contributions to a public good, knowledge that could also be relevant to enhancing collaboration. As we shall see, this kind of practical focus – how to improve online collaboration – is a characteristic of many Wikipedia studies.

Another study focusing on collaboration is by West, Weber, and Castillo (2012). They are not academic social scientists but rather computer science researchers working in the private sector (one of the authors works at Yahoo!, and all of them worked there while working on this paper). Their study is interesting in part, however, because the authors have data not just about Wikipedia, but also about who uses Wikipedia and how they use it (at least if they use Yahoo!).

One of the authors, West, comments that:

I wasn't so influenced by sociological theories, because I just don't have a big background in that. So it's more post hoc, that people tend to fill in, I think, the sociological things. ... People still mostly come from, like the data angle, and we have this data set, we want to make interesting findings. And then maybe afterwards, you try to fit it in with sociology. (West Interview, 2013)

Yet their analysis of Wikipedia editors clearly addresses social science questions, including which topics contributors are interested in contributing to, and how knowledgeable they are about the topic. They know how knowledgeable people are because they had access to data from Yahoo!'s toolbar, which captures data about each site that people visit (as long as users enable this feature). Thus, they could combine data about Wikipedia contributions of users with data about all other sites that these users visit (again, if they use Yahoo!, and if they have enabled the toolbar). The authors are aware of the potential bias of examining only Yahoo! toolbar users, but they make a strong case that this should not influence the results. We can note, however, that access to proprietary data is a feature that sets this study apart from the others that we consider here (and, it can be mentioned in passing, data about websites that users click on is rarely accessible to academic researchers).

Among the findings is that contributors to the entertainment-related part of Wikipedia, which makes up '7 of the 10 largest categories of article topics' (West et al., 2012, Section 6), look for more information on these topics than non-contributors; put differently, they seem to be more expert than others, which the authors describe as being more 'information hungry'. Furthermore, when they break this expertise down into 'science, business and humanities' as against 'entertainment-related' editors, they find that the former are more 'generalist', whereas the latter are 'from editors immersed primarily in popular culture'. Here, again there are practical findings from the research, which, according to the authors, could improve how Wikipedia encourages contributors to contribute (e.g. promoting contributions from contributors with certain types of more general or more specific types of knowledge).

As already mentioned, one of the reasons why this topic has practical significance is because there has been a concern that the number of Wikipedia contributors has been declining in recent years, so finding a match between what people know about and where they are likely to contribute most could have implications, for example, in campaigns to encourage new contributors. In this respect, there are various bodies of social science knowledge that this research might connect to, including about social mobilization and about the motivations for joining social networks. There is also a more general social science question that the authors address, in addition to a practical one, when they recommend that Wikipedia contributions could be enhanced by fostering 'diversity', for example via projects that appeal to subgroups with certain kinds of specialist knowledge. Furthermore, this recommendation that knowledge production is enhanced by greater diversity is potentially applicable beyond Wikipedia.

### Mapping knowledge, conflict and language

Studies related to geographical location have been another focus of Wikipedia research, and this can be illustrated first by the work of Graham (2011) who is interested in mapping the geographies of knowledge production. Graham is particularly concerned with the correspondence – or better, lack of correspondence – between offline and online knowledge, and how this place-relatedness is illuminated by Wikipedia content. These correspondences, or the lack thereof, according to Graham, are about power relations or about what is visible and invisible in relation to the world's places. To investigate this question, he limits himself to Wikipedia content that is 'geo-coded'; in other words, tagged as relating to particular places or events in places. Thus, he is able to show, for example, that there is a vast amount of geocoded content related to the United States as against the dearth of content about Africa. Seen in proportion to landmass, however, Central and Western Europe, Japan and Israel have the most articles, whereas large countries such as Canada and Russia have comparatively few. Finally, taking population size into account, the picture changes again, with Canada, Australia and Greenland having a large number of articles in respect to their relatively small populations.

Graham is aware of the limitations of focusing on geocoded articles (Graham Interview, 2013). Yet these data answer a problem geographers are currently debating: How to analyse spatial problems using the internet, which does not exist in physical space? Graham comments that 'the data itself is spatial, because it is attached to a place, it is fixed onto some part of the world, it has co-ordinates' (Graham Interview, 2013). His work demonstrates how this rootedness of digital information in physical places causes a creative disjuncture which opens up new perspectives on the study of the Web: 'These two things do not fit together, it is why we try and make this spatial, and it breaks open the idea, it is an interesting way. It sort of cracks the idea' (Graham Interview, 2013). Graham's results show that content relating to different parts of the world is highly uneven. He argues that while the specific findings that have just been mentioned (relation to population size or landmass) are perhaps what might be expected, the imbalance

between the global North and South is surprising. This imbalance supports ideas about the dominance of the global North and also ideas about the lack of representation among the more powerless parts of the world. There are a number of reasons why this imbalance matters, if we consider, for example, how certain places and the events associated with them are highly political (we can think here of Israel and the Palestinian territories as an obvious example). To this it can be added that, as mentioned earlier, when searching for information about places (as with other searches for information), Wikipedia entries are often highly visible among the search engine results, typically among the top results for Google searchers (though there are disputes about this visibility, compare <http://searchenginewatch.com/article/2152194/Wikipedia-Appears-on-Page-1-of-Google-for-99-of-Searches-Study> with <http://searchengineland.com/why-the-wikipedia-google-search-results-study-is-flawed-111628> both last accessed 20.10.2013).

