Humanities & Social Sciences Communications

ARTICLE

https://doi.org/10.1057/s41599-024-02619-z OPEN

Computational philosophy: reflections on the PolyGraphs project

Brian Ball[1✉], Alexandros Koliousis[1], Amil Mohanan[1] & Mike Peacey[2]

In this paper, we situate our computational approach to philosophy relative to other digital humanities and computational social science practices, based on reflections stemming from our research on the PolyGraphs project in social epistemology. We begin by describing PolyGraphs. An interdisciplinary project funded by the Academies (BA, RS, and RAEng) and the Leverhulme Trust, it uses philosophical simulations (Mayo-Wilson and Zollman, 2021) to study how ignorance prevails in networks of inquiring rational agents. We deploy models developed in economics (Bala and Goyal, 1998), and refined in philosophy (O'Connor and Weatherall, 2018; Zollman, 2007), to simulate communities of agents engaged in inquiry, who generate evidence relevant to the topic of their investigation and share it with their neighbors, updating their beliefs on the evidence available to them. We report some novel results surrounding the prevalence of ignorance in such networks. In the second part of the paper, we compare our own to other related academic practices. We begin by noting that, in digital humanities projects of certain types, the computational component does not appear to directly support the humanities research itself; rather, the digital and the humanities are simply grafted together, not fully intertwined and integrated. PolyGraphs is notably different: the computational work directly supports the investigation of the primary research questions, which themselves belong decidedly within the humanities in general, and philosophy in particular. This suggests an affinity with certain projects in the computational social sciences. But despite these real similarities, there are differences once again: the computational philosophy we practice aims not so much at description and prediction as at answering the normative and interpretive questions that are distinctive of humanities research.

[1] Northeastern University, London, UK. [2] University of Bristol, Bristol, UK. ✉email: brian.ball@nulondon.ac.uk

Introduction

Philosophy and computing have long and inter-related histories: for instance, the formal investigation of logic was initiated by the philosopher Aristotle over two millennia ago; and it was developments in this field in the late 19th and early 20th centuries that led quite directly to Turing's work, and the invention of the modern electronic computer (Ball and Koliousis, 2022). Nevertheless, while 'humanities computing' (McCarty, 2003) became common within the academy in the early years of the 21st century, philosophers have arguably failed to take full advantage of the opportunities afforded. Why? As Berry (2012) reports, the computing work in such endeavors has often been 'seen as [merely] a technical support to the work of the 'real' humanities scholars' (2012: p. 2). One hypothesis, then, is that philosophers are (or have been) particularly inclined to adopt such a view.

We will not assess this sociological conjecture here—after all, considerable empirical evidence that we do not possess would be required to either confirm or disconfirm it—but we will engage with the objection (concerning the role of computing in relation to the humanities) that underlies it. More specifically, in this paper, we describe a project we are pursuing in computational philosophy (Grim and Singer, 2022) about which the above complaint cannot be raised: the computational and humanities components of our project are thoroughly intertwined; and accordingly, there is no plausibility to the claim that the former play a merely supporting (i.e., non-intellectual) role.

We begin by outlining our project and reporting some of its initial findings. We then compare our approach to those pursued in other projects in the digital humanities and computational social sciences. In this way, we aim to situate the PolyGraphs project's computational philosophy within the intellectual landscape.

PolyGraphs

Background. Our project is entitled 'PolyGraphs: Combatting Networks of Ignorance in the Misinformation Age'. Funded for two years by the British Academy, the Royal Society, the Royal Academy of Engineering, and the Leverhulme Trust under the APEX award scheme, it brings together researchers from a range of disciplines (philosophy, computer and data science, economics) in order to explore information flow in social networks, and the concomitant dynamics of knowledge and ignorance in communities of inquiring agents. The topic is timely, as online misinformation (or even disinformation) about, e.g., coronavirus or the climate emergency can result in ignorance and polarization, preventing effective individual and collective action; and it might be hoped that investigations in this area will influence government and/or corporate policy to combat such pressing practical problems.

Nevertheless, it is worth stressing that our research ultimately has a broader scope than these particular applications suggest. To begin with, there is nothing in our approach that restricts attention to online communities: at its heart, PolyGraphs is a project in social epistemology (Goldman, 1999); and as such it concerns knowledge and ignorance in social contexts more generally, not just those that are technologically (let alone computationally and/or digitally) enhanced. Indeed, our guiding research question may be roughly formulated as follows:

> (Q) How ought we rationally to form opinions (i.e., beliefs), both individually and collectively?

This question is simultaneous: (i) *normative*—it asks not how things *are*, but how they *ought* to be; and (ii) *interpretive*—it requires us to consider how best to understand (e.g.) the notion of rational belief. We will return to these points below.

Also noteworthy for some readers may be our assumption that there are facts of the matter about what the correct answers are to the questions under investigation in the communities that interest us. While opinions may (reasonably) differ, we assume that some are ultimately correct, while others are erroneous. For example, vaccines *are* effective against coronavirus; and climate change really *is* caused by the consumption of fossil fuels—even if there are considerable bodies of (unscientific) opinion to the contrary. This is not to deny that social factors influence which opinions are adopted within a given community—indeed, our investigation explores precisely such influences. Nevertheless, knowledge entails truth—and so belief in a falsehood, whatever the cause, does nothing to alleviate ignorance.

Finally, and relatedly, note that we might seek to explain ignorance within our target communities by appealing to various irrationalities, including psychological 'heuristics' (Kahneman, 2011) that are deployed in everyday information processing and that depart from the ideal, or outright 'intellectual vices' (Cassam, 2018) that infect even conscious reasoning. Instead, we explore the possibility that such ignorance can be (at least partially) explained in terms of the social structures in which individuals are embedded. Even if our approach is somewhat unrealistic in assuming the rationality of the individuals that constitute our target communities, the idealization it involves has two virtues: first, it provides an opportunity to determine whether ignorance can arise, or persist, through no rational fault of the individuals involved; and second, it allows us to address our overarching research question (Q), given above, by exploring the effects of treating various different strategies as candidates for rationality.

