JOURNAL OF DIGITAL SOCIAL RESEARCH

WWW.JDSR.IO ISSN: 2003-1998

VOL. 5, NO. 1, 2023, 90–108

CENTRALISING QUALITATIVE RESEARCH IN BIG DATA METHODS THROUGH ALGORITHMIC ETHNOGRAPHY

Naomi Barnes[a] and Sam Hames[b]

ABSTRACT

Responding to the challenge for qualitative researchers to claim a central place in conversations about big data, analytics, datafication, data mining and the role of algorithms, this article describes a mixed-method research partnership focused on algorithmic ethnography. In the debates about the opacity of online algorithms, qualitative researchers typically advocate for access to code. This standard discourse centralises the technical aspects of big data and networked ethnographies. Instead, this article outlines a research methodology that analyses algorithmic discourses by working alongside the technical expertise of data scientists and utilizes the affordability of big data methods to do qualitative work. The potential for qualitative research skills to investigate the underlying technical processes that frame online social interactions is proposed as a way to place how people understand the world at the centre of big data research.

Keywords: big data; algorithms; ethnography; algorithmic discourses; mixed methods.

[a] Queensland University of Technology, Australia
[b] University of Queensland, Australia.

1 INTRODUCTION

Big data and their subsequent networks have historically been associated with qualitative research (Bancroft et al., 2014); however, the introduction of Internet-mediated big data has meant the methodologies have been taken over by quantitative, statistics-oriented, technology-enhanced methodologies (Sarkar, 2021). As such, qualitative research is faced with the challenge of re-centralising itself in this revolution of how "developing technology *and* the way people interface with that technology in their social contexts *and* what that interface enables or leads to in those contexts" (Cheek, 2021, p. 124). To rise to this challenge, the following article describes a research approach developed through a mixed-method partnership where the goal was to centralise qualitative research using contemporary big data technologies. In particular, we[1] focused on developing a technique to "measure the implicit meanings that occur in-between strings of words" (Mills, 2018, p. 599) by analysing algorithmic discourses.

This article has two purposes. Firstly, to explain a qualitative digital methodological approach that emerged through discussions between an education sociologist and a data scientist. By working together to understand the language of each other's fields, we outline below the first steps we have taken in re-centring big data towards qualitative, interpretive analysis. Secondly, we piloted what we have developed with a field of study— political and policy discourses associated with education. In particular, we look at the digital rhetoric that formed around a literacy policy deliberation process in Australia. Unlike an empirical research report where the methodology proceeds the findings, in this paper we weave the story of our field of study into the methodological explanation for illustrative purposes. You can read about the empirical study and findings in more detail elsewhere (Barnes, 2021). Before proceeding we explain the interpretive foundation to our study of the effects of algorithms, digital rhetoric.

2 DIGITAL RHETORIC

It has become increasingly important to consider the role of algorithms in the discourses influencing politics. A decade of research into the "black boxed" effects of algorithms (Pasquale, 2015) have led online communications scholars to argue that the opacity of algorithms are a key sociotechnical problem that requires transparency and regulation. Social algorithm researchers who have studied their effects have insisted that algorithms be opened up (Eubanks, 2018; O'Neil, 2016; Pasquale, 2020) for true social critique. As such, the direction of research has moved further away from the sociological and humanities approaches that have traditionally dominated an understanding of social issues and interactions, towards

[1] While each of us brought very different but complementary skills to the project, for ease of explanation, we use the pronoun "we" throughout, rather than specifically indicating how tasks were split.

technical approaches that required specialised computing skills (Mills, 2018). However, as algorithms are vastly differentiated, constantly evolving through machine learning, and intersecting across multiple platforms, with "long chains of actors, technologies and meanings" (Christin, 2020b, p. 897), it seemed reasonable to us to explore how qualitative approaches, well versed in collecting and analysing dynamic data, could be combined with specialist technical expertise to better understand the effect of algorithms.

To develop this methodology, we drew on the field of digital rhetoric. According to Eyman (2015) the term *digital rhetoric* is perhaps most simply defined as 'the application of rhetorical theory (as analytic method or heuristic for production) to digital texts and performances' (p.45). Some approaches are closely related to traditional rhetoric and composition studies including, how people use strategies to analyse digital texts, identifying how digital texts are constructed in order to produce more effective communication objects, and how people create digital authorship identities and audiences. These are all important to our project but are well established in qualitative studies and the leap from analysing terrestrial texts, audience and authorship to digital versions is not what we are concerned with in this article. Instead we are interested in explaining a methodology for considering the 'rhetorical function of networks' (Eyman, 2015, p. 45) by concentrating on a key mechanism for holding those networks together — algorithms.

When a political entity wishes to influence, then knowledge of how algorithms deliver that information becomes a rhetorical tool of influence. Education policy and politics is the field we chose to pilot our algorithmic rhetorical analysis. As education is a political field, it is then important to consider the role of algorithmic discourses in how education policy is developed and enacted. Close attention has been placed on algorithms in educational research in the fields that would be expected, such as the increasing reliance on machine learning and automated decision making in using data to construct educational futures (see for example Webb et al., 2020) and the calculation of A-Level results in the UK and Ireland (Kelly, 2021). As educational policy is developed in the public sphere, we contend that the discursive effect of algorithms through public-facing websites and applications, such as social media, are also important to consider. We hypothesised that educational rhetoric could be identified through manipulation of big data networking capabilities and forensic examination of the education and policy actors' rhetoric but also the algorithmic mechanisms that connected those policy actors. Unlocking algorithms' effect on educational political rhetoric is a broader project our collaboration is working towards and how we illustrate the methodological approach we explain in this article.

