Media Reviews

Annals and Analytics: The Practice of History in the Age of Big Data

Research in history is a very problematic issue. Theoretically, the best tool for historical research – a time machine – does not (and probably will not) exist. In such a situation, the problem of how we know about the past becomes a painful dilemma.

Nachman Ben-Yehuda[1]

As much as historians would like to add a time machine to their toolkit, research in history is not problematic only because of insufficient data, nor because most extant data have passed through the filtration and interpretation of second-hand observation. On the contrary, for many historians the sheer quantity of available information – what William Turkel terms the 'infinite archive' of digital materials – cannot be processed using traditional methods alone. Far from solving this problem, a time machine would exacerbate it, adding more and richer data into the mix. In addition, there are important historical questions that cannot be answered solely through close readings of texts or through direct observation of the past; both cases overestimate the historian's powers of observation, implying that critical analysis and ethnography can solve all historical puzzles.

Yet it is also dangerous to assume that more or more accurate data will necessarily lead to better understanding. The view that computers can take massive amounts of information and do most of our analytic thinking for us, a belief embraced by many data miners and glorified by tech evangelists, often yields statistically significant but conceptually meaningless results. We can and should outsource *some* of our thinking to smart machines, much as we have outsourced some of our memory to books and other media; but to do this well is to understand the limitations and leverage the affordances of different approaches to processing and analysing information, both human and machine.

The practice of historical research stands to benefit from, and may even require, a mixed-methods approach that incorporates the analytic strengths of human interpretation and computational processing. In this brief reflection, I explore one approach to mixed-methods history using *network analysis*: various statistical techniques with which the structure of connections among entities – people, places, concepts and so forth – can be modelled. Network analytic approaches are commonly applied when the connections among entities reveal more than analysis of the entities in isolation. Simply knowing which historians attended the opening reception at a conference, for example, does not contribute as much to our understanding of that scholastic community as knowing who talked to whom during the event.

To the extent that historians have used network analytic techniques, most have modelled *explicit historical networks*: correspondence communities, shipping networks, citation patterns and so forth.[2] These are models in which both the nodes (the entities connected

[1] Nachman Ben-Yehuda, 'History, Selection, and Randomness: Towards an Analysis of Social Historical Explanations', *Quality and Quantity*, 17, 5 (1983), 354.
[2] For an overview, see Shawn Graham, Ian Milligan and Scott Weingart, *Exploring Big Historical Data: The Historian's Macroscope* (London: Imperial College Press, 2016).

in the model) and the connections among them are defined by criteria for which there is direct, unambiguous evidence: epistlers and letters exchanged, ports and sailing routes, scholars and works cited. Yet there is considerable opportunity for modelling *implicit historical networks*, such as the conceptual connections that characterise rhetorical strategies or complex ideas. In his work on abolitionist arguments in nineteenth-century newspapers, for instance, Timothy Shortell argues that 'the sociocognitive structure of a discourse' can be modelled 'as a networked field of concepts from which arguments are fashioned'.[3] This approach allowed him to characterise and to quantify patterns of abolitionist argument, as well as changes in those patterns, in five newspapers over three decades.

Conceptual models can thus be abstracted from textual evidence as networks of relations among constructs. Disease, for example, is not simply a pathophysiological process; as Charles Rosenberg has argued, it is 'a biological event, a generation-specific repertoire of verbal constructs reflecting medicine's intellectual and institutional history, an aspect of and potential legitimation for public policy, a potentially defining element of social role, a sanction for cultural norms, and a structuring element in doctor/patient interactions'.[4] To understand disease as a discursive construct is thus to understand the interrelations among all these dimensions – in other words, to see it as a complex network of associations among biological, interpersonal, social, cultural, political, institutional and historical factors.

Now, suppose we want to understand *malnutrition* in the way Rosenberg suggests, that is, not merely as a pathological condition of the body but as a sociocultural construct characterised by a complex association of concepts. We could start, for example, by utilising any of a number of text mining techniques to determine which words most frequently occur in proximity to malnutrition across some corpus. (In all likelihood, the hard part would be building such a corpus in the first place, but for the present discussion, let's assume it already exists.) We could model the strength of association of all these terms as networks to explore different understandings of malnutrition, indicated by statistically significantly different patterns of association among terms, in various times, places or contexts. Text mining techniques have several distinct advantages – they are fast, consistent and replicable, and they produce quantitative data that can be used for statistical hypothesis testing – but when used indiscriminately, there is a significant disadvantage: such techniques are context agnostic, which may obscure critical issues in our understanding of how texts in the corpus should be interpreted.

For example, the term *deficiency* will be closely associated with the term *malnutrition*, but there are several important caveats that unsupervised text mining will not readily reveal. For one, there are significant differences between a holistic concept of deficiency (a lack of food) and a biochemical concept of deficiency (a lack of calories or specific nutrients). More problematically, malnutrition was often defined *in terms of* deficiency but *as distinct from* specific deficiency diseases, such as scurvy, rickets or pellagra. Lastly, we might find that many words associated with *malnutrition* do not provide meaningful insight. A topic model of Emily Dickinson's correspondence, for example, found that one of the words most indicative of erotic content was *mine*.[5] (Machines, it turns out, might have a sense of humour after all.) The point, here, is that we cannot abstract much

---

[3] Timothy Shortell, 'The Rhetoric of Black Abolitionism: An Exploratory Analysis of Antislavery Newspapers in New York State', *Social Science History*, 28, 1 (2004), 77.

[4] Charles E. Rosenberg, 'Disease in History: Frames and Framers', *Milbank Quarterly*, 67, S1 (1989), 1.