In any event, if there is no such place-related information, then information about these places will not be available; these places will be less visible or invisible online. However, the significance of location-related research, as with other Wikipedia research, could be highlighted much more with information about the popularity of Wikipedia, including the popularity of different topics and which readerships depend on this knowledge. Another obvious limitation specifically of this study is that it is difficult to know what to infer about geographical knowledge imbalances from this – albeit very important – online resource per se: How powerful of an indication is geocoded Wikipedia material, as opposed to Wikipedia content generally, or (as mentioned) rival online encyclopaedias which dominate in other parts of the world (Baidu Baike)? Furthermore, it could be asked: Why should Wikipedia be taken as an (albeit convenient) proxy for knowledge production, when so much is known already about global divides regarding where cultural goods and services are produced and consumed (e.g. Norris & Inglehart, 2009, pp. 82–83)?

Physics takes yet another approach to Wikipedia in relation to location. Indeed, one study with researchers whose background is in theoretical physics, Yasseri, Sumi, Rung, Kornai, and Kertesz (2012), uses the term ‘sociophysical studies’. Describing why physicists are working on data deriving from the internet, Yasseri notes many commonalities with the classic problems of physics:

I see many things in common, actually, like the main concepts, which is like building the whole system based on the features of the elements. This is what then we tried to do in studying Wikipedia, which is a collective behaviour, an emerging phenomenon, coming from like millions of people. Of course, we cannot get access to information for all these individuals, but we could characterise them, like with few features, few attributes, to each editor, and then based on that we wanted to see how the whole thing emerges, how the collective behaviour is governed. (Yasseri Interview, 2013)

Yasseri et al. are interested in a classical sociological question – conflict – and how this manifests itself in Wikipedia’s so-called edit wars: edit wars take place when different editors quickly change the content of an entry because they disagree over this content, typically because it is a controversial topic. Conflict can be examined by how frequently Wikipedia articles are edited, dividing these into relatively peaceful as against controversial ones. Controversial articles are few in number, but within these, Yasseri et al. focus on ‘mutual reverts’, which happens with articles where the changes to an article are made in a rapid back-and-forth manner: changing the content to re-instate the previous content because of disagreements. In other words, these are articles that are subject to highly conflictual editing. Based on examining all articles, they are able to establish that these articles, what they call ‘never-ending-wars’, are a tiny number – less than 100 in the set of 3.2 million articles, and that although these wars are carried on by a very small number of editors, they occupy a disproportionately large amount of editor’s time.

Apart from what these findings tell us about what people (or Wikipedia contributors) consider to be conflict-laden topics, they could also have implications for editors and how to resolve conflict more effectively (again, improving how Wikipedia works). It is curious, however, that the authors use ‘conflict’ here, rather than, for example, linking their research to the concentration of user-generated political content (e.g. Hindman, 2009, pp. 82–128), or about how conflicts are resolved in practice in Wikipedia (e.g. Reagle, 2010). These would be interesting links in view of how few articles are controversial. Further, there is a conflict tradition in sociology which applies mainly to violence, but which can also be related to knowledge production (Collins, 1975, pp. 470–523) and might also be useful in understanding the flipside of conflict; consensus, in Wikipedia (where conflict is very rare) and elsewhere in online collaboration. In short, putting this research into broader social science contexts could highlight interesting features about Wikipedia apart from locating rare conflicts.

Political science has also sought to map conflict and stability by means of Wikipedia, for example in the paper by Apic, Betts, and Russell (2011), where Wikipedia disputes are taken as an index for geopolitical instability. The authors situate their work at the intersection of biology and political science: all the three authors are biologists and they work in the private sector but also have academic affiliations. Their aim is to show that web content such as Wikipedia can ‘complement more arduous metrics’ (2011, p. 4) regarding conflict and instability. The paper takes a premise from biology that one can predict the role of a new molecule from the molecules it is associated with. The authors use this idea to look at whether online disputes about the content of Wikipedia articles about particular countries reflect actual conflict and instability in those countries. They test their hypothesis using existing indices of political conflict from the *Economist* magazine and from the World Bank, and find that Wikipedia content disputes do indeed correlate with actual political instability as measured by these indices.

The paper again illustrates some advantages and possible pitfalls of this kind of work. For example, language could be an important and possibly confounding aspect of the analysis since the authors include the English, German, Spanish and French-language Wikipedia versions while almost none of the countries classified by the analysis as unstable use these languages (and recall our discussion of the two versions of Chinese-language online encyclopaedias). There is another potential representativeness issue with the authors’ use of comparator indices drawn from differing, though overlapping, periods. Within the discipline of computer science, these representativeness issues are less important, because they do not affect the validity of the statistical and computational aspects of the analysis. However, as this is an analysis which also ventures into social scientific territory – whether international relations, political science or geography – the question of periodization is important.