2.1 Digital rhetoric and algorithmic ethnography

Algorithmic ethnography is the approach we took to understanding how algorithms hold networks together. Our approach involved analysing what information was fed into the so-called black box of the Internet and theorizing, using algorithmic metaphors, the educational discourses which emerge. Algorithmic ethnography has recently been defined as the "ethnographic study of the computational systems enabling and shaping online interactions" (Christin, 2020a, p. 109). Situated within the realm of online communication, Christin's research approach has the potential to be expanded into a sociological systems approach that considers the rhetorical role of algorithms in shaping online and offline interactions. By combining two methodological elements of Christin's (2020a, 2020b) proposal to enrol algorithms in established digital ethnographic approaches, and digital rhetoric (Losh, 2009), this paper outlines a methodology for analysing effects of algorithms in online educational discourse.

According to Christin (2020a), "adopting the lens of algorithmic ethnography entails paying close attention to the role of algorithms in structuring the back and front end of the digital platforms that increasingly mediate digital exchanges" (p. 109). While researchers like Pasquale (2015) advocate for making transparent the mechanisms which deliver information, algorithmic ethnography provides a way to theorise what is happening behind the forward-facing text to boost or block how that text is distributed around the Internet. Without the fine detail in the code which platforms keep black-boxed, it is still possible to hypothesise about the digital rhetoric of the algorithms. There are a finite number of categories of algorithm which means there are a finite number of potential interpretations for how information is being distributed online. We use Christian and Griffiths' (2017) popular explanation[2] (drawing non-exhaustively from many subfields) as a starting framework for our algorithmic thinking, with a particular focus on sorting and caching. We use these algorithms as building blocks to help us bridge the space between qualitative inquiry and the quantitative worldview underpinning the more complex algorithmic assemblages deployed in online systems.

Christin (2020a) outlines three necessary steps for designing a qualitative algorithmic ethnography: the type of data collected, the role of algorithms in sorting and organizing the data, and the effects of online metrics on how people interact online. To develop hypotheses about the role of algorithms we revisited a recent education policy deliberation study (Barnes, 2021). The research question we were interested in was: *What role can we see algorithms playing in affecting how online users influence online literacy policy deliberation?* Considering the so-called Reading Wars (see for example Pearson, 2004), we determined it was a useful place

[2] We have selected this popular guide to algorithms, rather than an academic text as 1) the book gives a clear explanation of how each algorithmic category works with accessible scenarios. We have not critically engaged with the work because 2) this methodological approach is just beginning, and we would hope other researchers will find a way into discovering the effects of the algorithms we did not note. Furthermore, 3) there is not enough room in this paper to effectively describe all the categories, so a popular explanatory text is a good place to direct anyone interested in pursuing this methodological approach. We would hope that more critical sociological work could emerge from this starting point.

to begin looking closely at the digital rhetoric. In brief, the Reading Wars are the academic, political, and public debate about the best way to teach children reading. Today the Science of Reading (Castles et al., 2018) has been determined by educational policymakers to be the best evidence-based approach. When we conducted this research, the Science of Reading had not yet gained policy status and the debate we analysed was part of the political process by which advocates of the program advocated for the program. The online engagement we captured through algorithmic ethnographic methods were between what we will refer to as Science of Reading (SOR) advocates versus socio-cultural literacy advocates (SCL). Very basically, SCL practitioners advocate for reading to be taught in the context of books, while SOR practitioners advocate for reading to initially be taught out of context through repetition using objects like flashcards and drills. This choice was also made because the Reading Wars have existed before the Internet was invented, meaning that the historical manifestations of the debate could be used to make sense of the debate as it occurred online.

Our method uses three analytical phases and one theoretical phase – we will start with a high-level overview of these phases to frame the detailed case study that follows. Of course, although we present these phases as a linear ordering for the convenience of the reader this is only an approximation of the actual reality of conducting such an analysis.

2.1.1 Phase 1: Selection of data

First it is necessary to determine which data is to be included in the study – this is also necessarily the first point of qualitative interpretation of the data. While this phase may initially begin with simple computational filtering, such as selection of documents containing a keyword, the selection phase would iteratively move towards more and more qualitative decision making about relevance of individual data points. Additionally, it is at this stage that the unit of analysis (that is - what is a data point?) is also chosen.

2.1.2 Phase 2: Sorting and searching through the data

Having defined what data is to be included in the analysis, the next phase delved further into the qualitative interpretation of the data. At this point we used the logic of an algorithmic cache to conceptualise the data coding procedure in a way that is consistent with the computational requirements of the next phase.

2.1.3 Phase 3: Conceptualising the social network

The third analytical phase uses network visualisations to map out the contours of the searched and sorted data. This phase brought together the different concepts identified by searching and sorting through the constructed caches of phase 2, along with the data selected as part of phase 1 into a single unified view.

2.1.4 Phase 4: Hypothesising the algorithmic discourses

Connecting the conceptualisations of the network to how the algorithm affected online rhetoric was determined through abductive reasoning. This phase did not intend to contribute to existing scholarship through deductive or inductive outcomes, rather develop hypotheses through experimenting with visualisations and various ways of analysing the data. The hypotheses form the basis of research questions for future inductive and deductive analysis framed by a relevant sociological theory.

