# Digital cultural sociology: scale, pattern and feedback loops

Adrian Mackenzie, Richard Mills, Stuart Sharples, Matthew Fuller and Andrew Goffey

## Introduction

Digital sociology focuses on culture as it plays out in the vast, expanding, power-laden and complex textual and media environments that have taken shape in the last few decades. The object of analysis is culture as it is re-written digitally in settings such as cities, shops, museums, clinics, offices or living rooms, and above all, in the transient publics of news, entertainment and social networking environments such as Twitter, Instagram, LiveJournal, YouTube, LinkedIn or Weibo. The methods of digital sociology vary widely, but nearly all involve using or making software to capture, organise, share, filter, search or sort data (hyperlinks, messages, transactions, text, images) (Lupton, 2012). Digital social researchers seek to learn about the coherence, modes of thought and value, practices, materials and forms of contemporary experience distributed in space and time. They are especially concerned today with the emergence of what recent observers have called 'a massive, culturally saturated feedback loop' (Schutt and O’Neil, 2013: 5) between what people do and what they see. In this 'culturally saturated feedback loop,' none of the objects, techniques and subjects of digital sociology are coherent, well-understood or stable. While digital sociology draws on methods such as ethnography, rhetorical, discourse and visual analysis, new skills and digital tools, borrowed or copied from domains of statistics, software development, hacking, graphic design, audio, video and photographic recording and editing constantly come into play with lesser or greater relevance. Social research in such settings, as Noortje Marres notes, 'becomes noticeably a distributed accomplishment: online platforms, users, devices and informational practices actively contribute to the performance of digital social research' (Marres, 2012: 139). The subject who researches, the social researcher, is affected by this digital feedback loop in a variety of ways.

Digital sociology overlaps with, borrows from and seeks to make sense of much more prestigious, well-financed and heavily equipped practices in marketing, in scientific research, in government administration and commerce that target the expanded media environments of contemporary culture (Cukier and Mayer-Schoenberger, 2013). *Scale* and *pattern* are key problems in any contemporary social research. Scale and pattern can also allure social researchers. As danah boyd and Kate Crawford suggest, 'Big Data tempts some researchers to believe that they can see everything at a 30,000-foot view' (boyd and Crawford, 2012). But the mode of ordering – 'recurring patterns embodied within, witnessed by, generated in and reproduced as part of the ordering of human and non-human relations (Law, 1994: 83) – of digital sociology can differ greatly from contenders and analogues such as 'data science' (Schutt and O’Neil, 2013), 'predictive analytics' (Prediction Impact Inc., 2009) or the official statistics of governments. Seeking to re-vitalise the sociological imagination (Uprichard, 2012) and aiming to pay close attention to the ways in which the description of culture has become a key economic and social concern (Savage, 2009; Thrift, 2006), digital sociology might take shape less as a social science of culture than as the study of events that pattern lives, experience, power and value on various scales. This apparently limited ambition harbours some surprising expansive possibilities. It might cultivate 'a sociological sensibility not confined to the predominant lines of sight, the focal points of public concern' (Back and Puwar, 2012: 12)able to develop novel empirical techniques of inquiry and evaluate the unprecendented volume of information we encounter. It might generate novel empirical practices and conceptions of the empirical (Adkins and Lury, 2009).

As we will see, particularly through the examples we draw from working with digital data from one moderately important social networking platform, GitHub.com, the process of finding important events on the various scales and amidst the intricate patterns of practice of culture is not at all straightforward. Patterns, it turns out, entwine with problems of scale in both contemporary culture and in its analysis. The case we draw on here – Github.com – is interesting precisely because what goes on there – open source software development – both typical of the vast and somewhat incoherent spaces of contemporary culture, and plays an important role in reshaping them (to the extent that software ever does this). Github has a broader interest. The popularity of Github as a way of storing and coordinating work by many people on shared texts has attracted other users who are not developing software: how-to guides, metadata on the Tate's art collection, the White House's open data policy, legal documents, recipes, books, and blogs are just some of the diversifying use-cases now found in repositories on Github.As a recent article in *The Atlantic* suggests, Github is increasingly of interest to non-programmers because the de-centralised, distributed and trackable collaborative processes it supports can be used for many kinds of documents: designs, legal documents, maps, images, books, blogs or websites (Meyer, 2013). Github as a social media platform for software development is one segment of the emerging massive feedback loop where assessments, adjustments, and re-alignments of what people do are actively pursued and fed back to them via manipulations, visualizations, or modelling of digital data. Github, we might say, is an 'indexical icon' (Lee and LiPuma, 2002) of the problems of scale and pattern that confront digital sociology.

## Scales, scaling and scale-free networks

The anthropologist Anna Tsing writes: 'scale is the spatial dimensionality necessary for a particular kind of view, whether up close or from a distance, microscopic or planetary. I argue that scale is not just a neutral frame for viewing the world; scale must be brought into being: proposed, practiced and evaded, as well as taken for granted' (Tsing, 2005: 58). *Scale* refers to the relative dimensions of enlargement or a reduction in a map, picture or model, as well as the marks or degrees used to measure intervals (as in the scale on the axes of a scatter plot). *Scale,* however,is a verb as well as a noun. In digital culture, re-scaling or re-dimensioning is common. The *scaling-up* of databases, of transactions, of geographies (e.g. Amazon's data-centres divide the globe into [eight regions](http://docs.aws.amazon.com/AWSEC2/latest/UserGuide/using-regions-availability-zones.html)), and capacities in many settings testifies to one aspect of this re-scaling. Like many digital media, Github offers analytic and empirical capacities that might help us engage with different scales and scalings. But wrangling these affordances in sociologically inventive ways entails much engagement with the practices and technical realities of digital platforms. We might seek to craft ways of dealing with different scales, with wholes, with events and relations in Github, but the allures of totality or exhaustiveness powerfully distort and twist our efforts.