The Zollman effect. Our approach involves first modeling, and then simulating, the social processes of opinion formation that interest us. Our basic model derives from economics (Bala and Goyal, 1998): rational agents conduct experiments to obtain new evidence; they share this evidence with their neighbors in the social network to which they belong; and they update their beliefs on the matter under investigation in light of the totality of the evidence at their disposal—including that which is provided by their neighbors. Following others (see below), we conduct philosophical simulations (Mayo-Wilson and Zollman, 2021) based on these models to see how inquiring communities of rational agents behave over time.

Zollman (2007) was the first philosopher to build simulations of the kind we employ. He imagined a community of scientists researching a particular disease, and testing which of two treatments, A or B, is more effective in combating it. It is known in this community that treatment A is effective with a probability of 0.5. Treatment B, however, is effective with probability $0.5+\epsilon$, and the agents need to determine whether ϵ is positive or negative in order to determine whether treatment B is better than treatment A. In fact, ϵ is positive in the models in question, and B is better (in this sense).

The individual scientists in this community are modeled as having a degree of belief, or credence, between 0 and 1 in the proposition that B is better, initially assigned at random from a uniform distribution. Those whose credence is above 0.5 are treated as believing that B is better; they accordingly administer treatment B to their n patients—and in so doing conduct an experiment that provides evidence of the effectiveness of treatment B. In particular, they are able to observe how many of their n patients recover. (It is assumed that recovery is an all-or-nothing affair.). Scientists who think (falsely) that A is better—

that is, those whose credence that B is better is below 0.5—administer treatment A; but as its effectiveness is known, this generates no new relevant evidence about the relative merits of A and B.

The community (of scientists working on this disease) as a whole is modeled as a graph, comprising (a set of) nodes and edges connecting them. The scientists at the nodes share their findings (if any) with those to whom they are connected. They then update their credences in light of the evidence at their disposal—this comprises their own findings, as well as the findings of those who are connected to them. Updating is performed using Bayes' rule:

$$P_f(h) = P_i(h|e) = \frac{P_i(e|h)P_i(h)}{P_i(e)} \tag{1}$$

In other words, the final (or posterior) probability function after updating on the evidence e assigns to a hypothesis h the initial conditional probability of that hypothesis on that evidence—which in turn is related to the other initial quantities as described (by Bayes' theorem).[1]

The entire process described above of performing an experiment (or not, for A believers), informing neighbors of the results (if any), and updating beliefs accordingly, constitutes a single simulation step. It is repeated until either all agents believe that A is better, and so generate no further evidence, or they all have credence above 0.99 in the proposition that B is better, making it exceedingly unlikely that they will go on to change their minds.[2]

Zollman generated his graphs artificially, subject to certain constraints. For example, in some simulations, he specified that the community of scientists should form a 'complete' network, with every node connected to every other by an edge. In others, he stipulated that each scientist should be connected to precisely two neighbors, with the first and last scientists in the network connected to one another as well, and the community as a whole, therefore, constituting a (ring or) 'cycle'.[3] What he found was that: first, more sparsely connected networks such as the cycle are more reliable in converging to the true belief that B is better than more densely connected ones; and second, more densely connected networks are faster at converging to the truth (i.e., they do so in fewer steps), so that there is a tradeoff between speed and accuracy/reliability.[4]

Comparing polarization models. O'Connor and Weatherall (2018) adapted Zollman's approach to accommodate scenarios under which it might be rational to distrust evidence provided by others. In their simulations, scientists update their beliefs using Jeffrey's rule:

$$P_f(h) = P_f(e)P_i(h|e) + P_f(\neg e)P_i(h|\neg e) \tag{2}$$

When the final probability of the evidence is equal to 1, this is equivalent to Bayes' rule; but in general, it allows uncertain evidence e to be discounted, with some weight given to the alternative possibility that $\neg e$. Of course, the amount of discounting applied to a given piece of evidence must be determined somehow—this is not set by the rule itself. O'Connor and Weatherall explore the idea that agents trust others more when they are more alike, and in particular when the absolute difference (or distance d) between their credences is smaller. More specifically still, they run simulations in which the final probability of the evidence e provided by a neighbor is set by the formula:

$$P_f(e) = 1 - \min(1, d \cdot m)(1 - P_i(e)). \tag{3}$$

Here the idea is that evidence is completely believed when it is supplied by someone who has the exact same credence as the agent does (e.g., the agent herself 'reporting' her own experimental findings)—and when the product of the distance between beliefs and the 'mistrust multiplier' m (which serves to amplify the effect of this distance) reaches (and then exceeds, but is replaced by) 1, the new evidence is completely ignored, having no effect on the final credence, leaving it exactly as it was. In between these extremes, the evidence e receives some boost in the agent's credence, but it is not treated as certain.[5]

O'Connor and Weatherall note that, when updating is done as indicated, polarization is a possible outcome: that is, a simulation can end up with some agents having credence above 0.99, while all others have credence below 0.5, yet no further evidence produced by the former will convince the latter to change their mind, since it is completely discounted (i.e., ignored) and $P_f(e) = P_i(e)$. 'In our models,' they report, 'over all parameter values, we found that only 10% of trials ended in false consensus, 40% in true consensus, and 50% in polarization.' (2018: 866) Unfortunately, this aggregate report ('over all parameter values') does not allow us to directly compare the prevalence of ignorance in Zollman's Bayesian models with O'Connor and Weatherall's polarization models in which Jeffrey's rule is employed.

As part of our PolyGraphs project, we built the Python code needed, and ran simulations on complete networks, using both Zollman and polarization models. We then compared: (i) the proportion of simulations (of a given size, and with a given ϵ value) that arrived at the consensus that B is better; and (ii) the number of steps needed to arrive at that consensus in those simulations that did so. We found that, comparing like for like, the models allowing polarization (i.e., those with mistrust multiplier $m > 1$) resulted in a lower proportion reaching consensus in the truth (i.e., more ignorance[6]), and an increase in the number of steps required to do so. Table 1 (below), for instance, shows the percentage of simulations converging to the correct consensus that B is better in relatively small (complete) networks (of size 16 and 64), and with relatively small values of ϵ (0.001 and 0.01), where the Zollman effect was known to occur. As can be seen, polarization models converged to the truth in a smaller percentage of cases, with this effect being more pronounced for larger (values of the mistrust multiplier) m. In short, the more that agents in our simulations distrusted others based on their divergent beliefs, the more ignorance resulted.