The study’s strength is in its simplicity – it uses a basic process of correlation to link online with offline disputes (as does Graham) – and finds that such a simple correlation comes close capturing the reality of disputes. The authors comment that ‘it is remarkable that so simple a metric can agree so well with more complex measures of political and economic stability’ (Apic et al., 2011, p. 4). Apic et al. treat Wikipedia as a single political entity, effectively a static whole, which is different from other studies here, such as Yasseri et al. (2012), who focus precisely on changes over time.

There is a further paper which examines conflict, or in this case overcoming language barriers (Bao et al., 2012), but which is different again from Apic et al.’s (2011) and Yasseri et al.’s (2013) research. Bao et al. focus on the differences in language across Wikipedia versions and analysed the user experience of reading across different cultures as reflected by language. The paper outlines the creation of the ‘Omnipedia’ system that gives users access to multiple language platforms within Wikipedia. Darren Gergle, one of the paper’s authors, describes the paper as:

essentially taking the idea that we need to retain these distinctions and differences in these different language editions and representations, and instead of translating across them or covering them up or taking the weighted average of the most common representation and presenting that, actually showing the overlap and distinctions and differences ... making that salient and then designing systems that actually retain that and highlight that as opposed to kind of masking it or covering it up or treating it as a bug. (Gergle Interview, 2013)

It is this approach to linguistic and cultural complexity which sets Bao et al.'s paper (2012) apart in disciplinary terms: although it is clearly situated in computer science in terms of its aims (system development and testing), and although the principal approach used in the paper is computational, describing how machine learning is used to bridge between different language editions of particular articles, it also relates to the study of cross-cultural communication in seeking to evaluate the experience of reading across cultures, and exploring information-seeking behaviour in a multilingual context.

The authors used human volunteers to evaluate the application, observing 'how people gained insights when viewing concepts of their choice through Omnimedia's hyperlingual lens' (Bao et al., 2012, p. 1082). The article thus moves from a focus on algorithm design and testing to evaluating how peoples' experience shifts across language and cultural perspective. The paper, which was published in the proceedings of a computer science conference, presents findings both about the technological system devised to bridge different language editions of Wikipedia and about the test subjects' experience of the system (they sought differences and similarities in perceptions of concepts, filtered with the influence of self-focus bias or bias based on the users' language, and sought a 'big picture view' of other cultures' treatment of topics). The paper's aim can be described as being to provide a multifaceted view of a new system which will serve to give that system, and its related worldview, traction among users, instead of merely acting as a proof-of-concept for the domain of computer science. Gergle notes that this mixed-method approach derives from the particular composition of his research group, which spans Communications Studies, Engineering and Computer Science. For him, the group focuses on a 'theory-driven design', using social theory and an understanding of human behaviour and user experience to inform the design and development of systems 'as opposed to just using theory to analyse and critique systems' (Gergle Interview, 2013). Yet again, the paper does not venture beyond the aim of improving Wikipedia uses to engage, for example, with wider questions about intercultural communication or the implications of Omnimedia for cross-cultural dialogue.

### **Predicting social and economic trends**

We have already encountered the 'sociophysical' approach of Yasseri et al. (2012). In a different paper, Yasseri and two colleagues used Wikipedia to predict movie box office success, which relates to business, marketing and economics rather than to 'conflict'. Mestyán, Yasseri, and Kertesz (2013) examined 312 movies released in the United States in 2010 to see if the level of Wikipedia activity (views of pages related to the movie plus three measures of editing levels) before the movie's release corresponds with the movie's earnings. Remarkably, they found that Wikipedia activity is a good predictor of box office success, and one indication of the accuracy of prediction here is that Wikipedia activity does a better job of prediction than Twitter did in a previous study (Asur & Huberman, 2010). In the case of box office prediction, the online world clearly does not just mirror the offline world, but can also be used in forecasting patterns in the offline world. In this sense of prediction, perhaps the label 'sociophysics' is appropriate inasmuch as it points to the scientific aspirations of social science.

A similar attempt at economic prediction is Moat et al.'s paper on Wikipedia usage patterns and stock market trends (2013). The authors seek to understand the role played by online sources of information in early stage decision-making processes. In doing this, they draw mainly on cognitive science, computer science and economics to analyse how Wikipedia searches for particular stocks can forecast the movement of those stocks on the market. Methodologically, the paper uses a quasi-experimental approach which brings together behavioural psychology with economics, comparing a strategy of buying and selling stocks based on Wikipedia page views to a hypothetical 'null' model where stocks are bought and sold randomly. The returns from the Wikipedia page view-based strategy are significantly higher than the random strategy. Contrary to other emerging studies by computer scientists such as Bollen, Mao, and Zeng (2012), which predict financial trends based on a positive correlation of online sentiment with investor behaviour, Moat et al.'s model treated increased page views as a sign of investor concern, and therefore related increased Wikipedia views to a greater likelihood that the stock in a company would be sold.