The outstanding issue for digital sociology to date has been the problems and potentials of studying culture on scales that do not align to either the individual or the whole culture. Given the scaling-up of digital infrastructure, there is a temptation to scale up empirical work. The conceptual poles of individual and whole culture have limited analytical purchase on open-ended, heterogenous and frequently re-scaling flows of contemporary cultural spaces. They are attached to more static concepts of agency and structure that have long been disputed (we are not re-visiting those debates here; see (Latour et al., 2012). For instance, while 'users' or consumers remain the focus of much market research and advertising, their demographic attributes of race, age, ethnicity, income and educational level are perhaps becoming less important as 'postdemographic' concerns of profiling in terms of groups, tastes, interests, device usage, etc., take hold (see (Rogers, 2013: 153–4)). The feedback loops between culture and technology mean that cultural processes change scale in many different ways. Digital sociology confronts the problem of working out if the scaling processes it observes are *isometric* (the same metric applies throughout) or *allometric* (different metrics are needed to make sense of the changes). On many accounts, network digital media display 'scale-free' characteristics – no matter what scale we look at, the same distribution of event types can be found. (We return to this issue below).

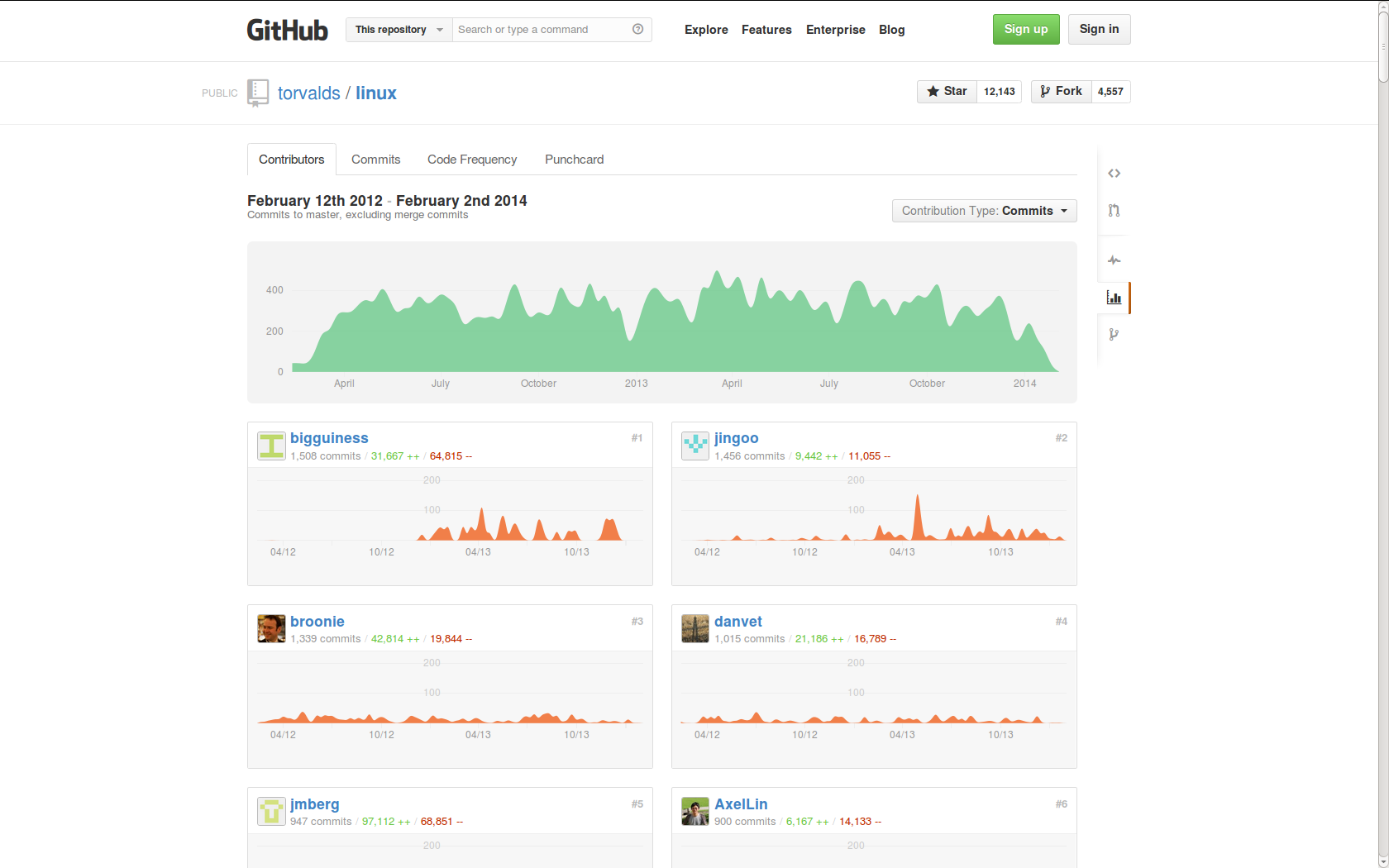
Given this uncertainty about the units and wholes of social fields, issues of scale and scaling processes are centre-stage in digital sociology. We can glimpse some of the problems of scale in Github. At the time of writing (February 2014), around 170 million events are recorded on the Github.com timeline since early 2012. (The timeline is the name for the comprehensive time-stamped series of user-generated events. The timeline does not include events staged by the people who manage and constantly tweak Github as platform. Platform-level events – a change in architecture, modifications to interfaces, shifts in underlying design or management practice – are much harder to see from the data.) Many of these events are highly emphemeral. They have little afterlife. Someone creates a repository and puts something there, and then never returns. Millions of such events occur. But other events represent the kernel of enduring, cascading or expanding transitions. In the flow of events marked on the Github.com timeline, there are some surprising features. For instance, the 18 months of Github event data graphed in Figure 1 shows growth.

Figure 1: Events on the Github.com timelineA description...