As for the number of steps required to arrive at the correct consensus (that B is better) in those that did so, we again found that, when comparing simulations with the same parameter values, ignorance persisted for longer, on average, in the O'Connor and Weatherall models than in the Zollman models. In particular, the number of steps required to achieve an accurate consensus was significantly ($p < 0.05$) greater in (small, low ϵ value) simulations based on the former models than in the latter, whether the mistrust multiplier was 1.1 or 1.5. In short, ignorance took much longer to eradicate in our simulations when agents discounted the (reliable) evidence provided by their peers (Table 2).

Group belief. We have seen that, for O'Connor and Weatherall, polarization is regarded as arising whenever there is a stable departure from consensus. In other words, when all agents' beliefs are stable, the community is polarized (on their account) provided at least one believes that A (has credence < 0.5) and one believes that B (has credence > 0.99). We assume that a group of agents cannot be said to believe something if it is polarized on the issue at hand: and, of course, belief is necessary for knowledge; so we (informally) classed simulations ending in polarization as ones involving ignorance on the part of the community.

Table 1 Ignorance in Zollman vs. polarization models.

Size	Epsilon	Trials	Model	Mistrust	Bs	No. of sims.	%	χ^2-value	*p*-value
16	0.001	16	Zollman		558	600	93		
			Polarization	1.1	300	460	65	128.5	0.00
				1.5	57	460	12.5	691.3	0.00
		64	Zollman		577	600	96		
			Polarization	1.1	300	460	65	172.4	0.00
				1.5	54	460	11.5	766.9	0.00
	0.01	16	Zollman		93	100	93		
			Polarization	1.1	307	460	66.5	26.5	0.00
				1.5	59	460	13	263.0	0.00
		64	Zollman		99	100	99		
			Polarization	1.1	137	460	70	35.1	0.00
				1.5	72	460	15.5	265.1	0.00
64	0.001	16	Zollman		597	600	99.5		
			Polarization	1.1	136	460	29.5	593.7	0.00
				1.5	4	460	1	1027.7	0.00
		64	Zollman		595	600	99		
			Polarization	1.1	127	460	27.5	610.6	0.00
				1.5	3	460	0.5	1023.7	0.00
	0.01	16	Zollman		98	100	98		
			Polarization	1.1	168	460	36.5	122.0	0.00
				1.5	3	460	0.5	520.0	0.00
		64	Zollman		100	100	100		
			Polarization	1.1	226	460	49	85.3	0.00
				1.5	5	460	1	521.1	0.00

We compare the outcomes of the former simulations with those of the latter, for each mistrust value ($m = 1.1$ and $m = 1.5$), using χ^2-tests, and show their significance ($p < 0.05$).

Table 2 Steps to convergence in Zollman vs polarization models.

Size	Epsilon	Trials	Model	Mistrust	No. of sims.	Mean steps	*U*-value	*p*-value
16	0.001	16	Zollman		558	5121		
			Polarization	1.1	300	10,584	138,141.5	0.00
				1.5	57	14,649	27,288.5	0.00
		64	Zollman		577	1287		
			Polarization	1.1	300	2880	144,059.5	0.00
				1.5	54	3223	26,473	0.00
	0.01	16	Zollman		93	53		
			Polarization	1.1	307	117	23,205.5	0.00
				1.5	59	159	4414	0.00
		64	Zollman		99	14		
			Polarization	1.1	323	29	26,313.5	0.00
				1.5	72	43	6137.5	0.00
64	0.001	16	Zollman		597	1710		
			Polarization	1.1	136	4776	72,458	0.00
				1.5	4	6497	2301	0.00
		64	Zollman		595	419		
			Polarization	1.1	127	1018	68,425	0.00
				1.5	3	2089	1765	0.00
	0.01	16	Zollman		98	16		
			Polarization	1.1	169	39	14,978.5	0.00
				1.5	3	116	287.5	0.01
		64	Zollman		100	5		
			Polarization	1.1	226	11	19,281	0.00
				1.5	5	22	4478.5	0.00

We show the results of the Mann–Whitney *U*-tests we ran, again with $p < 0.05$.

The definition of polarization, however, could be strengthened—and the requirements on group belief[7] concomitantly weakened. Thus, whereas O'Connor and Weatherall effectively require consensus before they are willing to say that the community believes that *B* is better, we might consider other accounts of group belief: for instance, it might be thought that a group believes something provided a simple majority of its members do; or provided a supermajority (of e.g., two-thirds, or three-fifths) does. In fact, we are interested in the possibility that whether a group believes something depends not only on how many of its members do so but also on how the members are related to one another—that is, on group structure. Accordingly, we wish to compare methods of aggregating individual beliefs into a group belief that is sensitive or insensitive to structure.

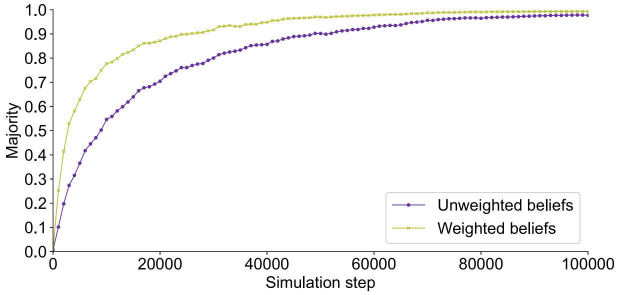

Fig. 1 The size of the majority (i.e., the proportion of votes) believing *B* is better in an EgoFacebook simulation over 100,000 steps. Votes are either *unweighted* (i.e., one node, one vote), or *weighted* to give each node a number of votes equal to its neighborhood size.