This research was conducted largely by physicists, but the lead author, Moat, is a behavioural psychologist with computer science and linguistics training, and her collaborator Preis is an economist who originally trained as a physicist. The process of the research involved a mixture of developing code to analyse Wikipedia page views, using the information on edits which Wikipedia itself makes available, and applying advanced statistical methods to compare the random with the Wikipedia-guided investment strategy. Although the statistics deriving from team member's physics training was central to the analysis, coding skills were central to accessing the data:

I wrote a script in Perl to download the data on how often people had looked at these pages on Wikipedia from a resource called stats.grok.se [an openly available online analysis tool for this purpose]. And my student, Chester Curme, wrote a script in Python and he parsed the pages, the history pages, on Wikipedia, which tell you who's edited the pages, to pull out the information on how many edits there had been to a page. (Moat Interview, 2013)

Although coding and statistics were central to accessing and analysing the data, the research was strongly influenced by Moat's interdisciplinary background.

I guess my instinct is very much to say, well, what are the processes behind this, why might we see these results, and to try and just automatically pull back what we see to models of decision-making or biases we know about decision-making, so loss aversion, etc., which is a concept we can use to try and explain some of our results ... [and] because I was interested in language, I'd done a lot of linguistics, and a lot of this information online – because it's communication data – is also via the medium of natural language, and so being able to quantify semantics or understand language processes has also been very helpful in trying to understand, for example, which pages people might be looking at before ... (Moat Interview, 2013)

The paper thus benefits from a highly interdisciplinary perspective, which facilitated the dialogue between theory and experiment in the research. Moat identified three particular advantages in using Wikipedia as an object for this research: first, that its size meant that the researchers could incorporate large numbers of people's activity in the analysis; second, that Wikipedia's accessibility meant that the team could access data on page views and edits very quickly; and third, validity. The latter operated both in terms of the size of the sample which could be accessed – since gathering data on as many users as possible gave the statistical approach greater analytical power – and in terms of the behavioural psychology aspect of the research, where accessing the data directly tells a more truthful story of people's online activities than, for example, reported behaviour does.

Linguistics is another area where it is possible to use Wikipedia for prediction, in this case for word frequencies. Serrano, Flammini, and Menczer (2009) conducted such a linguistic study of Wikipedia, along with two other large public online databases, in order to test a theory of the way in which words are correlated online and to rank those correlations (building on Zipf's law, which proposes more generally that different types of words can be ranked). The authors come from the disciplines of physical chemistry and computer science, and note that they are using a linguistic analysis to inform natural language processing, a subfield within computer science but also partly in linguistics. The border between linguistics and computer science is fuzzy here since both call on sophisticated statistical skills and both are aimed at understanding language as well as the workings of computing systems. The paper also draws on the idea of emergent topology of complex information, similar to the work of Akdag Salah, Gao, Scharnhorst, and Sucheki (2011), which is situated in the humanities.

Thus while the paper is situated in linguistics, it arguably also belongs in computer science (natural language processing) and uses network analysis, which is currently popular in many disciplines (Freeman, 2004). The study therefore highlights the multidisciplinary nature of 'big data' approaches particularly well, spanning more than the combination of computer science with one other field. The authors demonstrate a simple way of modelling highly complex attributes of natural language (bursts of rare words in particular texts and correlations within and across documents) with a simple generative approach which can provide insights for text mining and web analysis. The research has a predictive aspect, in that the model relies on the first few lines of any document as a proxy for the characteristics of the rest. Like other predictive work discussed here, however, it raises the question of whether this work is capable of predicting more than a narrow set of phenomena.

## Conclusion

The big data research on Wikipedia presented here provides a small but wide-ranging sample of studies in social science disciplines or in disciplines relevant to the social sciences. Our aim has not been an exhaustive review of this research (as mentioned, this would be impossible since research often does not self-identify as big data), but we have analysed some of the most prominent directions in this research area. Some researchers, as we have seen, use Wikipedia in order to shed light on phenomena other than Wikipedia, others examine the workings of Wikipedia in its own right. All of them provide examples of how Wikipedia has become a popular object of study because it contains data about so many interactions and so much content, all of which is readily available for computational analysis and for potential correlations with data about other phenomena of social science interest. The ready availability of big data about Wikipedia, as we have seen, has yielded a rich array of social science insights, as with big data studies about other social media.