We might not be surprised to see a steadily increasing count of events on a social media platform. (Github is, after all, a social media platform for coding). The pattern of growth is common to many digital media where isometric increases in scale are often promised. But here we see something anomalous. The number of events per week in April 2012 and especially September 2012 exceed the number of events per week at the end of 2013. Two peaks appear ealier than they should. These peaks reflect something about the way Github as a platform stored data rather than a dramatic turn of events (for instance, some kind of Github virality or contagion).

Even if we manage to filter our these platform effects and archival anomalies (statisticians and scientists often have to clean data before they analyse it), events envelope fractal complexities. 'Push' events, for instance, occur when software developers move code from a local repository to a Github repository. While they are atomic events in Github's terms, PushEvents may comprise complicated 'payloads'. The payload complexity of events poses analytical problems. It subducts analytical work into the depths of highly localised practices, and disorients attempts to keep things on one scale. Sociologists have long been aware of the problem of there being too much to analyse in contemporary culture. To take a somewhat pre-digital sociology example, Nick Couldry writing in 1999 discussed at length how cultural studies could respond to the problem of 'too many texts' and proposed the notion of *textual environments* as a way of addressing this. 'We are looking', he writes, 'at a textual environment comprising complex patterns of flows: flows of meanings, texts and potential readers' (Couldry, 2000: 87). Even in this somewhat premonitory characterisation of 'vast textual fields' (87) as 'patterns of flows', the conventional empirical material of social research, the texts (construed broadly to include audiovisual, visual, numerical, and alphabetic inscriptions), constitute only one dimension of the analytical problem. Flows of meaning and flows of 'readers' – and today we might want to add 'writers' – cross-hatch the flow of texts in textual environments. And *textual* here was already understood as referring to specific material practices that vary widely: the textuality of a printed house design magazine means it is likely to be read differently to the now-declining but still vastly popular SMS mobile phone messages. Given the complexity of textual environments, we should be less surprised by variation or change, Couldry suggests, than we are by the existence of order: 'in the vast textual fields we inhabit, it is order rather than uncertainty that we need to explain' (87).

How does digital sociology address vast order? On what scale do we look for order if, as Tsing suggests (and any phenomenology of vision would agree), seeing is already implicitly scaling? One of the problems with the re-scaling of practice is that this scaling is animated by a pursuit of order, and totality. As mentioned above, Github itself encourages a focus on individuals and individual repositories. It affords little scope for examination of transverse flows or the textual fields that Couldry adumbrates. Even the all-important APIs, the interfaces that GitHub offers to developers, focus on providing data about individual repositories and individual users. we encounter a repository that invites software developers and programmers to store, work on and retrieve the source code and associated texts associated with software on many scales, ranging from individuals to large organisations. The writing of the operational texts of code, and to a certain extent, how they are read, is legible in Github repositories woven in subtle ways through code (for instance, in comments, in organisation of the repository, in the references to other software projects, etc). This public legibility is typical of contemporary feedback culture: things are made to be readable by many. Github.com itself produces and encourages the production of various forms of visualisation and tabulation of what goes on there. The fact that we can access an archive of the Github event timeline attests to this. GibHub ran a data competition in 2012 in which data analysts sought to do something with the timeline data (Briandoll, 2012). But what is most available from that data is a set of pre-formed event types (as we saw in Figure 1) that more or less reinforce Github's conception of itself as first of all a *hub* and secondly as *social* media. We can easily for instance examine the most important or famous repository on Github – the Linux kernel, a much-vaunted, and now commercially and technically vital open source software project. Relatively quickly, individual contributions could be analysed, and we could begin to characterise the composition of the group of people who keep this important software object working and up-to-date. And indeed, this work is largely already done by Github.com itself (as the screenshot in Figure 2. shows).

Figure 2: Top Contributors to Linux on Github

It would even be possible through careful reconstruction to graph the changes in the network of relations that comprise the Linux kernel development teams, and perhaps to see how patterns of work on Linux kernel have changed over time as its economic and technical significance has extended through the popularity of the Android phones that rely on Linux and through the many corporate and industry users/producers invested in Linux. Issues, conversations, team and organisational changes could all be mapped. We could look further at the thousands of 'forks' and 'mirrors' of Linux to be found on Github, and characterise in fine-grained ways how processes of imitation generate flows of readers/writers, how meanings and practices are stabilised through repetition, and how invention might occur at intersections or overlaps between fields.

If even a single large repository on Github might require gargantuan effort to assimilate and analyse how it unfolds in cultural-economic-technical space, what hope do digital social researchers have in analysing even a platform such as Github? One response might be to say that we should be look for order in scale. That is, the forms of order digital sociology needs to analyse include the persistent *or isometric* scale rather than the uncertainties of scale (allometric scaling*)*. Github has millions of members, more than ten million repositories (according to its own reports), and is barely coherent because of the explosive growth in hardware platforms, media forms and ways of writing code found there. Yet great commonalities of practice occur on different scales. For instance, a particular code construct to retrieve a web page and extract its text content might be imitated or reinvented in thousands of different projects. The same kinds of software are re-implemented or imitated across many different projects, often with only small variations. Viewed from another another angle, those commonalities of practice point to astounding differences: the same code constructs might be found in projects focusing on an ecommerce platform, ecological models of rainforest species diversity, Mozilla Foundation's Firefox browser, the Linux kernel, a library to manipulate financial data, or a modified version of the Android operating system for smart phones. In short, in many of the settings it encounters, digital social research finds little coherence. It finds almost fractally dimensioned patterns of social action and transaction, linked by the common practices, formats and materials (code patterns, repository mechanisms), but varying widely in time and space, from ephemeral repositories that an individual creates and abandons, to large organisations coordinating projects of long duration.