It is worth noting that the effects of structure sensitivity are difficult to discern (if they exist) in the kind of small, artificial networks that have so far been our focus. Accordingly, our code is devised in such a way as to allow us to scale our simulations—and we can import large-scale, real-world networks to base them on. We ran our code on one such imported real-world network—though admittedly, EgoFacebook (Leskovec and Mcauley, 2012) is relatively modest, at approximately 4000 nodes.[8] In Fig. 1, we analyze the results from a simulation on this network over 100, 000 steps, looking at what size of majority (i.e., what proportion) of nodes in the network had credence above 0.99 every 1000 steps. In the 'unweighted beliefs' plot, 'voting' is unweighted, so that all nodes count equally. (This is a structure-insensitive aggregation technique.) In the second 'weighted beliefs' plot, the number of votes a node receives is weighted by the size of its neighborhood (i.e., the aggregation method is structure-sensitive in this way). As can be readily seen, the size of the 'majority' increases much more quickly when voting is weighted (reflecting the underlying fact that nodes with larger neighborhoods are reaching a credence of 0.99 more quickly than others are). Thus, if (for example) a three-quarters supermajority of votes is required for group belief, this is achieved (and group ignorance avoided) in less than 10,000 steps with weighting. It is not achieved in the first 20,000 steps without. And, of course, consensus is not achieved for tens of thousands more steps. In short, the aggregation technique matters when it comes to assessing group attitudes—and structure sensitivity in particular makes a difference.

Of course, other structure-sensitive aggregation methods are possible. But is structure sensitivity itself appropriate? In our view, it may well be. Beliefs enter into relations of two kinds—rational and causal. But when edges are undirected (as in EgoFacebook) nodes with larger neighborhoods are both causally more influential (affecting more neighbors) and rationally sensitive to more evidence (from more neighbors)—and their beliefs are therefore arguably more representative of the belief of the network as a whole. In future work, we will disentangle these two elements (causal influence and rational authority), exploring a range of structure-sensitive measures of group belief on large directed graphs.

We conclude this first part of the current paper by briefly summarizing our overview of the PolyGraphs project and its initial findings. We began by describing the models that we employ in our simulations, building on work by Bala and Goyal (1998) and others. We then sketched the Zollman effect, whereby there is a tradeoff between accuracy and efficiency in networks of various densities (Zollman, 2007). Next, we compared O'Connor and Weatherall's (2018) polarization models, in which agents mistrust others, using Jeffrey's (rather than Bayes') rule to discount the evidence provided by those who are unlike themselves. We found that simulations based on these models resulted in more ignorance overall than did those using Zollman's original models; and they took longer to overcome that ignorance, even in those cases in which they ultimately did achieve knowledge. Finally, we motivated the idea that we might wish to look at alternative ways of understanding what it is for a group as a whole to believe something, that does not require consensus, and which may be sensitive in some way to the network structure that is present in the group. We indicated that we will pursue a number of these strands further in future work.

Comparing digital humanities and computational social science

We turn now to the comparison of our approach in the PolyGraphs project with other related practices. We begin by sketching a taxonomy of work in this broad area where the human sciences meet digital technology. We then situate PolyGraphs relative to representative projects in the digital humanities and computational social sciences in turn—and in so doing draw out some of its distinctive features as a computational humanities project.

A taxonomy of approaches. What characterizes the digital humanities—'beyond being an encounter of some sort between the humanities and the digital' (Luhmann and Burghardt, 2022, p. 149)? At one extreme, some thinkers are skeptical, finding 'digital humanities' to be little more than a buzzword that masks poor quality research (Luhmann and Burghardt, 2022, p. 149), while ideological critics think the 'Digital Humanities appeal to university administrators, the state, and high-rolling funders because it [sic.] facilitates the implementation of neoliberal policies' (Neilson et al., 2018, p. 4), replacing socially progressive academic work with employment-oriented training. At another extreme (Luhmann and Burghardt, 2022, p. 149), there are those who hold that, presumably due to a certain methodological superiority, digital humanities will ultimately encompass or replace all work in the humanities.

We come not to evaluate the digital humanities, however, but to understand them—and to use that understanding to situate the approach taken in the PolyGraphs project. To this end, we suggest that a broad division of work in the area of the above 'encounter' can be effectuated based on *what* is being investigated and *how*. Thus, some research uses computational methods to address questions of traditional interest within the human sciences, while other work uses the techniques of these latter sciences, and takes some aspect of the digital realm as its object of inquiry. We can further distinguish, within the first category above, the digital humanities properly so-called on the one hand, from the computational social sciences (Lazer et al., 2009) on the other. The result is a three-way classification of work in this area, which is admittedly rough and ready, with fuzzy boundaries between categories, and some research projects no doubt displaying elements of more than one type of work. Nevertheless, we believe it will prove helpful in what follows.

Roth (2019) similarly discerns three kinds of work in this broad area of investigation—a fact that lends support to our analysis. Roth writes:

> The perhaps most widespread acceptation of 'digital humanities' relates to the creation, curation, and use of digitized datasets in human sciences and, to a lesser extent, social sciences. In broad terms, these approaches include the development and application of computer tools to, inter alia, digitize, store, process, gather, connect, manage, make available, mine, and visualize text collections and corpuses,

> image banks, or multimedia documents of various origins (2019: p. 616).

Roth uses the term 'digitized humanities' in connection with work of this kind. Nevertheless, it is this that we will be focusing on when we speak of the digital humanities—work that employs digital methods in service of academic goals that might be recognized by the traditional humanities disciplines.[9]

By contrast, according to Roth, researchers of a second kind 'develop mathematical frameworks and computer science methods with the specific goal of formalizing and stylizing some systematic social processes' (2019: 617–618), e.g., by building social simulations, or employing agent-based modeling. Here, she says:

> datasets are not anymore exploited as singular recordings corresponding to given empirical case studies, but simply as exemplar instances of a much wider and, more importantly, interchangeable phenomenon. This approach is not dissimilar from the one usually ascribed to natural sciences, in that [researchers] seek… general laws rather than local patterns' (2019: p. 618).