This research focuses on a single object, one which provides a focus of attention at the research front (Schroeder, 2014a, 2014b), which potentially makes for a highly integrated area of study. What we have seen, however, is that this research remains highly fragmented: How findings about Wikipedia relate to each other (in terms of building on each other, but also incorporating findings and insights from studies in other disciplines), how they are related to findings about new media, and how they could be related to other areas of social science research, is for the most part unrecognized in these disciplinarily disparate studies. This is to be expected in a very new field of research, yet it is also a limitation, partly due to the fact that Wikipedia data are taken as a given starting point, without considering how Wikipedia as a phenomenon fits among other, similar objects and findings in the social sciences. Indeed, this problem is highlighted by the fact that in this paper, we have labelled Wikipedia as an example of new social media,

which is a common way to categorize it (van Dijk, 2013, pp. 132–153 also calls them ‘connective media’). There are similarities insofar as other social media (Twitter or Facebook or LinkedIn) are similarly being studied using big data approaches and where there is similarly not yet a close embedding within the social sciences (though connections can be made more readily: for example, relating Twitter to the study of news dissemination, or Facebook to social network analysis, and the like; see Golder & Macy, 2014). In any event, it will be useful to pursue the fragmentation and greater or lesser of integration of Wikipedia further in order to provide an outlook on the shape and future directions of this research.

One topic concerning Wikipedia where studies could have a common focus is disagreements about content: Can contributors agree about controversial topics, given the neutral point of view (Reagle, 2010) of an encyclopaedia, and can topics therefore be finalized? The Apic et al. study (2011) uses ‘disputes’, while Yasseri et al. (2012) use the label ‘conflict’, and although they use different data (‘reverts’ as against the Wikipedia dispute index), there are clearly links between the two studies, such as that disputes might be focused on similar geographical places. However, while the findings show that Wikipedia conflicts shed light on contentiousness and disputes are concentrated on a few topics, it is not clear what to deduce from these quantitative analyses generally: the correlation that Apic et al. highlight is not surprising, and apart from the practical lessons of the Yasseri et al. (2012) study that the disproportionate effort of highly conflictual areas should be avoided, perhaps what is more interesting about Wikipedia is that it is an example of a great deal of agreement in the production of knowledge by a group of self-organized people, as Reagle (2010) points out. Or again, following a particular disputed entry in Wikipedia in detail from a qualitative perspective might shed light on the micro-level dynamics and reveal motivations that can be linked to macro-level processes: How stable is a Wikipedia entry for a contentious topic, and how is this stability accomplished in practice, which can be studied, for example, for a single, highly contentious entry about ‘neoliberalism’ (see Tatum & LaFrance, 2009)? Yet this would also require situating Wikipedia among other collaborative knowledge-producing efforts or the contentious aspects of knowledge more broadly.

At the same time, Wikipedia cannot, as is done in a number of studies discussed, be taken as a model of social relations generally. Yasseri et al. (2012, p. 6) point out that Wikipedia ‘is one of the few human societies that [sic] the history of all actions of its members are recorded and accessible’. They say ‘few’, but this statement also applies to Facebook, Twitter, virtual worlds and online games, and more. However, it is also worth pointing out that these are not ‘human societies’, but rather specialized groups: collaboration in creating a joint knowledge repository (Wikipedia and similar efforts), or interactions among groups of connected people (Facebook), and the like. ‘Societies’ are defined by social scientists in various ways, but they comprise many multi-faceted interactions. Wikipedia and similar sources of big data research are not ‘societies’ but rather microcosms, smaller subsets of social interactions that illustrate in the first instance only a special-purpose set of interactions among a limited group of people.

A different kind of fragmentation or non-integration can be illustrated by the Zhang and Zhu (2011) and West et al. (2012) studies in terms of where they were published and how they address different audiences. The two studies both have findings about collaboration, but they are published in venues that are disciplinarily so remote – an economics journal and a WikiSym Proceedings – that the authors and other researchers working on the topic are unlikely to come across each other. Even assuming that researchers interested in collaboration do come across both studies, however, the findings are addressed to different audiences: economists with interests in formal models in the one case, and systems developers who are interested in implications for design in the other. Yet not only do both studies have interesting insights about collaboration, which could be compared, but they could also be brought to bear on each other to guide future research: for example, in relation to different populations of Wikipedia contributors, where connections

could potentially be made between Chinese contributors (Zhang and Zhu) and contributors who specialize in editing content about China (a subgroup in the West et al. study). Another overlap is the practical aim of improving Wikipedia collaboration and contributions: Can understanding (economic) incentives be combined with understanding the degree to which contributors specialize in particular areas? Again, this connection could be made between the two studies (though of course these are only examples: many more connections could be instanced), but it is likely to be overlooked where there is a disciplinary division of labour whereby quantitative analysis is pursued in relation to highly specialized questions based largely on data availability. Perhaps it is not surprising that publications are often bound to disciplines, in view of academic reward systems. Nevertheless, in an interdisciplinary area such as big data Wikipedia research, this siloing also creates barriers.

Similar points could be made about the publication venues and audiences of the other studies described here. Whether there is in future more potential for ‘travel’ across different disciplinary communities or not remains to be seen. Yet there are further limitations to how social science disciplines inform each other in this case: one example is the constraint that, in geography, the data often come from a subset of Wikipedia entries – edits from geocoded contributors, place names, geocoded entries and the like. Large parts of Wikipedia are related to place, and so there is potential scope for much cross-fertilization; for example, including Yasseri et al., Bao et al., Zhang and Zhu, Graham, and others. There would be even more scope if place and the data related to it could be seen in the context of what is known about how peoples’ knowledge of different places relates to their awareness about these places, which is, after all, what Wikipedia is aimed at.