In much social research, problems of scale largely related to scarcity. Where data was abundant, scale problems were handled by setting limits on data through sampling strategies, research methods and research designs that allowed social researchers to be more or less confident that their research covers the social field of interest (for instance, making sure that differences in age, gender, ethnicity, sexuality, nationality, education-level, or income are represented in the data; many of the chapters in a typical introductory statistics textbook for social science address these problems of scarcity and represetativeness). Selecting and sampling strategies seem to work differently in digital sociology as it encounters expanding textual environments where data scarcity is rarely an issue. The much discussed problem is how to cope with the vast amount of material. What's worse, textual environments such as social media are explicitly expansive. Their expansion is a material component of their mode of existence and they offer many different affordances for expansion. For instance, every Github repository has prominent web-buttons 'clone this' or 'forkme' that allow visitors or viewers of the repository to either copy the repository or make a new version of it under their own name. (By contrast, older-style code repositories such as SourceForge.com only allowed people to download copies of the code or software.) While the textual environment for online code repositories such as Github includes the many forms of text and graphic visible on Github webpages, it also includes the flows of data that these platforms generate for use by others. As millions of people interact even with mid-level platforms such as Github, their actions generate large volumes of data that can be streamed as time-stamped events. So, like many other social media platforms, Github publishes all of these individual time-stamped events to anyone who wants to use them. (While some users may pay to have their repositories and working practices remain invisible, very many do not.) The events are available more or less as they happen (through the Github API – application programmer interface) or in bulk through various archives ([http://githubarchive.org](http://githubarchive.org/); a mirror of the data is also published by Google as a demonstration of their 'BigQuery' cloud computing service). The data derived from people what do on social media is mixed in form. It includes when things were done (date-time), various free textual forms (descriptions, tags, titles, etc), structured text (links or URLs of associated webpages; names of associated organisations and groups; names of user), categorical attributes (on Github these are limited: the event type, and the programming languages used), and often a mathematically encoded summary of the result of actions that change the contents of the repository (in the case of Github, the *hash digest* of any new content or change to existing contents of the repository).

Although it might seem a purely technical or practical issue, the problem of actually traversing such data flows lies at the heart of digital sociology as well as many parts of contemporary culture. The constant updating of events, the relatively frequent advent of new flows of data, the teeming and burgeoning ways of inhabiting the reefs of digital infrastructures, many of which are developed and available on Github itself, confront digital sociology with challenges and alluring possibilities. Unlike the major social media platforms or even the legion of startup companies in London, New York, San Francisco, Shanghai, Berlin or Amsterdam who offer ways of packaging, summarising, monitoring or shaping flows of data in networks, digital sociologists do not have hundreds of developers to wrangle data, dozens of computers and disk-drive arrayed in racks to expedite the process of searching or exploring the data. Given the billions of events welling-up in diverse APIs, what can digital sociology do? Should it sample and filter according to the criteria of adequacy and representativeness?

The expansive forms of textual environment we have just been describing are writ large at the moment under the broad banner of 'big data.' The proliferating discussions of 'big data' need to be analysed in their own right in terms of how they intensify desires to connect information flows, previously disparate infrastructures and systems (energy, telecommunications, entertainment, transport, retail and manufacturing), and how they actually reorganise work, domestic lives, forms of sociality and value in the name of flows of data. 'Big data' is certainly part of the feedback loop or accumulation strategy in which social practices recorded as data become the basis of new textualities that seek to enrol further readers or writers, to align reading, viewing and buying, writing and working, as well as other forms of value. The physical, life and environmental sciences offer both a lead and something distracting here. The term 'data' carries with it an aura of scientificity, objectivity or neutrality that digital sociology is still wrestling with. 'Big data' is an expansive grouping and its memebership continues to grow: house prices, clicks on hyperlinks, vehicle detection loops on roads, mobile phone call details, satellite photos of crops, electronic payment, stellar images from orbiting telescopes, transactional data such as credit card authorisation or supermarket checkout scanners – these are just examples in a list that keeps growing. The listing of data sources is a interminable feature of most talk about big data (and digital sociology shares this habit). Scientific data, however, has a particular resonance and perhaps anchors some forms of referentiality in data talk. The standard reference to allude to transformations in scientific data is *The Fourth Paradigm: Data-intensive Scientific Discovery*  (Hey et al., 2009), a book published by Microsoft Research Press (a publisher that probably lies quite close to the source of much business data practice). This book furnishes vignettes of a range of scientific enterprises ranging across physical, earth, environmental and life sciences in which flows of data have transformed knowledge-making practices. The data intensive sciences authorise data practices in specific ways. The auratic power of scientific instruments such as DNA sequences or infrared satellite photos differ, as Mike Savage observes, from the 'mundane descriptions, … ordinary transactions, websites, Tesco loyalty cards, CCTV cameras in your local shopping centre, etc., that are the stuff of the new social' (Savage, 2009: 171). It may be that this auratic/mundane difference, important though it is in differentiating certain practices, also usefully links different domains of the social. Auratic scientific instrument data, with its referential links to the diversity of life, the fate of the planet or the conundrums of missing matter in the universe, rivets data to things. The scientific examples allow data more generally in all its administrative, transactional or media-derived forms to carry universalising epistemic value. It suggests that the birth of stars in remote galaxies can be analysed in similar terms to the birth of stars in the media environments of Xfactor or reality TV shows. In this respect, the popularity of the term 'data *science*' suggests that the referential power of science matters to business, commerce, industry and government as they seek to commodify, extract or regulate contemporary cultural spaces.