But Roth notes that 'in practice, [work of this kind] generally builds more often on social science research issues than humanities' (2019: 618): thus, whereas she speaks of the 'numerical humanities', we follow Lazer et al. (2009) in referring to this and related work as 'computational social science'.[10]

Finally, the 'humanities of the digital' as Roth calls the third category of work, 'focuses on computer-mediated interactions and societies, such as the Internet and other online communities' (2019: p. 623). This may suggest a relatively restricted field, including only, e.g., work on human-computer interaction and/or the philosophy or sociology of technology; though we propose that any work employing the methods of the humanities or social sciences that makes the digital into the object of inquiry is of this character. Work of this third kind is of considerable interest: Roth herself, for instance, concludes by 'insist[ing] on the possible broker role of the. humanities of the digital bridging the gap between digital humanities and numerical humanities' (2019: p. 629).

Our proposed threefold taxonomy can accommodate other (e.g., historical) accounts of work in this area. Berry (2012), for instance, suggests three periods (or 'waves') in the development of the digital humanities. In the first wave, traditional objects of humanistic inquiry were digitized, allowing them to be explored using computational techniques. In the second, humanists turned their attention to an expanded range of cultural artefacts, including those that were 'born-digital' (2012: p. 4). Berry then suggests 'a tentative path for a third wave of the digital humanities, concentrated around the underlying computationality of the forms held within a computational medium' (2012: p. 4) One might expect that this 'computational turn' (Berry, 2012: p. 4) would be akin to Roth's 'numerical humanities'; but in fact it appears to be closer to her 'humanities of the digital'—for Berry says that in this endeavor, 'code and software are to become *objects* of research for the humanities and social sciences, including philosophy' (2012: p. 17, our emphasis). In short, the methods of the human sciences are used to investigate digital/computational objects in the third wave (as in, e.g., explorations of algorithmic bias).[11] Meanwhile, work in Berry's first two waves is clearly of the 'digitized humanities' variety. The computational social sciences are simply ignored.

Given that it is well-suited to the task (e.g., successfully subsuming Berry's divisions), in what follows we deploy our threefold taxonomy, with its similarities to that of Roth (2019), in order to compare the approach of the PolyGraphs project with other, related practices. We set aside the humanities and social sciences of the digital as involving a fundamentally different sort of encounter between the digital and the humanities than the other two, and one that is broadly irrelevant to our current purpose of situating the approach taken in the PolyGraphs project.[12] This leaves us with two comparisons to make, which we undertake in turn: first, with the digital humanities; and then, with the computational social sciences.

Digital humanities and PolyGraphs. For better or for worse, philosophers have not, it seems to us, been ready adopters of the methods employed in the digital humanities. We suspect that there are two central reasons for this. First, philosophers do not typically think of the subject matter of their discipline as consisting primarily of texts (or other human artifacts, such as images). Insofar as texts are investigated in philosophy, this is in order to glean insights into the true subject matter of the field, which is—for want of a better phrase—the human condition; that is to say, at least roughly, various aspects of human experience, the nature of the world we navigate, and how this affects us (both morally and cognitively/epistemically). This leads to the second point. For, insofar as the techniques of the digital humanities are oriented towards the investigation of digital artifacts (e.g., texts) and/or repositories (e.g., journal archives), their investigation may be thought to be at best incidentally related to, and ultimately separable from, philosophical inquiry properly so-called. In short, digital activities may appear to be simply grafted onto a humanistic one. Allow us to give a representative example of where this charge might be levied—whether fairly or not.

Alfano (2018) aims to 'explain a synoptic Digital Humanities approach to Nietzsche's interpretation and demonstrate its explanatory value' (2018: p. 86). In particular, Alfano is interested in Nietzsche's views on moral psychology, and specifically how he employs the notions of drive, instinct, and virtue; and he explains, in effect, that after choosing these notions to focus on, he then operationalizes them with words and word stems that are expressive of them, searches a repository of Nietzsche's texts for occurrences of those textual elements, cleans, analyzes, and visualizes the data he obtains, and then engages in a close reading of relevant passages in Nietzsche's work that are revealed by that data. As a result of his research, he concludes that, for Nietzsche: (i) instincts and virtues are kinds of drives; (ii) drives are dispositions to perform particular action types; and (iii) drives cannot be easily changed.

It is perhaps worth remarking that in this case, even if there is a broader interest in whether Nietzsche's moral psychology is ultimately correct (and so in human nature—i.e., an aspect of the human condition), the immediate object of investigation *is* a body of texts, namely the corpus of Nietzsche's writings. For this reason, the techniques of the digital humanities are perhaps especially well-suited to the investigation at hand (whereas they might not be appropriate for other philosophical projects). Nevertheless, there is a way of thinking about the project as described in which the specific digital techniques employed are ancillary to the central interpretive work that constitutes the proper humanistic investigation. In effect, there is some 'humanities computing' that plays a supporting role in allowing Alfano to identify passages in Nietzsche's writings to look at; and he then engages in the proper philosophical work of interpreting those passages (through 'close reading'). From this perspective, the ('tech support') role played by the digital element of the project is not unlike that played by a steam-powered train in getting a 19th-century researcher to the library—it may enhance efficiency, but is hardly integral, or essential, to the intellectual work it supports.

This is no doubt an unfair characterization of Alfano's project, and of the variety of digital humanities work it is here representing. For one thing, part of the *argument* for the interpretation given concerns the distribution over time of the keywords that express the target notions, and this distribution is discerned through the digital humanities techniques employed. Nevertheless, it is safe to say that the role of the computational methods employed in PolyGraphs is unlike that described in this caricature: they are certainly not dissociable from the intellectual work of the research in which we are engaged.[13] Our simulations generate evidence that bears directly on philosophical questions. What might happen if a community of agents conducted an inquiry in the manner specified in one of our models? Would knowledge be achieved within the community? Or would ignorance persist? These are questions that interest philosophers—and the computations performed in our simulations are integral to our attempts to address them, not mere *addenda* to those inquiries.

It is perhaps worth commenting on one further point in connection with the digital humanities, before comparing PolyGraphs to work in the computational social sciences. We have hitherto focused on the use of digital techniques in the early stages of research—roughly, in (or as preparatory to) investigation. But as Neilson et al. (2018) point out, some think of the digital humanities as disciplines 'in which students and faculty make things, not just texts' (2018: p. 3). In this 'maker turn' (2018: p. 7), as they call it, 'publicly available Digital Humanities projects are often part of the demand to retain ownership over one's work, disseminate information freely, and reach audiences outside of the university.' (2018: p. 7) Indeed, they note that some in this camp (e.g., futurists) hold that 'critique now takes place through the design and implementation of new systems' (2018: 7). In this way, those supporting the maker's turn might be thought to address the charge of regressive neoliberal appeasement discussed above—on the contrary, it is the digital humanities that are progressive, possibly even revolutionary!