[5] D. Sculley and Bradley M. Pasnek, 'Meaning and Mining: The Impact of Implicit Assumptions in Data Mining for the Humanities', *Literary and Linguistic Computing*, 23, 4 (2008), 410.

meaning from the association of terms in the decontextualised analytic space produced by quantitative techniques applied without any human supervision.

How, then, can we leverage the advantages of a historical network analysis without obscuring critical interpretive issues or wading through statistically significant but meaningless results? Put another way, what are the elements of a good conceptual network analysis?

First, if we want to understand a historical concept in the way that Rosenberg suggests we think about disease, it is not a network of terms in which we are interested but a network of *codes*: concepts that have meaningful interpretations in some context. The concept DEFICIENCY in the context of malnutrition, for example, is not simply equivalent to the word *deficiency*. It also includes related words, such as *insufficient*, *lack of*, *inadequate*, and so forth. Based on the caveats given above, we may also want to distinguish between deficiencies of specific nutrients and general or latent deficiencies, as the two types of inadequacy were often deemed categorically different. Text mining can help identify the seeds of potential codes, but an understanding of the corpus is needed to select appropriate seeds and refine them into useful codes. In other words, the combination of human interpretive understanding with computational pattern recognition can produce better outcomes than either approach alone.

Identifying appropriate codes, then, requires deep engagement with the source material and the context in which it was produced, but a machine's-eye view of the corpus may also reveal patterns that the human eye misses. Thus, a second critical element of a network analytic approach to historical research is that the qualitative and quantitative aspects need to be part of a closed interpretive loop. We might begin with a close reading of a sample of texts, which allows us to develop and refine a set of relevant codes, from which we can develop and validate automated coding algorithms that can be used to code the entire corpus, which enables us to model quantitatively whether, how and to what extent the various codes are associated in various contexts. But critically, we also need to be able to validate the resulting network model, to ensure that the process of coding and modelling does not introduce some interpretive distortion, which may manifest as statistical significance without meaningfulness simply due to the volume of data. In other words, we want to be able to go from the model back to the qualitative data that produced it in a way that allows us to see how the qualitative data correspond to the quantitative model. This triangulation is part of what characterises a *mixed*-methods approach, rather than simply two parallel analyses using different methods.

Lastly, a good network analysis does more than simply visualise patterns of association in the data. A key affordance of network analysis over traditional qualitative research methods alone is that the former can quantify patterns of association, which enables rigorous statistical hypothesis testing and facilitates comparison of different networks. For example, we might want to compare conceptual networks of malnutrition at different points in time, in different national contexts or in different types of source (e.g., professional vs. popular literature), and to determine when the observed differences are significant. To facilitate comparisons such as these on both a qualitative and a quantitative level, it is important that the statistical and graphical properties of the networks correspond in a way that is mathematically and visually meaningful.

It is beyond the scope of this brief discussion to examine more fully – and with concrete examples – the use of conceptual network analysis in historical research, which I will leave for a future paper. It is important to note, however, that the best practices outlined above are of little value if they cannot be implemented, and most network analysis tools are ill suited

to the kinds of conceptual network analyses that historians may want to conduct. In current work, we use a technique known as *epistemic network analysis* (ENA) that is optimised for modelling, visualising and comparing conceptual networks, which are typically small, densely connected networks with a fixed number of nodes.[6] Of particular value, ENA models and visualises networks in a metric space, which enables analysis and comparison of networks both visually and statistically, and the interface allows researchers to see the original data that contributed to any connection in a given network graph, facilitating triangulation between the quantitative model and the qualitative data that produced it.

Network analysis is, of course, only one example of a mixed-methods approach to historical research, and there are certainly more aspects of network analysis worthy of serious discussion by historians. It is our hope that this paper and the others that accompany it, much like the panel from which they emerged, will stimulate further discussion about how we can incorporate new approaches and tools into our historical toolkits in order to better understand the past.

A. R. Ruis and David Williamson Shaffer
University of WisconsinMadison, USA

Bedlam: the Asylum and Beyond, Wellcome Collection, London, 15 September 2016–15 January 2017.

Mike Jay, *This Way Madness Lies: The Asylum and Beyond* (London: Thames and Hudson, 2016), pp. 255, £24.95, hardback, ISBN: 978-0-500-51897-7.

The Wellcome Collection's current exhibition *Bedlam: The Asylum and Beyond*, examines the institutional treatment of mentally ill people in Western culture, taking the metaphor of the asylum as its central motif. The asylum is all but synonymous with Bethlem Royal Hospital, colloquially known as 'Bedlam', the first institution devoted to the treatment of those suffering from mental illness. Bedlam, in both its literal and metaphorical manifestations, functions as the contextual backbone of the exhibition. Its changing material and spatial dimensions (the hospital has occupied four different sites over the course of its history, and still provides mental health care today) are represented though a range of architectural imagery. These reflect shifts in the institution's function, which in turn mirror changing attitudes towards the understanding, care and treatment of mentally unwell people within a broader cultural and historical context.

The vagrancy acts of the eighteenth century formalised the institutionalisation of people deemed 'insane' for the first time. Indeed, the primary role of the asylum was to manage and contain the troublesome presence of mentally unwell people in the interests of maintaining civic order. As historical documents testify, these inmates were often subjected to various forms of inhumane treatment. Remedies for mental illness were restricted to purging and bloodletting, and patients were often physically restrained with

---

[6] For more information about ENA, including user guides, worked examples, and research publications, or to access the tool itself, visit http://www.epistemicnetwork.org/. For a detailed description of the technique, see David Williamson Shaffer, Wesley Collier and A. R. Ruis, 'A Tutorial on Epistemic Network Analysis: Analyzing the Structure of Connections in Cognitive, Social, and Interaction Data', *Journal of Learning Analytics*, 3, 3 (2016), 9–45.