At the same time, important differences are at work. For instance, people's interactions with Github.com mostly appear in the API data stream as time-stamped events. These events are categorised according to 14 different event types (PushEvent, CommitEvent, AddUserEvent, DeleteEvent. These event types are organic to Github.com, but as variables in any data analysis they are defined by the platform designers rather than by any questions that social researchers might bring to bear on what people do on Github. 'Source-defined variables' are a central concomitant of data analysis practice in digital sociology. We might say that in the new social fields, data has an 'organic' aura: it is generated and collected by virtue of the existence of the infrastructures and platforms that are part and parcel of the social field rather than from instruments or measuring devices introduced by market or social researchers. But even this relationship is becoming increasingly complicated by virtue of the intricate and shifting relations between 'organic data' and 'paid for data' (Google Inc., 2009)). Even if this data is 'observational' rather than experimental, the fact that is generated intrinsically as part of a social field has powerful referential attractions, and attracts much work. The scaling up of the information infrastructures needed to deal with it is a constant feature of the software projects on Github.com. Many of them deal with problems of scale (for instance, over 1500 repositories relate to 'MapReduce,' a parallel-processing algorithm first developed at Google to improve web search engine response times: <https://github.com/search?q=map+reduce>; 3700 repositories engage with Hadoop, a single open source implementation of the MapReduce algorithm: https://github.com/search?q=hadoop;). This scalin produces auratic effects that relate to the infrastructures themselves. The infrastructures that increasingly bring people into various forms of relationality are difficult to concretely imagine. We see somewhat sublime images of earth-spanning information networks or visualizations of networks of hundreds of millions of Facebook users, but it is the data flows they produce, even if they are largely inaccessible or discarded by all but the most data-hungry government intelligence agencies, that promise to tell us most about how ideas and values move and differentiate, sometimes coherently and often incoherently. This too is something that digital sociology might address: the organic-infrastructural aura associated with large-scale data flows. Why does Google make available via their 'BigQuery' service a massive public archive of all the data produced by Github since early 2012? While we, as social researchers, might find it is enormously useful to have an aggregated, hourly-updated timeline of all Github actions, it is very unlikely that some part of Google concerned with data architectures for cloud computing has social researchers particularly in mind. Much more likely, they seek to attract the attention of the millions of software developers who use Github.com for coding work. Trying to see themselves and others in this high volume datastream, software developers and programmers familiarise themselves with Google's BigQuery architecture, and perhaps use it more thereafter. This is certainly our experience as social researchers working on Github.com. The Google BigQuery datasets empowers a different scale of exploration, at a certain cost of being drawn in to large scale tabulations of results and away from the singularity and variability of practices.

## Deriving patterns in digital sociology

An interest in patterns lie at the heart of digital sociology as it is currently taking shape. A crucial question here is how to think about pattern. As we have seen, in sociological thought more generally, pattern is a long-standing concern. John Law speaks of 'recurring patterns,' Nick Couldry of 'patterns of flow,' and Andrew Abbott suggests that 'if most things that could happen don’t happen, then we are far better off trying first to find local patterns in data and only then looking for regularities among those patterns' (Abbott, 2001: 241). A broader philosophical re-conceptualisation of patterns runs through some social and cultural theory (for instance, in recent work influenced by A.N. Whitehead, who writes 'beyond all questions of quantity, there lie questions of pattern' (Whitehead, 1956: 195). Patterns and the learning of patterns are the central pre-occupation in many contemporary sciences, in financial markets, in biomedicine, and business analytics (see for instance, the fields of 'pattern recognition' and 'machine learning' (Hastie et al., 2009)).

In many cases, people craft these patterns in visualizations. Much digital sociology seeks to either make or reuse visual patterns of cultural processes. The principal visible forms of pattern include plots that show lines, curves, peaks and clusters of points, networks, trees, and maps. Drawing on the spectrum of plots, graphs and diagrams developed in the last few centuries (see <http://www.datavis.ca/milestones/> for a catalogue; Edward Tufte's work (Tufte, 2001) is a standard reference for quantitative digital data), contemporary visual displays of pattern abound in 'predictive analytics and the move back to visualization in social statistics, the new cartography and associated Web 2.0 innovations, [and] visual montages designed to represent amalgams of “variables”' (Burrows, 2012: 585). Sociological research visual culture of data visualization is surprisingly scarce. The renewed emphasis on visualization in digital sociology differs somewhat from adjacent efforts such as computational social science (Giles, 2012; Housley et al., 2013) where visualization is usually closely coupled to statistical or predictive models. In digital sociology, visual forms and the practices of viewing visual forms are taken seriously as both an object of inquiry and as a challenge to methodological practices precisely because the visual forms attest to a shift away from some traditional sociological concerns with abstractions, models and structures as deep explanations of social processes, and a lighter, perhaps more responsive attunement to patterns, groupings and flows. This contrast has been extensively debated in sociology. (See Andrew Abbot's discussion of patterns versus causes; the 'empirical crisis in sociology' literature; as well as the explicit focus on digital devices in recent sociology (Abbott, 2001; Ruppert, 2013; Savage and Burrows, 2007)) and we will not rehearse these debates in great detail here. They have been debated fairly widely elsewhere (see (Burrows, 2012) for an overview). The central point is that patterns have begun to supplant causes as modes of explanation in many predictive settings. Although patterns still usually need to be interpreted in some way, the main mode of interpretation is not in terms of causal models.

These patterns can take the form of plots (bar graphs, scatter plots, lines and curves drawn through clouds of data points, network diagrams), maps on many different scales (especially maps that superimpose different geo-located datasets), information visualizations (typically combining data graphics, text and typographic design elements), as well as tables and textual-graphic forms (word clouds), with varying degrees of animation and with many different forms of scale. These descriptive visualizations might prompt some causal interpretation in their viewers but they are not premised on any such abstraction. While models of various kinds may have been involved in producing them (for instance, many smoothing algorithms used to draw lines through points effectively fit a series of local linear models – splines – in building a smooth curve), the models themselves are not interpreted as such by act in the world more like things than thoughts. A related point is that techniques of description that rely on pattern matching and pattern recognition have implicitly become mundane parts of contemporary culture in many different ways (for instance, in the face recognition logic now built into many digital cameras, or the much-discussed recommendation systems typical of online commerce). And they are in any case widely distributed through various social fields where visual displays, metrics, dashboards, graphs, and visualizations form part and parcel of social life, whether in the 'dashboard' that users of the microblogging site Tumblr see, the graphic displays that users of supply chain and inventory management systems or financial traders gaze at (Knorr Cetina and Bruegger, 2004), or in the many news-related and particular sport-oriented visualizations produced by data journalists for news sites such as *The Guardian* datablog (Guardian News, 2009) or the *New York Times'* highly visible news visualizations *(New York Times, 2012)*. Some commentators suggest that the production and use of such visualizations is a key concern for contemporary sociology: 'the discipline … will have to take visualization methodologies far more seriously than we have hitherto' writes Roger Burrows (Burrows, 2012: 585) because of the ways that they are being used and could be used to understand 'particular patterns of association that exist between persons, objects, symbols, technologies and so on' (585).