As an example of a project that might be thought to exhibit some of these characteristics, consider Slave Voyages, described on its website as 'a collaborative digital initiative that compiles and makes publicly accessible records of the largest slave trades in history'.[14] This is a valuable (and progressive) project, and we ourselves have learned important truths from engaging with it. Nevertheless, it may strike (certain) philosophers that the digital elements here are incidental to the research. In particular, the digital outputs produced—e.g., the two-minute video of Kahn and Bouie (2021) depicting the voyage of each ship carrying slaves across the Atlantic over a 315-year period—may be thought to primarily facilitate the dissemination of findings, rather than being integral to the research.

Allow us to elaborate on this line of thought. If research is a structured activity aimed at the production of knowledge, then whether that knowledge is disseminated in journal articles or in some other way is not directly relevant to that research. Philosophers in particular may be inclined to hold that propositional, or declarative knowledge (i.e., knowledge *that*)—rather than either texts or other artifacts —is what research aims to produce. Arguably, such knowledge is most naturally expressed linguistically (rather than, say, graphically, or in terms of images); but there is no inherent reason why it should be expressed in English, for example, rather than French—and so there is no special connection to texts, any more than there is to, e.g., videos. (We might compare Socrates here, who famously never made any of his philosophical contributions in writing.) Philosophers may even be inclined to go so far as to isolate the propositions known as a result of inquiry from the actual knowing of them by specific individuals.

Again, without assessing the merits of this philosophical line of argument, we simply stress that the computational elements in PolyGraphs are not merely supporting dissemination. It is true that we are producing data visualizations as part of the project, and we are releasing the code that performs our simulations on GitHub. The former, we hope, will facilitate the communication of our findings; and the latter constitutes a piece of digital infrastructure that may allow others to conduct further research and obtain new findings. But at its core, PolyGraphs is a computational humanities project (as we will see). How this compares to a project in the computational social sciences is a question to which we now turn.

Computational social sciences and PolyGraphs. PolyGraphs employs models and seeks generalizations, just as certain computational social science projects do. Indeed, the models of information sharing at its heart derive from the social science of economics (Bala and Goyal, 1998); and as we have emphasized, even when we apply them to online social networks (as in our analysis above of the EgoFacebook network), our findings should generalize beyond any such particular application to illuminate the phenomena of social epistemology more broadly. Nevertheless, the computational philosophy we practice aims not so much at empirical description and prediction as at answering the kinds of normative and interpretive questions that are distinctive of humanities research.

Computational social science projects typically aim to achieve empirical validation through descriptive accuracy and/or predictive success about some social phenomenon—e.g., the rate at which fake news articles spread on social media. However, they often involve highly simplified 'agents'—for instance, ones whose actions are restricted to either sharing/re-tweeting a story or not (Menczer and Hills, 2020). Plausible causal mechanisms—such as attentional overload (Weng et al., 2012)—may be identified; however, the nodes of the networks in these studies cannot be readily regarded as occupied by human subjects, with beliefs and desires of their own, who may behave rationally or not.[15]

By contrast, PolyGraphs is concerned with precisely such issues. Can individual agents plausibly be interpreted as having credences that they update using Bayes' rule? Ought they to use Jeffrey's rule instead? PolyGraphs addresses these (and other) interpretive and normative questions. For instance: are we able to understand collective action in terms of group attitudes—including beliefs? If so, how ought groups to aggregate their attitudes from those of their members? Such questions are paradigmatically humanistic—and we use computational techniques (specifically, simulations) to investigate them. In other words, PolyGraphs is a humanities project with a computational methodology.

In comparing PolyGraphs to research in the computational social sciences, we have stressed both the character of the questions involved and the corollary that validation is not straightforwardly empirical.[16] There has been some recent discussion of modeling in philosophy which may illuminate these points. Thus, Williamson (2017), for example, notes that in the natural and social sciences models are often tested by way of measurable quantities and that this is not possible for (at least some) models in philosophy. However, he stresses that scientific models are also sometimes tested through qualitative predictions —and that philosophical models can and do yield such predictions. Crucially (from our point of view), when it comes to qualitative distinctions of category, some judgment may be required to apply them—and thereby gain the 'model-independent knowledge of the target phenomenon' that, as Williamson notes, is required for the testing of those models. In our case, for instance, the prediction of a given model (using Bayes' or Jeffrey's rule) might be that a community of rational agents in certain specific circumstances that aggregates its members' attitudes in some particular way will be ignorant (rather than knowledgeable) of the fact that treatment B is better than treatment A after exposure to this or that course of evidence. If we can independently ascertain whether that would indeed be the case, we can use this knowledge to test our model's assumptions surrounding the nature of (individual and group) rationality (e.g., whether the update and aggregation rules it employs are the ones that ought to be used in a community of that kind in those circumstances).[17] But of course, the categorical difference between knowledge and ignorance is quite high-level, and not 'observational': an exercise of judgment is required in order to determine how to apply it in a given case.[18]

We have emphasized not only that our investigation employs modeling, but also that it addresses normative questions. In recent work, Titelbaum (manuscript) discusses normative modeling. He suggests that normative models are distinguished from descriptive models by the character of the facts they aim to capture—namely, normative, rather than descriptive, facts.[19] We note, however, that such normative facts cannot be simply 'observed'. Yet perhaps this point is more readily made in connection with the account of normative modeling given by Colyvan (2013). 'Normative models, Colyvan notes, 'are not supposed to model actual behavior or explain actual behavior; rather, they are supposed to model how agents *ought* to act.' (2013: p. 1338, emphasis original) Since, unlike actual behavior, how agents ought to act (including what opinions they ought to form) cannot be directly detected by empirical methods, normative models (including, arguably, those we employ) cannot be validated (or refuted) through overly simplistic ('positivistic') appeals to empirical evidence. The judicious exercise of judgment is required.