Another way that research could become more interrelated is if studies build on each other, or the ‘mutual dependence’ (Whitley, 2000) of research that we discussed in the section about previous research. An exception among the studies discussed here is where they try to outdo each other’s predictions: we have mentioned Mestyán et al.’s (2013) Wikipedia prediction competing with versus Asur and Humberman’s (2010) Twitter prediction, but equally Moat et al.’s (2013) Wikipedia prediction could be compared as competing against Bollen et al.’s (2011) prediction of stock market movements based on Twitter analysis. Many social scientists would reject the idea that social science should aspire to predictive power. Yet the studies presented here that do so are associated with the more quantitative disciplines (physics, economics) and aimed at very narrow goals. Indeed, big data, insofar as it is a growing area of social science research, raises the question of whether there is a new turn towards or emphasis quantitative and scientific research in the social sciences, with all that this implies for the richness or otherwise of these and other social science traditions. The turn to quantification and scientificity is bound to remain contested, and perhaps it can simply be noted that among the Wikipedia studies discussed here, in some cases, predictions derived from Wikipedia forecast real-world trends (Mestyán et al., 2013; Moat et al., 2013), while other studies argue that Wikipedia distorts phenomena in the world-at-large (Graham, 2011). In view of the fact that these are also studies with greater and lesser specificity or generality and apply to different subsets of data, this is a topic where synthetic analysis has the potential to provide insightful ways forward.

We have seen a high degree of interdisciplinarity in Wikipedia research, research which can be found at the intersections of many disciplines (we discussed interdisciplinarity in the section on previous research). Wikipedia research is unlikely to become a subdiscipline in its own right, though it has some of the trappings of being organized as ‘Wikipedia research’ (as evidenced by the Wikipedia conferences and the bibliographies of Wikipedia research). As we have seen, there are many potential areas of cross-fertilization between these studies, but much of this cross-fertilization is unlikely to take place because the specialized nature of the research, the findings, and how they are disseminated in publications. Specialized studies and a focus on readily available data also leave out key topics which could form bridges to other topics and social

science questions, for example, if more was known about the Wikipedia editors and readers? Or, what is the landscape of online encyclopaedia production world-wide, including not just into different language versions but also alternatives that are used (Baidu Baike)? How important is this source of knowledge, not just in search engine rankings that have been mentioned but more generally (the question is starting to be asked in surveys, for example, 7% of Swedish internet users use Wikipedia daily and 66% do so occasionally [Findahl, 2012, p. 15])? If more was known about Wikipedia as an object within a larger context, then we would also be able to better situate the significance of specific findings about this data object. This limitation also applies to big data research in other areas of the social sciences (Schroeder, 2014a), but it is important to point out that it applies in an area where there are no restrictions on data access, and where there are studies that use many techniques and cover many topics apart from those that lend themselves to computationally and data-intensive approaches.

All of the papers that we have examined have been carried out by researchers with considerable computer science skills, and, as Golder and Macy (2014) point out, these skills are only just beginning to be recognized as important for social scientists. Another feature of the studies discussed is that although we have focused on research relevant to social science, most of the papers we discuss here have not been published in social science journals. The papers also evince a variety of aims, from contributions to systems design, to forecasting for business purposes, and studies that use Wikipedia to illuminate traditional social science questions like knowledge, power and geography. Big data research here, as elsewhere, tends to be multidisciplinary, while non-big data research on Wikipedia (as with other online media) is more bound by traditional disciplines.

It is also plausible that Wikipedia research using big data is more oriented to questions that can be answered using the available data rather than being guided by traditional questions within disciplines. This point also applies to other big data studies of new media (Schroeder, 2014b). Other big data research on new media, on the other hand, is constrained by the fact that the data are either proprietary (as in the case of Facebook) so that access requires a connection to a private company, or for example in the case of Twitter only limited samples are available for free. Wikipedia (again) is an interesting case of big data research since the data are freely and completely available, and findings can therefore be replicated and built upon. One prognosis that can thus be ventured is that Wikipedia is likely to continue to be exploited for its data, and more so than other sources, at least by researchers who do not have access to commercial data sources (thus also providing an important exception to the argument made by Savage & Burrows, 2007, that research in the commercial sector is increasingly outpacing academic social science). A point that follows directly is the imbalance created here: regardless of the significance of Wikipedia in society in comparison to other online media, media with proprietary data will be studied less by academic researchers. In this respect, it is interesting to consider that this imbalance did not apply in the same way to other social science research where data gathering faced different kinds of challenges (e.g. conducting large-scale surveys commercially or via grant funding).

