The Social Life of Data Clusters: The Potential of Sociomaterial Analysis in the Critical Study of Educational Technology

Carlo Perrotta

School of Education, University of Leeds, c.perrotta@leeds.ac.uk

Abstract

This paper draws on Actor-Network Theory to argue that methods used for the classification and measurement of online education are not neutral and objective but are involved in the creation of the educational realities they claim to measure. The paper examines Cluster Analysis (CA) as a 'performative device' that, to a significant extent, creates the educational entities it claims to objectively and neutrally represent through the emerging body of knowledge of Learning Analytics (LA). In the conclusion, the paper suggests that those concerned with social justice in educational technology need not limit themselves to denouncing structural inequalities and ideological conflicts. At the opposite end of the 'critical spectrum' there is the opportunity to analyse in a more descriptive fashion how hegemonic discourses in education are legitimated through techniques and devices.

Keywords

Cluster Analysis, Learning Analytics, Actor-Network Theory.

Introduction

In this era of austerity and generalised uncertainty, education is also undergoing a profound crisis where widening rifts across the whole spectrum of social justice are construed as inevitable. A new hegemonic form of 'business as usual' has taken hold to which, we are led to believe, there are no viable or even responsible alternative (Peters & Bulut, 2011). In this paper, I argue that technology and methods of data analysis are instrumentally implicated in this hegemony and that alternative narratives are needed. Yet, faced with the complexity of the task at hand, there is a real risk to fall into the traditional quagmire of social critique, whereby established conceptual categories and discursive strategies lead to familiar arguments or claims which, irrespective of their moral worth, struggle to persuade and to cause visible ripples across the broader academic community, let alone the public discourse (Latour, 2004). There lies the real challenge that critically minded academics face: to reclaim relevance and legitimacy in the face of the instrumentalist and utilitarian consensus of the globalised 'education market'.

This is, first and foremost, a methodological challenge concerned with the ways in which claims are construed and given authoritativeness in the networked discourse of global academia, with its range of mediating factors and arbiters: funding bodies, ranked journals, league tables and so forth.

A few years ago, Neil Selwyn argued that research in educational technology needs to develop more vigorously along social scientific lines (Selwyn, 2010). It is a hard to ignore exhortation in the context of this symposium – not least because social science has traditionally provided the conceptual and methodological means to 'expose' the contradictions of modernity. This is exemplified in the more sociological schools of thought which over the past two centuries have accumulated a portentous arsenal for critical thinking and argumentation, assisted by empirical and conceptual categories that have become unavoidable referents for anybody with a 'social scientific' interest in matters of equality or social justice. Two things effectively symbolise such influence and continued relevance: the reality of social and geopolitical stratification and the dialectic view of collective interactions, whereby 'antagonism' (between classes and social strata, between geopolitical conditions, and above all between those who own the means of production and the exploited) is still a very effective framing to explain inequality and conflict in most social arenas (e.g. Fuchs, 2010).

However, embracing social science means also acknowledging its own crisis and its struggle for authoritativeness which became apparent at the turn of the century. I am referring in particular to recent calls for a 'methodological renewal' spurred by a perceived obsolescence of the main methods in social research, with

ISBN 978-1-86220-324-2

their susceptibility to the risks of self-reporting and interviewer effects, and the related invitation to examine the large amounts of 'data trails' left behind by people as they go about their routine transactions mediated and recorded by digital devices (Savage & Burrows, 2007). The initial provocation by Savage and Burrows developed into an ambitious programme of scholarship at the Centre for Research on Sociocultural Change (CRESC), which set out an agenda to respond to the 'exhaustion' (Savage, 2013: 8) of cultural theory and the growing theoretical traction of sociomateriality, viewed as a framework to study how digital devices and data are 'simultaneously shaped by social worlds, and can in turn become agents that shape those worlds' (Ruppert, Law & Savage, 2013: 31).

Building on the recent enthusiasm for Actor-Network Theory (ANT) and its subsequent elaborations an growing number of social scientists are interested in the study of digital methods and devices and keen to examine their role in a 'new ontology of the social', in which technologies and analytic methods are 'materially implicated in the production and performance of contemporary sociality' (ibid: 34).
Against this background, it is increasingly hard to ask questions about educational equality from a social scientific point of view, without a concern for how the social and the technical become entangled in the reproduction of the world with all its contradictions. The task is made even more arduous by the fact that critically minded educational researchers (the author of this paper included) are still drawn to hermeneutic accounts of learning and education, given the proven potential of qualitative methods to interrogate forms of educational consensus. We also cannot ignore that that these methods give empirical currency to the lived experiences of those who are often at the margins of mainstream education. Although hermeneutic research still has a crucial role to play, there is no denying the need for a complementary empirical language to attend to the current socio-technical reality of national and supranational education, in which technology and data are often being recruited to authorise the same forms of governance, control and surveillance which are heavily implicated in the reproduction of educational inequalities (Ozga, 2009).