In comparing PolyGraphs with other projects in the computational social sciences, we have attempted to show that, while there are similarities in approach, subtle differences remain. Our computer simulations rely on (what are intended to be) generally applicable models, but the models involved are arguably normative in character, and accordingly cannot be tested in a flat-footedly empirical manner. We have argued that this befits the humanistic nature of our inquiry.

Conclusion

We began with an overview of the PolyGraphs project, covering prior results (the Zollman effect), and comparing polarization models (due to O'Connor and Weatherall), before briefly considering (our innovative, structure-sensitive approach to) group belief. We then gave a three-way distinction amongst aspects of the 'encounter' between the digital (on the one hand) and the humanities and social sciences (on the other). In particular, we distinguished digital humanities, computational social science, and the investigation of the computational and digital using the methods of the human sciences. We argued that whereas some digital humanities projects (appear to) merely append some computational elements either before or after a thoroughly humanistic investigation, in PolyGraphs the computational elements are integral to the research itself. But in contrast to certain computational social science projects, the research questions in PolyGraphs are both normative and interpretive in character. In short, PolyGraphs is a computational humanities project.

Data availability

The datasets generated and analyzed during the current study are available in the GitHub repository, as is our source code.

Received: 18 March 2023; Accepted: 8 January 2024;
Published online: 31 January 2024

Notes

[1] P_i is the initial probability function (prior to update), P_f the final probability function (afterwards).

[2] This is the stopping condition we have employed in our simulations, following O'Connor and Weatherall (2018). Zollman himself originally required B believers to have credence above 0.9999 (2007: 579); and in Zollman (2010) he allowed simulations to stop after 10, 000 steps. The simplification in the text does not affect our discussion.

[3] He investigated various further network structures as well.

[4] Subsequent work by Rosenstock et al. (2017) found that these results held only for relatively small networks, with small numbers of (patients, or more generally) trials, and small values of ϵ. Nevertheless, in such cases, Zollman's two findings were confirmed—and of course, many social epistemological phenomena are approximated by the (small) parameter values in question (e.g., those involving families, committees, or scientific communities with limited evidence-gathering resources).

[5] This means that there is no 'anti-updating'—receiving the uncertain evidence that e never makes an agent give *eless* credence than they previously did. Discounting without anti-updating might be an appropriate attitude to take towards 'bullshitters'—cf. Frankfurt (2005). With known liars supplying one's evidence, by contrast, anti-updating might be appropriate. While O'Connor and Weatherall explore an implementation of Jeffrey's rule with anti-updating, we do not consider it here.

[6] Knowledge requires justified true belief. The consensus belief that B is better is true when it arises in our simulations—and the agents involved update their beliefs in a rational manner, based on the evidence available to them, so their beliefs are justified. Thus, we here treat the consensus that B is better as group knowledge, and its absence —whether through error (false consensus) or omission (e.g., through polarization)—as group ignorance.

[7] more careful discussion would distinguish (i) what a group believes from what its members (ii) severally and (iii) collectively believe (Ball, 2021). We do not believe the neglect of this distinction in the main text affects our central points here.

[8] In future work we intend to run our simulations on much larger, real-world networks; but the EgoFacebook graph discussed in the main text already suffices to make our main point here.

[9] This chimes with Luhman and Burghardt's finding—based on a computational analysis of research articles across a range of journals—that 'textual data. continue [sic.] to be the predominant object of study in DH' (2022: p. 167).

[10] Luhman and Burghardt identify Roth's numerical humanities with what they call 'computational humanities'—which, they say, 'approaches humanities research questions through computational models' (2022: 149). This would be an apt description of the PolyGraphs project—but when we look, for instance, at the website for the research group Burghardt leads, we are told that humanities computing asks, inter alia, 'How can humanities data—which is traditionally interpreted in an idiographic, hermeneutic way—be modeled in a way it becomes available for computational, empiric analyses?' (See https://www.mathcs.uni-leipzig.de/en/ifi/research/computational-humanities. Accessed: 14/02/23.) As we will see below, the computational approach to philosophy practiced on the PolyGraphs project preserves a role for interpretation (and, indeed, normativity); and the quote above in any case seems to suggest only a more computationally sophisticated/intensive version of Roth's digitized humanities.

[11] Note that there may also be a hint here of a connection to the maker turn discussed below. As we indicated, our proposed taxonomy is rough and provisional, with some projects lying between, or even spanning boundaries.

[12] That said, the present work may belong to this third category (even if PolyGraphs itself does not).

[13] We do not wish to suggest that it is *only* in PolyGraphs that computational work is integral to philosophical investigations. Many others have done work with this character—for just a few examples, in addition to the work by O'Connor, Weatherall, Zollman, and others discussed above, see e.g., Hegselmann and Krause (2002); Mayo-Wilson (2014); Olsson (2013); Pollock (1989); Skyrms (2010). For an overview of work in this area, see Grim and Singer (2022).

[14] See https://www.slavevoyages.org/. Accessed: 14/02/23.

[15] In a similar spirit, Lazer and co-authors note that, amongst thousands of recent papers drawing on the platform's data, 'the large majority of Twitter research is making inferences about accounts or tweets; very little of Twitter research can reasonably claim to be making statements about the behaviors of humans' (Lazer et al., 2021, p. 191). But even 'very detailed agent-based approaches' (Balcan et al., 2009, p. 21848) in the computational social sciences, which do tell us about the behavior of human beings, often fail to illuminate personal level motivations that would allow us to regard those behaviors as actions. Instead, we get, e.g., 'realistic estimates of population mobility' (Eubank et al., 2004, p. 180). This may be appropriate given the research aims—in this case understanding 'the relative merits of several proposed mitigation strategies for smallpox spread' (Eubank et al., 2004, p. 180). Our point here is simply to contrast the impersonality of such research with that undertaken in PolyGraphs.