Finally, a notable feature of most of the studies examined here is that they all have a practical aim; that is, they could be used to improve Wikipedia or similar socio-technical artefacts, or to correlate Wikipedia characteristics with problems to be addressed in the world-at-large. At the same time, they are all oriented to illuminating social science topics: contributions to the public good, the characteristics of online contributors, how online content maps to offline realities, and conflict in online discussion or deliberation. Yet the studies discussed do not build on each other (again, apart from the specific aim of outdoing each other in terms of predictive power), in terms of common methods, or findings that advance on each other, or of a community where researchers engage with each other – in respect to these three kinds of integration, we see only limited evidence. Indeed, despite having a common object, this research remains disparate, a

pluralism in the social sciences which has advantages and drawbacks. This disparateness would at least be partly overcome if there was a shared understanding of the significance of Wikipedia (and cognate sources of knowledge and information) in society, which would lend a shared focus to the research beyond the common use of a source of data.

Thus, to move away from the language of integration and fragmentation for a moment, there are centrifugal and centripetal forces in social science research about Wikipedia using big data. What is perhaps more notable is the rapid growth that Wikipedia research shares with other big data social science research. And while this is a thriving area, there are also, as we have seen, limitations, some of which could be overcome by greater awareness of other research in this field, as well as awareness of research in neighbouring disciplines and other work on Wikipedia and the wider significance of this and other social media. What we see is that new territory is being opened up for social science through a new source of data about a social phenomenon that is, moreover, readily analysable on a large scale. What we also find is that many new questions are raised and limitations encountered: What can be learned from these online ‘microcosms’? Further research, but also synthetic analysis across disciplinary boundaries, is needed if these limitations in research about a particularly rich object of data (or objects, if we consider particular language versions or other online encyclopaedias) are to be overcome, and if we should simultaneously understand the role of these microcosms and their significance in society.

### Disclosure statement

No potential conflict of interest was reported by the author.

### Funding

This work was supported by the Sloan Foundation under the grant ‘Accessing and Using Big Data to Advance Social Science Knowledge’.

### Notes on contributors

Ralph Schroeder is Professor at the Oxford Internet Institute at the University of Oxford. He is director of its Master’s degree in ‘Social Science of the Internet’. His books include ‘Rethinking Science, Technology and Social Change’ (Stanford University Press 2007), ‘Being There Together: Social Interaction in Virtual Environments’ (Oxford University Press 2010) and ‘An Age of Limits: Social Theory for the 21st Century’ (Palgrave Macmillan 2013). Before coming to Oxford, he was Professor at Chalmers University in Gothenburg, Sweden. His current research is focused on the digital transformations of research.

Linnet Taylor is a Marie Curie research fellow in the University of Amsterdam’s International Development faculty, with the Governance and Inclusive Development group. Her research focuses on the use of new types of digital data in research and policymaking around issues of development, urban planning and mobility. Previously she was a researcher at the Oxford Internet Institute on the project ‘Accessing and Using Big Data to Advance Social Science Knowledge’. Linnet studied a DPhil in International Development at the Institute of Development Studies, University of Sussex where she was also part of the Sussex Centre for Migration Research.

### References

- Akdag Salah, A., Gao, C., Scharnhorst, A., & Suchecki, K. (2011). *Design vs. emergence, visualization of knowledge orders*. Retrieved from [http://scimaps.org/maps/map/design\\_vs\\_emergence\\_127/](http://scimaps.org/maps/map/design_vs_emergence_127/)
- Apic, G., Betts, M. J., & Russell, R. B. (2011). Content disputes in Wikipedia reflect geopolitical instability. *PLoS ONE*, 6(6), e20902. doi:10.1371/journal.pone.0020902
- Asur, S., & Huberman, B. (2010). *Predicting the future with social media*. arXiv:1003.5699. Retrieved from <http://arxiv.org/abs/1003.5699>