2.2 Phase 1: Selection of data

The selection of data involved a collaboration between the authors: our first author being the educational sociologist intent on understanding online educational policy advocacy; our second author being a research data scientist. The collaboration began with a feasibility conversation, trying to find the middle ground of what our first author wanted to research and what our second author was able to extract from the available databases. This process comprised of discussions about research questions, how the extraction process worked, and what was available for extraction. At the time of data extraction, the Australian Twitter database was the most comprehensive social media database available for use. Today other databases are also possible; however, the database had multiple holes in time that data were not archived, so a phenomenon needed to be selected that aligned with the available timeframes. After some initial searches, the literacy policy debate, that became known as the 2018 #PhonicsDebate, was identified as a viable study for two reasons. First, it was selected because audience members were encouraged to tweet using the hashtag #PhonicsDebate during a live debate on YouTube meaning there would be a lot of tweets within a short timeframe. Secondly, it became a multiplatform (Twitter, Facebook, YouTube, blogs, podcasts, petitions, static websites, in person debates) event requiring qualitative skills to connect them all together. Quantitative data collection can only connect between platforms via similarity of key words. Meaning needs interpretive skills that software cannot yet do. This choice meant that qualitative work was centralised.

The key to centralising qualitative research in this research approach was that although this initial starting point is still a quantitative content-based selection, it was conducted in the context of a quantitative/qualitative collaboration with clear parameters and mutual discussion of goals. This created a solid starting point for further qualitative refinement and exploration outside of those parameters.

It was also at this stage that we decided on the initial ethical framework for approaching this data. In doing so we considered multiple aspects, including the public (and publicised) nature of the debate, the technically public nature of the tweets, the nature of the expected audience authoring those tweets, and the difficulty of paraphrasing or other types of anonymisation when reporting on social media data. Weighing up these considerations, we decided that when reporting on this material would not identify or use material on individual participants, not even in paraphrased or "anonymised" forms – all reporting would focus on aggregated or abstracted views of data, and high-level descriptions of content that cannot be linked to individual accounts. This study was also approved by the QUT Human Research Ethics Committee.

2.2.1 Extraction

A spreadsheet (mocked up[3] in Table 1) was provided, drawn from QUT's Digital Observatory's longitudinal Australian Twittersphere database using #phonicsdebate, #phonicscheck, associated key words "teaching +reading", and "phonics". These were collected from the two-week period surrounding either side of 31st July 2018 when a debate about the value of universal synthetic phonics was live streamed on YouTube. The back and forth between us continued once the phenomenon and timeframe was selected, refining the inclusion and exclusion criteria. Hames engaged in computational activities like pruning keywords that were overly broad or were capturing unrelated tweets.

Table 1

Mockup of the initial extraction of tweet data, including initial fields considered, and the structural data describing the tweets place in the Twitter conversation.

tweet_id	user	text	created_at	reply	retweet	quote
XXXXXX177 1462070000	@usera	Phonics is the best way to teach reading! #phonicsdebate	26/07/2018 7:23	0	0	0
XXXXXX485 1394900000	@userb	Phonics should not be separate to language #phonicsdebate	26/07/2018 8:15	0	0	0
XXXXXX785 6229600000	@userc	@userf Don't teachers usually do both? #phonicsdebate	26/07/2018 9:46	1	0	0
XXXXXX249 9598490000	@userd	RT @usere After the #phonicsdebate, come over to #AussieEd!	26/07/2018 11:24	0	1	0

[3] As the ethical clearance does not allow for the direct quotation of tweets, the tables are illustrations of the EXCEL sheets. The tweet identification numbers have been anonymized to ensure the original tweets are not searchable.

The unit of analysis was the tweet (or the text written by Twitter users to comment on the phenomenon), and the associated Twitter users *handle* (online unique identifier), and associated metadata (tweet id and time tweeted). The tweet was chosen as the unit of analysis because this aligns the qualitative and computational components of this work directly with the fundamental unit of communication on Twitter.

Alternative units of analysis were considered, including user and conversation focused approaches. A user focused approach would consider the unit of analysis to be a user profile and an aggregation of all of their tweets, representing a particular accounts communications relating to the subject of interest. A conversation focused unit of analysis would aggregate users and tweets replying to a specific thread into a single unit, representing a specific exchange relating to the topic of interest - at the time of this case study the Twitter API did not provide the information needed to ensure that complete threads of conversation could be reconstructed, ruling out this approach. Importantly aspects of the user and conversation focused units can be aggregated from the tweet level representation, but the disaggregation to the tweet level is more difficult.

Once the tweets were computationally identified, the associated information attached to each tweet was also extracted, such as who wrote the tweet and what time it was broadcast. The tweets were those from identified Australian Twitter accounts, but the dataset did include international participation if an Australian account retweeted a tweet by an international user. Tweet extraction did not include tweets or profile information from Twitter accounts that were 'protected', that is, those whose tweets and profile details are accessible only to other users approved by the account holder.