Any attempt to challenge this consolidating hegemony needs to engage in an informed fashion with the epistemic assumptions that underpin instrumentalist readings of education and technology: how is knowledge in the 'global education marketplace' created, and how are technologies, devices and data involved in its legitimisation? The first step in answering these questions is to reject the view of methods and 'digital data analysis' as 'pure' technical devices, and subject those very methods and devices to a more rigorous form of study: methods and tools must become objects of inquiry. The aim of this paper is to provide an example of how this can be achieved in the context of networked learning.

In the remainder of the paper, I will examine Cluster Analysis (CA) as a 'performative device' that, to a significant extent, creates the educational entities it claims to objectively and neutrally represent through the emerging body of knowledge of Learning Analytics (LA). In doing so, I aim to problematize the process through which a particular version of networked learning is being created and reproduced. Cluster Analysis will be considered as an 'apparatus': an assemblage made of networks of expert knowledge, technologies and algorithms that translates clusters of digital data about learners into socially negotiated 'materializations' - what Latour (1998) has described as an expression of the 'traceable social that is rendered visible'. My attempt is also indebted to the work of Callon, Millo and Muniesa (2007) on 'market devices', in which a collection of technologies and data analysis tools (such as index-based derivatives and pricing techniques) were considered as objects of sociological inquiry, and as 'material and discursive assemblages that intervene in the construction of the market' (Muniesa, Millo and Callon, 2007: 2).

In a similar fashion, I contend that methods used for the classification and measurement of online education are not neutral and objective but are, to varying degrees, involved in the creation of the educational realities they claim to measure. This social construction is the result of an epistemic negotiation across heterogeneous networks of people, organisations, technologies and analytic techniques - a process in which methods operate as 'inscription devices' that turn sometimes nebulous and open-to-interpretation 'learning phenomena' into easily readable materialisations (Latour & Woolgar, 1986), which are then treated as real and consequential. This view draws in equal measure on Actor-Network Theory and Bowker and Star's influential analysis of how classificatory systems become embedded in institutional settings, acquiring taken-for-granted, almost invisible qualities and contributing to shape those very settings with significant consequences for the people and the objects being classified (Bowker & Star, 1999). The key aspect in this process is the political nature of classification procedures, which always serve more than one purpose or group and, as such, should never be reified but kept open to contestation and re-formulation.

ISBN 978-1-86220-324-2

For example, the case of international benchmarking in education can be thought of as a high profile example of how an epistemic network tries to advance a specific version of data-intensive global education. In this network, The Organisation for Economic Co-Operation and Development (OECD), mainly through its Programme for International Student Assessment (PISA), acts as the chief epistemic guarantor in a global configuration of education ministries, interest groups and policies – a configuration organised around a set of 'core' transcultural educational values, mainly the 'real life application' of literacy and numeracy skills, 'inscribed' through standardised, 2-hour testing sessions performed in more than 70 countries. Although of great relevance to the argument discussed here, this network is also very complex and overly reliant on a single mediator whose weight and influence needs to be dealt with caution and with an analytic depth simply not possible in this paper. I shall focus instead on a 'closer-to-home' example: the emerging field of Learning Analytics (LA) in the context of e-learning and Massively Open Online Courses (MOOCs).

Learning analytics and Cluster Analysis

Learning Analytics (LA) has been described as the 'measurement, collection, analysis, and reporting of data about learners and their contexts, for the purposes of understanding and optimizing learning and the environments in which it occurs'[1]. A key concern in LA is the use of insights derived from data to generate 'actionable intelligence' to inform tailored instructional interventions (Clow, 2013; Cambpell, Deblois and Oblinger, 2007). According to Clow (2013), LA is less a solid discipline than an eclectic 'jackdaw' approach that picks up methods from other areas, as long as they serve its overarching pragmatic aims. The field is trying to establish a distinctive academic identity in relation to a range of contiguous approaches which use the same techniques but for more reductionist, managerial or 'educationally irrelevant' purposes. These approaches include business intelligence (which is becoming fairly established in the HE sector), machine learning, web analytics and educational data mining. Such a stance is particularly interesting, as it presumes that methodologies are indeed 'pure instruments' which transfer from one field to another without bringing along a heavy baggage of epistemic and ontological assumptions. One of these methods is cluster analysis, which represents an excellent case study to illustrate the entanglement of social construction and objective technical expertise.

Although cluster analysis originated in anthropology and, most notably, psychology (Cattell, 1945), it was never a particularly popular approach among social researchers, and it developed mostly outside of social science circles. Savage and Burrows (2007) argue that social scientists looked with suspicion at the adoption of cluster analysis in market research since the 1970s, as it was perceived as a reductionist, overly descriptive technique that avoided the 'hard' questions (answerable through more traditional multivariate analyses), in favour of 'visualisations' that made statistical information accessible to a wide audience of corporate marketing departments. Aside from market research, CA saw significant developments in Computer Science (e.g. Bonner 1964), reflecting a growing interest in using computational techniques to make sense of diverse types of scientific (e.g. epidemiologic) and digitised industrial data. The rise of artificial intelligence also reflects the growing diffusion and sophistication of clustering algorithms, which allow computers to modify their behaviour and make 'intelligent' decisions on the basis of actual and predicted patterns in the data.