Sociological work on the transformations of data visualisation is still rather scarce. Although network visualisations, tag clouds, stream graphs and the like have abounded in data visualisation on the web, especially with the growth of graphics libraries and packages such as Hadley Wickhams *ggplot2 (Wickham, 2009)*,Mike Bostock's *d3.js*, IBM's older *ManyEyes* or widely-used scientific plotting packages such as *matplotlib*  and *Matlab,* and many others, we have little sense yet of the visual culture of these devices and their visual forms. Sociological work on scientific visualization offers some leads here (Latour et al., 1990; Myers, 2008) alongside work on visualization in finance (Pryke, 2010), but the visual culture of data as it moves out of scientific publication has received little attention. There is much scope for investigation of the seeing in data visualisation if, as Gillian Rose writes, 'different ways of seeing are bound up into different, more-or-less conscious, more-or-less elaborate, more-or-less consistent practices' (Rose, 2012: 549). Practices of data visualisation are routinized through the proliferation of certain visual forms (the network diagram, bubble plots and chloropleth maps are widely found) in different places. What do they make make visible? Both in the visual culture of data, and in its own visualisation of digital data, digital sociology faces the problem of describing how patterns are produced at the intersection of various concretisations and abstractions. As Mike Savage suggests, we need to understand 'how pattern is derived and produced in social inscription devices' (Savage, 2009: 171), whether these devices are objects of analysis or part of our own methods. (Again, overlaps between the object of digital sociology and the methods of digital sociology occur at every turn.)

Patterns arise in very different ways. On the one hand, many data visualisations seek to show something that is hard to see because it occurs on spatio-temporal scales that are difficult to directly see. For instance, Figure 3 seeks to convey something of the patterns associated with different scales of activity in Github repositories by counting events that appear in the Github timeline over an interval of two years (2012-2013). The general pattern shown here is the somewhat annoyingly ubiquitous 'power law' distribution of events that often shows up in social media data. At the left hand end, the high point refers to the millions of Github repositories consisting of one or two events. At the low end on the right, the curve approaches the x-axis, a small number of repositories receive many thousands of events. The power law distribution of events in social media often vexes data analysis and data visualization. Many social media processes yield heavy-tailed distributions when graphed. This common scaling of events across the 'many-some-rare'scales (Conte et al., 2012: 334) requires site-specific work. The visualization of repository event counts begins to do this by showing something of the different composition of the repositories on the different scales. The many small repositories mainly consist of the basic content-creating Push events. Mid-scale repositories show the presence of more social events such as Watch and Fork. The rarer very large repositories attract many more social events – Fork, Watch and PullRequest.

Figure 3: Patterns of repository events on GithubA description...

But if this pattern of scaling between many-some-rare is so common ('In recent years, due to ubiquitous computerization, networking and obsessive data collection, reports of heavy-tailed distributions have almost become a routine' (Muchnik et al., 2013: 1)), what can digital sociology do? On the one hand, many data visualisations whether in the form of networks, scatter plots or line graphs today present the power-law or scale-free pattern of digital media. What then of the cultural saturated feedback loop described above? As technological concretisation binds practices, habits, emotions, and interactions through digital devices and infrastructures and devices, the derivation of patterns increasingly depends on abstractive modes of knowing that classify, cluster, calculate and predict events precisely in order to shape them. Predictive analytics, as demonstrated in Google Research's work on how users' searches foreshadow airline ticket bookings or car sales (Varian and Choi, 2009), derives patterns from data using a much more technical armature of machine learning techniques. This modelling is an increasingly dense force affecting the feedback loop between people and digital infrastructures (Pariser, 2011). Techniques such k-means clustering, nearest neighbour classification, linear regression, logistic regression, principal component analysis, neural networks, decision trees, random forests, and support vector machines are rapidly becoming an integral part of every level and niche of digital assemblages, ranging from playful mundane devices such as 'kittydar, 'a neural network to detect cat photos (<https://github.com/harthur/kittydar>) through to thousands of projects implementing 'face detectors' or 'motion detectors' for smart phones, web browsers, and for different operatings systems. Even a single technique like the popular random forest classifier (Breiman, 2001) can be found hundreds of thousands of times in gitbhub.com and tens of thousands of times in a popular programming languages such as R (R Development Core Team, 2010). The proliferation of these techniques is perhaps much less visible, and the ways in which they imprint or weave through flows of meaning and things is harder to analyse. They are somewhat withdrawn elements in the feedback loops of cultural space. These predictive models and classifiers sometimes operationally shape the experience and action (as in Netflix or Amazon recommendation systems, or in the classifiers that detect and classify body gestures in the Microsoft Kinect game controller), and sometimes they are analytic tools used by people working on platforms trying to make sense of emerging or divergent patterns in practice. As always, the feedback loops between knowing and acting are hard to disentangle precisely because they are becoming more tightly coupled. If decision trees were an analytic techniques developed by statisticians in the late 1970s trying to make sense of air pollution measurements in Los Angeles (Breiman et al., 1984), in the Kinect game controller they become predictive devices that intensify the immediacy of computer game play. Online learning – the constant updating of predictive models in response to the flow of current events – is increasingly common in social network media and online transactions.