2.2.2 Refinement and initial exploration

The refinement and exploration process are the key qualitative approaches within the selection phase. The refinement process involved a manual reading of each of the tweets to ensure that they were all part of the phenomenon under scrutiny. What became evident was that the key words "teaching +reading" extracted multiple tweets that were unrelated to the policy debate or education all together. These tweets were removed from the spreadsheet. The final list of tweets was refined from N=2232 to N=2150 tweets.

The initial exploration also revealed that a number of tweets were repeated in the list because each time a user retweeted a tweet by another user that was already in the dataset, that retweet also became a part of that users' timeline of tweets (effectively, if a tweet was retweeted ten times in the dataset, it would appear as ten distinct rows in the spreadsheet, despite having the same content). A qualitative decision was made to leave them as separate tweets rather than collapse them into one tweet with a retweet count. This work could have been done computationally, but we decided that because the research was a sociology it was important to retain the interactions between users to better understand the social dynamics[4].

The initial exploration also allowed for a qualitative coding of the tweeters' experiences of the phenomenon. By keeping the primary research question in mind (What are the social dynamics of the online literacy policy pipeline?), tweets were roughly coded using qualitative decision making about how each tweeter experienced, understood, comprehended and/or conceptualized the phonics debate. For example, some tweeters were positively on one side or the other, some were diplomatic, and some were using the opportunity to use the hashtag for promotion of other online educational events and hashtag chats (See table 2 as a mockup of this exploration).

Table 2

Explorative coding of tweets. The tweet_id, user and Text columns are as shown in Table 1, and indicate the "as collected" state of the data, the remaining four columns indicate a binary coding (code is present or absent) of the associated tweet. The developed categories identify the position of the tweet with respect to the sides of the Debate.

tweet_id	user	Text	SCL	SOR	Diplomatic	Marketing
XXXXXX1771462070000	@usera	Phonics is the best way to teach reading! #phonicsdebate	0	1	0	0
XXXXXX4851394900000	@userb	Phonics should not be separate to language #phonicsdebate	1	0	0	0
XXXXXX7856229600000	@userc	@userf Don't teachers usually do both? #phonicsdebate	0	0	1	0
XXXXXX2499598490000	@userd	RT @usere After the #phonicsdebate, come over to #AussieEd!	0	0	0	1

This coding prepared the data for further qualitative targeted coding, later in the analysis by making chunks of data sortable. This binary coding scheme is commonly

[4] If a similar study were to be conducted again, the list of "likes" would also have been useful to look at the social dynamics of phonics debate sentiment (which users have liked which tweets only became comprehensibly accessible via the Twitter API as of early 2022 and was not possible at the time of the study).

used for statistical and machine learning representations of data for its efficient numerical representation – for the qualitative use case here it uses the affordances of spreadsheets for data entry, while allowing easier integration of the coded data into other analytical tools.

2.3 Phase 2: Searching and sorting

Searching and sorting are often used interchangeably in regard to the Internet, as "search engines" are actually sorting engines. For example, Google uses personal data to *sort* the millions of websites and deliver suggestions in response to a *search* term. In the second analytical phase, these processes are two separate but overlapping activities aimed at developing rigorous categories and themes that could be used to describe the experiences of the phonics debate.

2.3.1 Key terms, hashtags, and time

While the tweets had been extracted via umbrella terms, like #phonicsdebate, other keywords and phrases were evident in the tweets. These keywords gave clues about the themes and categories that could be extracted from the data. Tableau was used to quantitatively sort through the tweets and identify key terms which indicated which experiences of the phonics debate required deeper qualitative analysis. For example, in analysing the secondary hashtags (see Figure 1), "dyslexia", "science" or "evidence" might be clues for sorting the tweets into themes.

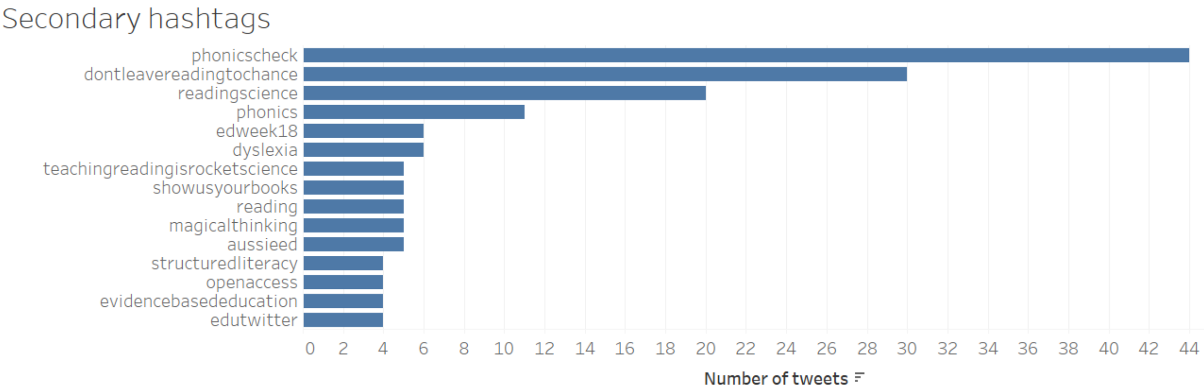

Figure 1. An example of sorting secondary hashtags in the twitter data by frequency to complement the close reading with existing context about the data.

Other key term analysis software, like Leximancer or Excel, could also be used to deal with the quantity of tweets, but Tableau was chosen because it could be used to build a visualisation of the data contemporaneously, rather than after, the coding of the data. As the point of the research was to experiment with what different visualisations large amounts of data could produce, such a tool was more useful for developing hypotheses than the others on offer.