[16] In her influential discussion of the humanities, Small (2013) likewise notes that they 'focus... on interpretation and critical evaluation' (2013: 23) and involve 'an ineliminable element of subjectivity' (2013: p. 23). Specifically, on this last point, she claims that there is a need in humanities research for an exercise of judgment rather than 'positivistic appeals to evidence' (2013: p. 23). We take this to vindicate our claims in the main text—particularly once it is recognized that 'critical evaluation' is what ultimately underpins normative assessment.

[17] That ignorance is (epistemically) *worse* than knowledge is an evaluative claim—and therefore relevant to (strictly) normative questions about how agents ought to behave in relation to opinion formation.

[18] Indeed, judgment can sometimes be required even to determine whether supposedly 'observational' categories apply: think of the task of determining whether some color sample that borders on being orange counts as red.

[19] Titelbaum explains that the normative facts, in his view, may be general or particular, and include both prescriptions and evaluations. Others—e.g., Dietrich and List (2017)—regard the normative as strictly distinct from, though related to, the evaluative. We incline slightly towards this latter view, but we do not think anything of significance turns on the issue here.

References

Alfano M (2018) Digital humanities for history of philosophy: a case study on Nietzsche. In: Neilson T, Levenberg L, Rheams D (eds.) Research methods for the digital humanities. Springer, Cham, pp. 85–101

Bala V, Goyal S (1998) Learning from neighbours. Rev Econ Stud 65:595–621

Balcan D (2009) Multiscale mobility networks and the spatial spreading of infectious diseases. Proc Natl Acad Sci USA 106:21484–21489

Ball B (2021) Groups, attitudes and speech. Analysis 81:817–826

Ball B, Koliousis A (2022) Training philosopher engineers for better AI. AI Soc 38:1–8

Berry DM (2012) Introduction: understanding the digital humanities. Palgrave Macmillan UK, London. pp. 1–20

Cassam Q (2018) Vices of the mind: from the intellectual to the political. Oxford University Press

Colyvan M (2013) Idealisations in normative models. Synthese 190:1337–1350

Dietrich F, List C (2017) What matters and how it matters: a choice-theoretic representation of moral theories. Philos Rev 126:421–479

Eubank S (2004) Modelling disease outbreaks in realistic urban social networks. Nature 429:180–184

Frankfurt H (2005) On bullshit. Princeton University Press

Goldman AI.(1999) Knowledge in a social world. Oxford University Press, Oxford, England

Grim P, Singer D (2022). Computational philosophy. In: Zalta EN, Nodelman U (eds.). The Stanford Encyclopedia of Philosophy. Metaphysics Research Lab, Stanford University, Fall 2022 edition

Hegselmann R, Krause U (2002) Opinion dynamics and bounded confidence models, analysis and simulation. J Artif Soc Soc Simul 5. https://jasss.soc.surrey.ac.uk/5/3/2.html

Kahn A, Bouie J (2021) The Atlantic slave trade in two minutes. Available at https://slate.com/news-and-politics/2021/09/atlantic-slave-trade-history-animated-interactive.html. Accessed: 14/02/23

Kahneman D (2011) Thinking, fast and slow. Macmillan

Lazer D (2021) Meaningful measures of human society in the twenty-first century. Nature 595:189–196

Lazer D (2009) Computational social science. Science 323:721–723

Leskovec J, Mcauley J (2012) Learning to discover social circles in ego networks. Adv Neural Inf Process Syst 25. https://papers.nips.cc/paper_files/paper/2012

Luhmann J, Burghardt M (2022) Digital humanities–a discipline in its own right? an analysis of the role and position of digital humanities in the academic landscape. J Assoc Inf Sci Technol 73:148–171

Mayo-Wilson C (2014) Reliability of testimonial norms in scientific communities. Synthese 191:55–78

Mayo-Wilson C, Zollman K (2021) The computational philosophy: simulation as a core philosophical method. Synthese 199:3647–3673

McCarty W (2003) Humanities computing. Encyclopedia Libr Inf Sci 2:1224

Menczer F, Hills T (2020) The attention economy. Sci Am 323:54–61

Neilson T, Levenberg L, Rheams D (2018) Introduction: research methods for the digital humanities. Res Method Digit Humanit 1–14

O'Connor C, Weatherall JO (2018) Scientific polarization. Eur J Philos Sci 8:855–875

Olsson E (2013) A Bayesian simulation model of group deliberation and polarization. In: Zenker F (ed.) Bayesian argumentation, Springer

Pollock J (1989).How to build a person: a prolegomenon. MIT Press. Cambridge, MA: MIT Press

Rosenstock S, Bruner J, O'Connor C (2017) In epistemic networks, is less really more? Philos Sci 84:234–252

Roth C (2019) Digital, digitized, and numerical humanities. Digit Scholarsh Humanit 34:616–632

Skyrms B (2010) Signals: evolution, learning, and information. Oxford University Press, Oxford, England

Small H (2013) The value of the humanities. Oxford University Press

Titelbaum MG (manuscript) Normative modeling

Weng L, Flammini A, Vespignani A, Menczer F (2012) Competition among memes in a world with limited attention. Sci Rep 2:335

Williamson T (2017) Model-building in philosophy. Philosophy's future: the problem of philosophical progress. John Wiley & Sons, Inc. pp. 159–171

Zollman KJ (2010) The epistemic benefit of transient diversity. Erkenntnis 72:17–35

Zollman KJS (2007) The communication structure of epistemic communities. Philos Sci 74:574–587

Acknowledgements

This work was supported by the British Academy, the Royal Academy of Engineering, the Royal Society, and the Leverhulme Trust, under the APEX Award scheme, grant number APX\R1\211230, and by Northeastern University's NULab for Text, Maps, and Networks. The authors also gratefully acknowledge support for this publication from the Digital Academy at the Academy of Sciences and Literature Mainz, Germany.

Author contributions

AK developed the simulation framework. All authors made substantial contributions to the conception or design of the work and/or the acquisition, analysis, and interpretation of the data. They contributed to drafting the work or revising it critically for important intellectual content and have given final approval of the version to be published. They agree to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

Competing interests

The authors declare no competing interests.

Ethical approval

This article does not contain any studies with human participants performed by any of the authors.

Informed consent

Not applicable.

Additional information

Correspondence and requests for materials should be addressed to Brian Ball.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.