- Bao, P., Hecht, B., Carton, S., Quaderi, M., Horn, M., & Gergle, D. (2012, May). Omnipedia: Bridging the Wikipedia language gap. In *Proceedings of the 2012 ACM annual conference on human factors in computing systems* (pp. 1075–1084). New York: ACM Press.
- Bar-Ilan, J., & Aharoni, N. (2014, June 23–26). Twelve years of Wikipedia research. In *Proceedings of WebSci'14*, Bloomington, IN. Retrieved from <http://dx.doi.org/10.1145/2615569.2615643>
- Becher, T., & Trowler, P. (2001). *Academic tribes and territories: Intellectual inquiry and the culture of disciplines* (2nd ed.). Milton Keynes: Open University Press.
- Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1), 1–8.
- Collins, R. (1975). *Conflict sociology: Toward and explanatory science*. New York: Academic Press.
- Findahl, O. (2012). *Swedes and the Internet*. Retrieved October 20, 2013, from [http://worldinternetproject.net/files/Published/oldis/120\\_engsoi2012\\_web\\_121214.pdf](http://worldinternetproject.net/files/Published/oldis/120_engsoi2012_web_121214.pdf)
- Freeman, L. (2004). *The development of social network analysis*. Vancouver: Empirical Press.
- Gartner. (2011). *Gartner says solving 'big data' challenge involves more than just managing volumes of data*. Retrieved August 4, 2014, from <http://web.archive.org/web/20110710043533/http://www.gartner.com/it/page.jsp?id=1731916>
- Golder, S., & Macy, M. (2014). Digital footprints: Opportunities and challenges for online social research. *Annual Review of Sociology*, 40, 129–152.
- González-Bailón, S., Wang, N., Rivero, A., Borge-Holthoefer, J., & Moreno, Y. (2014). Assessing the bias in samples of large online networks. *Social Networks*, 38, 16–27.
- Graham, M. (2011). Wiki space: Palimpsests and the politics of exclusion. In G. Lovink & N. Tkacz (Eds.), *Critical point of view: A Wikipedia reader* (pp. 269–282). Amsterdam: Institute of Network Cultures.
- Hindman, M. (2009). *The myth of digital democracy*. Princeton: Princeton University Press.
- Klein, J. T. (1996). *Crossing boundaries: Knowledge, disciplinarity, interdisciplinarity*. Charlottesville: University Press of Virginia.
- Lazer, D., Kennedy, R., King, G., & Vespignani, A. (2014). The parable of Google flu: Traps in big data analysis. *Science*, 343(6176), 1203–1205.
- Liao, H.-T. (2009). Conflict and consensus in the Chinese version of Wikipedia. *IEEE Technology and Society Magazine*, 28(2), 49–56.
- Mestyán, M., Yasseri, T., & Kertész, J. (2013). Early prediction of movie box office success based on Wikipedia activity big data. *PLoS ONE*, 8(8), e71226.
- Moat, H. S., Curme, C., Avakian, A., Kenett, D. Y., Stanley, H. E., & Preis, T. (2013). Quantifying Wikipedia usage patterns before stock market moves. *Scientific Reports*, 3, Article number 1801.
- Norris, P., & Inglehart, R. (2009). *Cosmopolitan communication: Cultural diversity in a globalized world*. Cambridge: Cambridge University Press.
- Okoli, C., Mehdi, M., Mesgari, M., Nielsen, F., & Lanamäki, A. (2012). *The people's encyclopedia under the gaze of the sages: A systematic review of scholarly research on Wikipedia* (available at SSRN).
- Park, T. K. (2011). The visibility of Wikipedia in scholarly publications. *First Monday*, 16(8–1). doi:10.5210/fm.v16i8.3492
- Puschmann, C., & Burgess, J. (2013). The politics of Twitter data. In K. Weller, A. Bruns, J. Burgess, M. Mahrt, & C. Puschmann (Eds.), *Twitter and society* (pp. 43–54). Oxford: Peter Lang.
- Reagle, J. M. (2010). *Good faith collaboration: The culture of Wikipedia*. Cambridge, MA: MIT Press.
- Rule, J. (1997). *Theory and progress in social science*. Cambridge: Cambridge University Press.
- Savage, M., & Burrows, R. (2007). The coming crisis of empirical sociology. *Sociology*, 41(5), 885–899.
- Schroeder, R. (2014a). Big data: Towards a more scientific social science and humanities? In M. Graham & W. H. Dutton (Eds.), *Society and the Internet* (pp. 164–76). Oxford: Oxford University Press.
- Schroeder, R. (2014b). Big data and the brave new world of social media research. *Big Data and Society*, July–December, 1–11.
- Serrano, M. Á., Flammini, A., & Menczer, F. (2009). Modeling statistical properties of written text. *PLoS ONE*, 4(4), e5372.
- Tatum, C., & LaFrance, M. (2009). Wikipedia as a distributed knowledge laboratory: The case of neoliberalism. In N. Jankowski (Ed.), *e-Research: Transformation in scholarly practice* (pp. 310–27). Abingdon: Routledge.
- Van Dijk, J. (2013). *The culture of connectivity: A critical history of social media*. Oxford: Oxford University Press.
- West, R., Weber, I., & Castillo, C. (2012, August 27–29). Drawing a data-driven portrait of Wikipedia Editors. In *Proceedings of WikiSym'12*, Linz, Austria.

- Whitley, R. (2000). *The intellectual and social organization of the sciences* (2nd ed.). Oxford: Oxford University Press.
- Yasseri, T., Sumi, R., Rung, A., Kornai, A., & Kertész, J. (2012). Dynamics of conflicts in Wikipedia. *PLoS ONE*, 7(6), e38869.
- Zhang, X., & Zhu, F. (2011). Group size and incentives to contribute: A natural experiment at Chinese Wikipedia. *American Economic Review*, 101, 1601–1615.

#### Interviews

- Gergle, D. (2013). Associate Professor of Communication Studies, Northwestern University, Interviewed 28.8.2013.
- Graham, M. (2013). Senior Research Fellow, Oxford Internet Institute, University of Oxford, Interviewed 29.5.2013.
- Moat, S. (2013). Researcher, Behavioural Science, Warwick University Business School, Interviewed 16.5.2013.
- West, R. (2013). Ph.D. Candidate, Computer Science, Stanford University, Interviewed 28.5.2013.
- Yasseri, T. (2013). Research Officer, Oxford Internet Institute, Interviewed 13.1.2013.
- Zhang, M. (2013). Professor, Hong Kong University of Science and Technology Business School, Interviewed 10.5.2013.