Tableau also allowed for counting the number of tweets broadcast by users over the course of the debate, which allowed judgement about intensity of sentiment, or potential bot engagement[5] with the hashtag. For example, the most visible user (who we later referred to as the hyper-connector below) retweeted multiple tweets (N=476) from the SOR side of the debate. Basic analysis of the tweets from the most visible users showed that parents of children with dyslexia were the most engaged groups of actors in the dataset.

2.3.2 Qualitative caching

The reason we have termed this stage "caching" is because the technical structure of the Internet uses caching to speed up an individual's access to information. In computing, a cache is a copy of some data stored in a temporary (usually ephemeral) location for either easy access or to avoid repeating an intensive operation. A physical analogue would be the sorting trolley in a library before books are returned to the stacks. The caches of information also work to define what a term will come to mean for each Internet user. For example, when you search for a term on Google and explore the initial offerings, Google's search engine will have a cache of websites you clicked through to in case you would like to visit the website again. In other words, Google's search engine begins to build a personalised "meaning" of a search term for each user with each website they visit that uses those terms.

We felt this action is a good description of the initial sorting of themes when coding qualitative research. For example, when a qualitative researcher sorts their sticky notes, decides on codes for NVivo, or, in our case, uses a binary coding system to organize themes, the most recent piece of data added to a pile is the pathway into the clearest definition of the theme. When a qualitative researcher cannot choose a pile or *cache*, they will either manipulate the definition of an already existing cache or start a new cache. Either way, the most recent piece of data is the clearest clue for the defining features of the cache.

Unlike computational caching where the first pieces of information eventually drop out of the cache in a purely mechanical process, *qualitative caching* is an iterative process which occurs throughout all analytical phases. A qualitative researcher will return to all the pieces of data in a cache in order to develop a definition of that theme, discarding or shifting data points between caches, until all relevant pieces of data have a home, and the qualitative researcher is satisfied that all the pieces of data within a cache are representative of the theme. In this framework the qualitative choices an analyst makes are: 1) which items are worth including in the cache for further consideration and 2) in which cache (or caches) should they be included.

[5] No bots were detected in this study after investigating the top tweeters.

The binary coding (1 and 0) in our project was used to develop the caches and iteratively revisit them until satisfied with the themes. A binary code means the Excel spread sheet could be sorted and resorted and checked and rechecked each time a cache meaning shifted. This also meant that the close reading and rigorous coding could be done in chunks, rather than having to read the entire spreadsheet every time a new concept was noted. Considering that a debate is a structured genre where points are made, illustrated, and rebutted, it was straight forward to align tweets with the argumentation. Not all tweets engaged directly with the live debate, and those were put aside for future sorting (see Table 3).

Table 3

Argumentation coding of tweets – this extends on Table 2 to show the additional binary codes used for the argumentation analysis. The additional columns in Table 2 are omitted for brevity.

tweet_id	user	Text	Claim	Warrant	Rebuttal
XXXXXX 1771462070000	@usera	Phonics is the best way to teach reading! #phonicsdebate	1	0	0
XXXXXX 4851394900000	@userb	Phonics should not be separate to language #phonicsdebate	0	1	0
XXXXXX 7856229600000	@userc	@userf Don't teachers usually do both? #phonicsdebate	0	0	1

Each cache organised the vast number of initial tweets into manageable sized groupings for later forensic qualitative analysis.

2.4 Phase 3: Conceptualizing the network

The searching and sorting phase of data analysis provided the foundation for interpreting the social network analysis and conceptualising the digital rhetoric. The #phonicsdebate social network analysis was used to render an initial network visualisation, which helped us analyse that network and subsequently used the analysis to begin to explore variations within the network.

2.4.1 Mapping the network

The Twitter data was also rendered for a social network analysis using the open-source tool Gephi and its Force Atlas 2 algorithm to lay out the network. Interpreting what was computed by this algorithm (see Figure 2) showed that there was a distinct binary between each side of the debate.