Patterns, then, are made, acted upon, altered and re-imprinted in much more rapid cycles. In such cases, pattern recognition is no longer a practice conducted at leisure by expert interpreters or analysts. They are operational components of the textual environment. The transformation of abstract analytic devices of many different kinds (linear regression models, clustering algorithms, Bayesian models, etc.) into things that either circulate much more widely in the world in gadgets and devices or into meta-things such as search engines that modulate flows on a large scale is a central component of the flow of texts, meanings, audiences, viewers, visitors, spectators, readers and players in many settings. These abstractions do exercise an auratic effect akin to the data infrastructures we were discussing above. Journalists, social researchers and commentators on digital technology tend to attribute great potency to algorithms in general. The proliferation of algorithms is hard to deny, but digital sociology might play an important role in describing what happens as these algorithms shift shape and move into different settings ranging from pay day loans to computer game play, from web search engines or ecommerce recommendations to face recognition in digital cameras. The aura of algorithms as epistemic prime movers akin to steam power or electricity covers over their diverse provenance (they do not come just from computer science but also from statistics, psychology, cognitive science, ecology, archaeology or geology) and their diversity in practice. Perhaps more significantly the patterns that these algorithms produce or derive from data are neither obviously legible in descriptive devices, nor is their relation to existing structures, groupings, or classifications direct.

For digital sociology, recognising the effects of this pattern-making is a significant challenge. The imprinting of flows of meaning, media and practices is sometimes legible (for instance, in the 'induced viralities' seen in various social media platforms that identify trends using pattern recognition and then feed them back into the flow of messages or network connections), but not always. Another difficulty is much more challenging. While descriptive devices can relatively easily slip into the analytical toolkit of social researchers – the spread of tag-clouds or Wordle graphics would be a typical instance of this – the appropriation and re-purposing of machine learning, pattern recognition or data mining approaches is more problematic. Many of these devices rely on formidable mathematical apparatus, ranging across linear algebra, probability theory, function analysis and numerical optimisation. The diverse provenance of the techniques means that, although they operate abstractly (that is, with little regard for the concrete specificities of a given situation), they handle notions of group, classification, difference and similarity heterogeneously. Nearly all of them bring to bear powerful scaling processes that reduce the high dimensions and volume of data to legible forms of variation and pattern. Again, scale and pattern are intimately interwoven here.

## Conclusion

Organised attempts to analyse and study social patterns have animated the growth of large-scale digital infrastructures and social media. The algorithmic elements of search engines such as Google Web Search's *PageRank*  algorithm or Facebook's *Social Graph* suggest that the detection of social patterns and flows of meanings, texts and readers has been a pivotal device in the growth of digital culture. Where does this leave digital sociology? Several years ago, one of the editors of the current volume wrote:

It follows that a core concern [for social research] might be to scrutinize how pattern is derived and produced in social inscription devices, as a means of considering the robustness of such derivations, what may be left out or made invisible from them, and so forth. We need to develop an account which seeks to criticize notions of the descriptive insofar as this involves the simple generation of categories and groups, and instead focus on the fluid and intensive generation of potential (Savage, 2009: 171).

Patterns in cultural life today derive from social inscription devices assembled in wide-ranging feedback loops. Feedback runs between recording what people do, visualising or graphing what they do, finding/generating patterns in the recording, and shaping what they encounter next. If digital sociology attempts to describe what is happening in contemporary textual and media environments, it needs to map the paths of these feedback loops running across publics, infrastructures, techniques, textual and media forms, and diverse, expanding practices. At the same time, pattern themselves are increasingly heavily analysed and modelled inside the culturally saturated feedback loop between people and social inscription devices. Digital social research is not alone in its interest in this processes. Reputational, attentional and sentiment economies (Arvidsson, 2011) directly act on that patterning.

We suggest identifying events that link this patterning and scaling is a key concern for digital sociology. The 'massive, culturally-saturated feedback loop' produces massive feedback-saturated cultural loops. As the case of Github shows, the process of implementing social inscription devices is itself a highly social field filled with its own patterns and problems of scale. The 'simple generation of categories and groups' that Savage refers to can certainly be found there, replicating and propagating at scale. Furthermore, digital sociology and certainly social research more generally is not immune from the aura of 'big data.' It is always grappling with the reflexive-recursive problem of its own implication in methods, techniques and infrastructures for deriving pattern. We can easily, as digital social researchers, find ourselves lost in the labyrinth of technical possibilities opening up around platforms, tools, visual forms and data flows. In this labyrinth, we have been suggesting, the entanglements of pattern and scale present the primary objects and means of analysis of saturated feedback loops typical digital cultures. These entanglements between scale and pattern, and especially between the different ways in which pattern might be found could offer a way to avoid the homogeneity and uniformity of predictive analytics with their highly constrained commitment to increasing advertising revenue or sales.

## References

Abbott A (2001) *Time matters: on theory and method*. University of Chicago press.

Adkins L and Lury C (2009) Introduction: What Is the Empirical? *European Journal of Social Theory*, 12(1), 5–20.

Arvidsson A (2011) General Sentiment: how value and affect converge in the information economy. *The Sociological Review*, 59, 39–59.

Back L and Puwar N (2012) *Live Methods*. Wiley-Blackwell, Available from: http://library.wur.nl/WebQuery/clc/2016616 (accessed 28 November 2013).

boyd danah and Crawford K (2012) Critical Questions for Big Data. *Information, Communication & Society*, 15(5), 662–679.

Breiman L (2001) Random forests. *Machine learning*, 45(1), 5–32.

Breiman L, Friedman J, Olshen R, et al. (1984) CART: Classification and regression trees. *Wadsworth: Belmont, CA*, 156.

Briandoll (2012) Data at Github. *GitHub*, Available from: https://github.com/blog/1112-data-at-github (accessed 3 March 2014).

Burrows R (2012) Digitalization, Visualization and the ‘Descriptive Turn’. In: Heywood I, Sandywell B, Gardiner M, et al. (eds), *The handbook of visual culture*, London: Berg, pp. 572–588.

Conte R, Gilbert N, Bonelli G, et al. (2012) Manifesto of computational social science. *European Physical Journal-Special Topics*, 214(1), 325–346.

Couldry N (2000) *Inside culture : reimagining the method of cultural studies*. London: SAGE.

Cukier KN and Mayer-Schoenberger (2013) The Rise of Big Data. How It’s Changing the Way We Think About the World. *Foreign Affairs*, Available from: http://www.foreignaffairs.com/articles/139104/kenneth-neil-cukier-and-viktor-mayer-schoenberger/the-rise-of-big-data (accessed 10 February 2014).