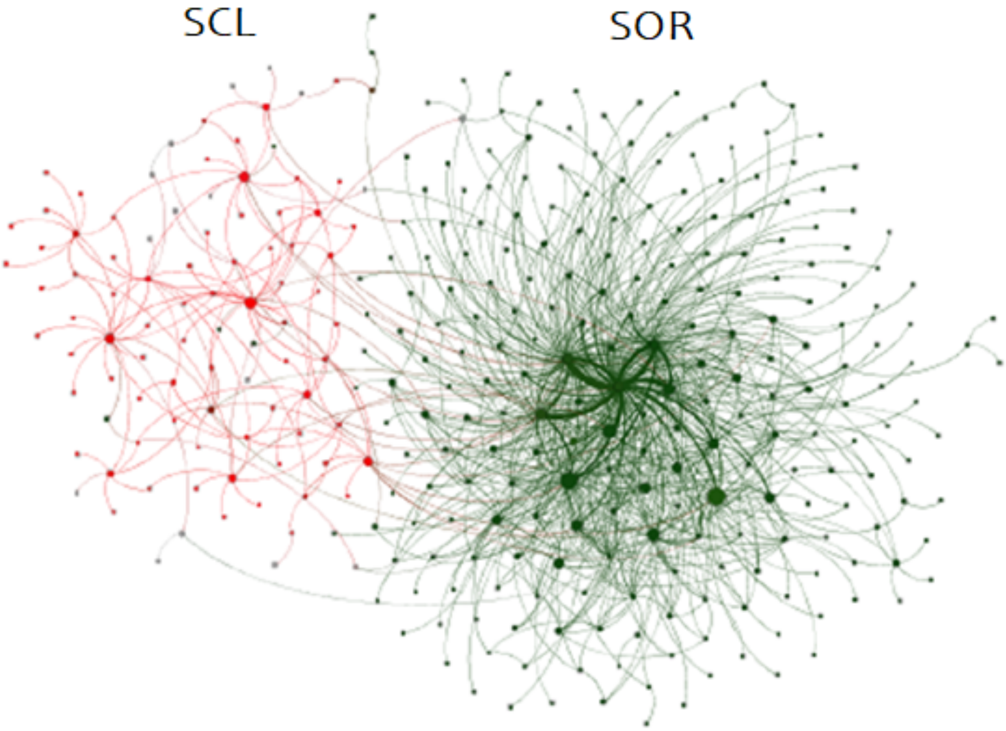

Figure 2. Initial social network analysis. Each node in this network is a Twitter account, the edges between nodes indicate the volume of engagement (retweets and replies) between accounts – thicker edges indicate stronger engagement. Nodes are coloured using the sentiment assigned to each profile to indicate positioning of that account with respect to the debate (red - SCL, green – SOR, grey - diplomats and marketers).

Each node was given a colour according to the side of the debate their tweets indicated they supported. One challenge for this process was that the qualitative analysis was conducted at the granularity of the tweet, but this visualisation was created with the nodes as users. In other words, to colour the nodes, the tweets needed to be read and interpreted because the position of the individual users was in what they wrote in their tweets (refer back to Table 3). Such coding of data would not be possible with quantitative analysis only. This manual annotation was also only possible due to the relatively small number of nodes. This process would not have been infeasible for a larger or more complex dataset and how to address this issue for larger datasets is the focus of our ongoing collaboration.

We noted that the SOR side of the debate had a hyper-connector, or someone who was using the functions of Twitter to distribute a huge amount to tweets from the SOR perspective. This was the same parent who retweeted 476 times. In Figure 2, this hyper-connector is the node in the centre of the SOR network surrounded by a daisy shape. This daisy means they were actively tweeting, retweeting and replying to multiple accounts. Moreso that anyone else but not alone in their activity as is indicated by the thicker connecting lines (or edges) within the network visualisation.

2.4.2 Variation of experience

Much analytical work will stop at the visualization of a dataset, but our process began after all the possible visualisations were developed from the raw data. In other words, the visuals helped us make sense of a massive amount of data to begin hypothesising about the role of algorithms in distributing information on the Internet. The initial social network analysis produced a stripped version of the phenomenon (See Figure 2). It showed that there were two sides of the literacy debate and that they were quite obviously on two ends of a spectrum. However, the social network analysis raised questions which sparked further investigation: These included:

- Who is the hyper-connector and is their activity driving the movement of information around Twitter or are they being assisted in any way by other groups?
- Why are there so few people tweeting about one perspective and so many about the other?
- Is it an accurate representation of online engagement with the debate?

The third question led to a re-coding of the data to create a simple tweak to the network map. In the first and second phases of the analysis, we noticed the tweet texts did not always directly link into the debaters. For various reasons that range from privacy to non-existent Twitter accounts, some debate participants were named but not coded in a way that automatic coding could render them into a social network. Automatic coding needed each member of the network to have a consistent username. As such, we searched the tweets for variants and adjusted the text for consistency (see Table 4) so that the automated systems would identify all the interlocutors.

Table 4

Inconsistent username examples

Hypothetical Twitter account	Possible Variants of Name and title (could include misspelling)
@ProfessorX	Professor X, Prof X, PX, Dr X, Speaker 1, X, Professor Ex, etc
@HarleyQuinn	Professor Quinn, Prof Quinn, Dr Quinn, Harly Quin, Dr Q, Q, HQ, etc

This process transformed the initial network representing only the concrete, platform mediated traces of communication to a more nuanced network that incorporates references to people, not just social media accounts. In the resulting social network one side of the debate all but disappeared from the network map, being absorbed into the other (see Figure 3).

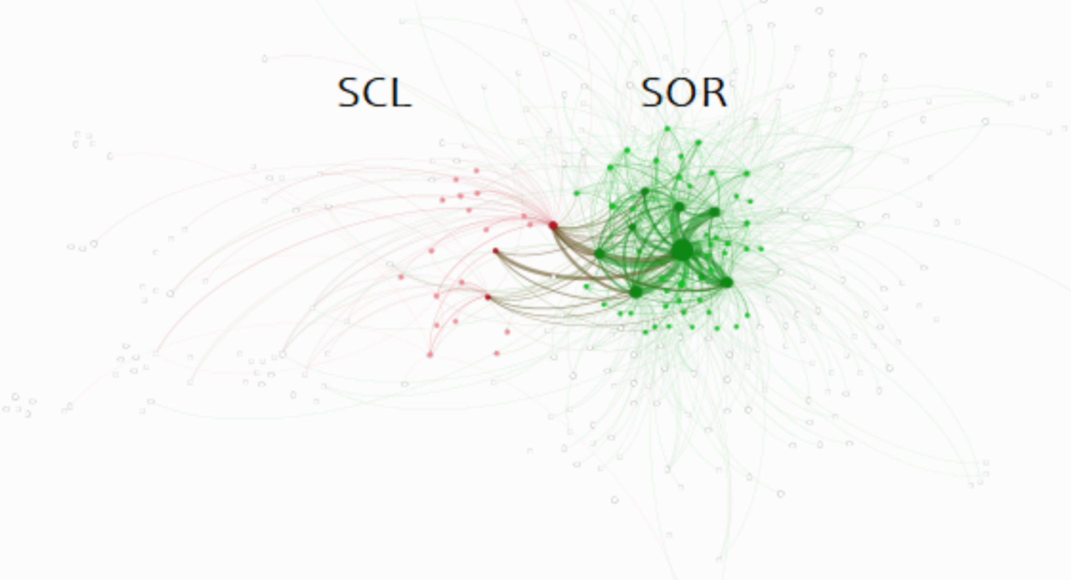

Figure 3. Manipulated social network – As in Figure 2, each node in this network is a Twitter account, the edges between nodes indicate the volume of engagement (retweets and replies) however this time edge weights include mentions of debate participants by name rather than just by Twitter handle. Thicker edges indicate stronger engagement. Nodes are coloured using the sentiment assigned to each profile to indicate positioning of that account with respect to the debate (red - SCL, green -SOR, uncoloured – diplomats and marketers).

This raised new questions about how representative the debate was on Twitter. The reconfiguration of the social network analysis showed that the same accounts were still hyper-connectors, initiating an investigation into why their Twitter behaviour worked so powerfully on the network.

Hypothesising the different experiences of the phonics debate from different interlocutors within the debate provided a clue for where to search for data next. Conceivably, the extra data might have come from any source, including interviews, but as the purpose of the research was to discover how information about a policy moves through the Internet, we chose to stay with data available in Internet archives. However, this article is about centralising qualitative research in big data research and how we worked together to conceptualise how data science and digital sociology can work together. As such we have chosen to not explain this later forensic stage, but it can be read about elsewhere (Barnes, 2021). Essentially, analysis of other online objects led us to return to the Twitter data set and more closely consider the connection between the tweets and the debate. What became apparent was that the hyper-connectors and the bloggers were seeing the phonics debate from a parent's perspective. Knowing that parents were the key hyper-connectors we began to theorise what was rhetorically happening behind the forward-facing text.

Hypothesis building: When a child finds it difficult to read, the response from the parent (or family system) is very different from the response of the school system or the literacy research system. Coming to understand how a parent might use the platform is one research question we drew from the data (as outlined in the next section).

2.5 Phase 4: Hypothesising the algorithmic discourses

In this final phase of the process, we walk through the process of hypothesising using the outputs of the analytical phase and our algorithmic metaphors to arrive at a narrative description and map of the logics at work. We provide this as more of an extended example because this phase is expected to be the most specific and sensitive to the particular research questions of each project.

The initial hypothesising was informed by the algorithmic sorting digital rhetoric. Sorting algorithms, including the commonly used *Mergesort,* alongside caching algorithms, consider the optimal organization of information on the Internet. Christian and Griffiths (2017) provide the recursive logarithmic pattern behind the process and explain why it was so revolutionary by comparing it to how a human might organize their personal library. In terms of Mergesort, a practical and near-optimal way to organize a bookshelf is to invite friends around and divide the books evenly between them. Each friend is asked to organize their own stack, then stacks are combined – because each stack is already sorted combining them is easier than trying to sort everything at once. Using this concept, we can consider algorithmic organisation on the Internet as a collaborative effort – but rather than individual decisions about specific orders, we see individuals curate an organised view of their corner of platforms implicitly via the logic of what they consume and engage with. Platforms, attempting to "personalise" content then mediate the final merging of individual stacks not only based on personal use of the Internet but also the groups of users one might engage with the most — friends. While a human might sort their information into alphabetical order or by genre, theme, or topic, to access it at a later date, an algorithm deployed on the Internet will generally sort information using the logic behind caching. Algorithmically, caching is the most efficient way to find information. Sorting things into categories is less efficient that creating stacks of recently used information. The computer algorithm uses the logic that someone is more likely to want information closely related to the information they just consumed and engaged with. As such the quickest way to organize information is via a logic which directs the human to the last piece of information they used.

Adding the Mergesort logic to caching algorithms, an Internet site or application that sees its major function as searching (whether explicitly user directed or not) will recommend information closely related to the last piece of information extracted and connected to the groups a user is most likely to interact with.

In terms of the phonics debate study, we hypothesised that if a parent were to search Google, Facebook, or another social media site for why their child cannot read, they are very likely to come across other parents experiencing the same issues. Those parents then share their links with each other. These links are most likely to have been established before children attend school and be authored by health professionals like psychologists and speech therapists. Effectively the Mergesort friendship group grows but remains constrained to the presentation of information that aligns with what has already been seen. Eventually a parent may end up on Twitter, which has a strong teacher presence, and encounter the Reading Wars. However, by the time a parent arrives at the Reading Wars, they are more likely to side with the debate that is closer to their Mergesort friends – the psychologists and speech therapists. From a traditional rhetorical point of view, the parents have a logic of affect and comradery attached to their argument because of the Mergesort friends they gathered before coming across an alternative point of view.