Giles J (2012) Computational social science: Making the links. *Nature*, 488(7412), 448–450.

Google Inc. (2009) Back to Basics: Direct, referral or organic - definitions straight from the source - Analytics Blog. *Google Analytics Blog*, Available from: http://analytics.blogspot.co.uk/2009/08/back-to-basics-direct-referral-or.html (accessed 7 February 2014).

Guardian News (2009) Data journalism and data visualization from the Datablog | News | The Guardian. Available from: http://www.theguardian.com/news/datablog (accessed 30 January 2014).

Hastie T, Tibshirani R and Friedman JH (2009) *The elements of statistical learning: data mining, inference, and prediction*. New York: Springer.

Hey T, Tansley S and Tolle K (2009) The fourth paradigm: data-intensive scientific discovery. *Microsoft Research*.

Housley W, Williams Matthew, Williams Malcolm, et al. (2013) Computational Social Science: Research Strategies, Design and Methods Introduction. *International Journal of Social Research Methodology*, 16(3), 173–175.

Knorr Cetina KK and Bruegger U (2004) Traders’ engagement with markets: a postsocial relationship. *The Blackwell cultural economy reader*, 121–142.

Latour B, Lynch M and Woolgar S (1990) Drawing things together. In: Anonymous (ed.), *Representation in Scientific Practice*, Cambridge, MA London: MIT Press, pp. 20–68.

Latour B, Jensen P, Venturini T, et al. (2012) The Whole is Always Smaller than its Parts. How Digital Navigation May Modify Social Theory. *British Journal of Sociology*, 63(4), 590–615.

Law J (1994) *Organizing modernity*. Oxford, UK ; Cambridge, Mass., USA: Blackwell.

Lee B and LiPuma E (2002) Cultures of Circulation: The Imaginations of Modernity. *Public Culture*, 14(1), 191–213.

Lupton D (2012) Digital Sociology: An Introduction. Available from: http://prijipati.library.usyd.edu.au/handle/2123/8621 (accessed 14 November 2013).

Marres N (2012) The redistribution of methods: on intervention in digital social research, broadly conceived. *The Sociological Review*, 60(S1), 139–165.

Meyer R (2013) Github, Object of Nerd Love, Makes Play for Non-Programmers. *The Atlantic*, Available from: http://www.theatlantic.com/technology/archive/2013/08/github-object-of-nerd-love-makes-play-for-non-programmers/278971/ (accessed 13 February 2014).

Muchnik L, Pei S, Parra LC, et al. (2013) Origins of power-law degree distribution in the heterogeneity of human activity in social networks. *Scientific Reports*, 3, Available from: http://www.nature.com.ezproxy.lancs.ac.uk/srep/2013/130507/srep01783/full/srep01783.html (accessed 1 March 2014).

Myers N (2008) Molecular embodiments and the body-work of modeling in protein crystallography. *Social Studies of Science*, 38(2), 163–199.

New York Times (2012) 2012: The Year in Graphics. Available from: http://www.nytimes.com/interactive/2012/12/30/multimedia/2012-the-year-in-graphics.html (accessed 30 January 2014).

Pariser E (2011) *The filter bubble: What the Internet is hiding from you*. Penguin UK, Available from: http://books.google.co.uk/books?hl=en&lr=&id=-FWO0puw3nYC&oi=fnd&pg=PT3&dq=eli+pariser&ots=g2PoCtpQV-&sig=3\_CftKt2BPOwLVpT\_OFzizJr\_-c (accessed 13 February 2014).

Prediction Impact Inc. (2009) Predictive Analytics World Conference: Agenda. Available from: http://www.predictiveanalyticsworld.com/sanfrancisco/2009/agenda.php#usergroup (accessed 24 May 2011).

Pryke M (2010) Money’s eyes: the visual preparation of financial markets. *Economy and Society*, 39(4), 427–459.

R Development Core Team (2010) The R Project for Statistical Computing. Available from: http://www.r-project.org/ (accessed 11 June 2010).

Rogers R (2013) *Digital methods*. Cambridge, Massachusetts; London: The MIT Press.

Rose G (2012) The Question of Method: Practice, Reflexivity and Critique in Visual Culture Studies. In: Heywood I, Sandywell B, Gardiner M, et al. (eds), *The handbook of visual culture*, London: Berg, pp. 542–558.

Ruppert E (2013) Rethinking empirical social sciences. *Dialogues in Human Geography*, 3(3), 268–273.

Savage M (2009) Contemporary Sociology and the Challenge of Descriptive Assemblage. *European Journal of Social Theory*, 12(1), 155–174.

Savage M and Burrows R (2007) The Coming Crisis of Empirical Sociology. *Sociology*, 41(5), 885–899.

Schutt R and O’Neil C (2013) *Doing data science*. Sebastopol, Calif.: O’Reilly & Associates Inc.

Thrift N (2006) Re-inventing invention: new tendencies in capitalist commodification. *Economy and Society*, 35(2), 279–306.

Tsing AL (2005) *Friction : an ethnography of global connection*. Princeton, N.J.: Princeton University Press.

Tufte E (2001) *The visual display of quantitative informations 2nd ed.* Cheshire, Conn.: Graphics Press.

Uprichard E (2012) Being stuck in (live) time: the sticky sociological imagination. *Sociological Review*, 60, 124–138.

Varian H and Choi H (2009) Official Google Research Blog: Predicting the Present with Google Trends. Available from: http://googleresearch.blogspot.com/2009/04/predicting-present-with-google-trends.html (accessed 23 May 2011).

Whitehead AN (1956) *Modes of thought; six lectures delivered in Wellesley College, Massachusetts, and two lectures in the University of Chicago*. New York,: Cambridge University Press.

Wickham H (2009) *ggplot2: elegant graphics for data analysis*. Springer New York, Available from: http://had.co.nz/ggplot2/book.