3 REFLECTION

While our analysis has shown the more simplified algorithmic discourses (sorting and caching) are at work, we believe that there is enough evidence to justify future algorithmic ethnographies considering the role of algorithms in digital rhetoric. This is particularly important given the increasing role machine learning and artificial intelligence is taking in decision-making. Although this initial work has focused on two foundational classes of algorithms as organisational tools, algorithms as deployed in the real world can be much more complex – sorting and caching are much more likely to be used as building blocks of larger systems. Despite this limitation of this work, we think using these algorithmic tools to inform our analysis and theorising is a useful for ensuring that the qualitative and quantitative components of such work can be mutually grounding rather than separate.

Through our exploration of the possibilities of big and small data network analysis, we have shown that algorithmic ethnography that includes algorithmic digital rhetorical analysis, is a useful way forward in centralising qualitative research in big data methods. It should be noted that the dataset presented here is small enough to work forensically with each node, edge, and network representation. Larger datasets will require more comprehensive qualitative and quantitative efforts: for this study off the shelf tools and simple data formats worked, but "scaling up" to map out a larger phenomenon will require more detailed attention to the modelling of data, the representation of qualitative labelling efforts and how the components are drawn together into the final map. This study is a first step in how mixed-methods teams can work together, with the purpose of understanding each researcher's field enough to solve such problems.

As moderation and connection of the different platforms became too unwieldly and enormous to be conducted via human labour, algorithms became the vehicles responsible for doing the work, becoming increasingly sophisticated by applying machine learning to distribute information and users more quickly and efficiently around the Internet. Algorithms became the lifeblood of the Internet, and increasingly tangled and rooted in how people navigate information about society, including education, and use that knowledge to make decisions. Now digital platforms linked to, but separate from, the political system, are woven into how and why political decision making is performed. In this reality transdisciplinarity becomes essential for understanding the effects of the Internet on policy deliberation and politics. There are too many systems at play for one qualitative researcher to come to understand and those systems are too dynamic for one data scientist to adequately capture.

4 REFERENCES

Bancroft, A., Karels, M., Murray, Ó. M., & Zimpfer, J. (2014). Not Being There: Research at a Distance with Video, Text and Speech. In *Big Data? Qualitative Approaches to Digital Research* (Vol. 13, pp. 137–153). Emerald Group Publishing Limited. https://doi.org/10.1108/S1042-319220140000013009

Barnes, N. (2021). The social life of literacy education: How the 2018 #phonicsdebate is reshaping the field. *The Australian Educational Researcher*. https://doi.org/10.1007/s13384-021-00451-x

Castles, A., Rastle, K., & Nation, K. (2018). Ending the Reading Wars: Reading Acquisition From Novice to Expert. *Psychological Science in the Public Interest*, *19*(1), 5–51. https://doi.org/10.1177/1529100618772271

Cheek, J. (2021). Big Data, Thick Data, Digital Transformation, and the Fourth Industrial Revolution: Why Qualitative Inquiry Is More Relevant than Ever. In *Collaborative Futures in Qualitative Inquiry*. Routledge.

Christian, B., & Griffiths, T. (2017). *Algorithms to Live By: The Computer Science of Human Decisions*. HarperCollins GB.

Christin, A. (2020a). Algorithmic ethnography, during and after COVID-19. *Communication and the Public*, *5*(3–4), 108–111. https://doi.org/10.1177/2057047320959850

Christin, A. (2020b). The ethnographer and the algorithm: Beyond the black box. *Theory and Society*, *49*(5), 897–918. https://doi.org/10.1007/s11186-020-09411-3

Eubanks, V. (2018). *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. St. Martin's Publishing Group.

Eyman, D. (2015). *Digital Rhetoric: Theory, Method, Practice*. University of Michigan Press.

Kelly, A. (2021). A tale of two algorithms: The appeal and repeal of calculated grades systems in England and Ireland in 2020. *British Educational Research Journal*, *47*(3), 725–741. https://doi.org/10.1002/berj.3705

Losh, E. M. (2009). *Virtualpolitik: An electronic history of government media-making in a time of war, scandal, disaster, miscommunication, and mistakes*. MIT Press.

Mills, K. A. (2018). What are the threats and potentials of big data for qualitative research? *Qualitative Research*, *18*(6), 591–603. https://doi.org/10.1177/1468794117743465

O'Neil, C. (2016). *Weapons of Math Destruction*. Crown Publishing Group.

Pasquale, F. (2015). *The Black Box Society*. Harvard University Press.

Pasquale, F. (2020). *New Laws of Robotics: Defending Human Expertise in the Age of AI*. Harvard University Press.

Pearson, P. D. (2004). The Reading Wars. *Educational Policy*, *18*(1), 216–252. https://doi.org/10.1177/0895904803260041

Sarkar, S. (2021). Using qualitative approaches in the era of big data: A confessional tale of a behavioral researcher. *Journal of Information Technology Case and Application Research*, *23*(2), 139–144. https://doi.org/10.1080/15228053.2021.1916229

Webb, P. T., Sellar, S., & Gulson, K. N. (2020). Anticipating education: Governing habits, memories and policy-futures. *Learning, Media and Technology*, *45*(3), 284–297. https://doi.org/10.1080/17439884.2020.1686015