In essence, CA operates by partitioning a given data set into groups (clusters), in such a way that all the units in a cluster are more similar to each other than the units in other clusters. The degree of similarity between data points is generally computed in terms of Euclidean distance, based on the assumption that measurements are at least on an interval scale. CA is, in principle, an 'unsupervised' technique, which means that clustering should not occur on the basis of predefined classes that reflect expected or desirable relations amongst the data. In actuality, it is very hard to undertake clustering without some notion of the grouping criteria and without establishing a number of parameters, such as the number of clusters and their density. As a consequence, there always remains a degree of uncertainty as to whether the partitions reflect the overall structure of the data, or if the process has produced artificial groupings. Given the nature of clustering algorithms, results will always be obtained irrespective of the number of variables used and sample size. As notably observed by Aldenderfer and Blashfield (1984, p.16): 'Although the strategy of clustering may be structure-seeking, its operation is one that is

1 See https://tekri.athabascau.ca/analytics/

ISBN 978-1-86220-324-2

structure-imposing'. The challenge is compounded by the existence of a range of different algorithms, each with its specific properties and the potential to produce different outcomes. As a social research method, CA is therefore exploratory rather than confirmatory and, more importantly, it reflects specific inductive principles which are nothing more than the 'mathematical formalisations of what researchers believe is the definition of a cluster' (Estivill-Castro, 2002, p. 65) - as such, 'Clustering is in the eye of the beholder' (ibid).

A crucial implication begins to emerge from this discussion: the (relative) ontological indeterminacy of CA and its relationship with the apparatus of socio-technical and human factors that underpin the choice of grouping criteria and attributes. While this choice is contentious even in the case of biological variables (such as gene expression levels or tissue types), as evidenced in the growing application of CA in biomedical 'big data' research; it is significantly more problematic when based on social or educational attributes. We can accept that basic biomedical features are 'essential' in an Aristotelian sense, i.e. that they reflect distinct characteristics and refer to discrete classes of phenomena. However, this view cannot be easily extended to socioeducational phenomena, which can be better described from a Wittgensteinian perspective, less concerned with how the world actually is than how it's represented symbolically, never in terms of distinct categories, but as overlapping, fuzzy and 'polythetic' relationships.

This is to say that, for the most part, learning environments and the relationships therein are not 'naturally occurring' but are the result of a complex interplay of choices and negotiations, many of which are contingent and draw on a broad palette of cultural factors. A range of antecedent and concurrent factors (educational, technological, epistemological and so on) influence the range and types of attributes around which groupings may or may not form. This argument makes perfect sense from an educational design perspective, as it rests on the rather uncontentious assumption that learning is, to a degree at least, shaped by the pedagogic and epistemic conditions put in place and reinforced by instructional designers, teachers and learners. However, the significance of these networked negotiations between agents – these 'agencements' (Callon, 2007) to remain faithful to a sociomaterial terminology - is sometimes lost in the more instrumentalist readings of learning analytics and educational data mining, even in otherwise theoretically informed accounts. The problem arises again from believing in the neutral and 'pure' nature of tools and methods of data analysis – a belief which inevitably leads to reifying the outputs of those analyses as equally neutral, objective and natural phenomena. This confusion is apparent, for instance, in Siemens (2013) when he suggests that the techniques shared by Learning Analytics and Educational Data Mining can be placed on a conceptual continuum (possibly borrowed from biomedical research) from basic to applied research. Forms of learning are thus 'discovered' in the same way as epidemiological subpopulations:

> 'Through statistical analysis, neural networks, and so on, new data-based discoveries are made and insight is gained into learner behavior. This can be viewed as basic research where discovery occurs through models and algorithms. These discoveries then serve to lead into application' (Siemens, 2013: 7)

By treating clusters of users (or any other analytic output) as essential entities, analysts run the risk of crystallising knowledge about those groups. As a result, deeply contextual knowledge about patterns of engagement with digital content in an online course - for example about 'completing' modules by watching videos and performing assessments - turns into a factual, universal account of learning and accomplishment. The outputs of the analyses are no longer considered as contingent, but as totalising formulations of the social order of digital learning.

There are various examples in the LA and EDM literature which illustrate the range of antithetic and circumstantial criteria chosen for creating clusters: frequency of accessing course materials (Hung and Zhang, 2008); choice of synchronous vs. asynchronous communication during online collaborative work (Serce et al., 2011); strategies used by learners during one-on-one online mentoring (Del Valle & Duffy, 2009). However, one case in particular exemplifies the issue being discussed here. A well-received paper by Kizilcek, Piech and Scheineder (2013), uses CA to identify four prototypical trajectories of engagement in three MOOCs offered by Stanford University on the platform Coursera: Completing, Auditing, Disengaging and Sampling. The objective and 'natural' quality of the resulting clusters is then emphasised by virtue of their 'making sense from an educational point of view' (p.172). The clusters are therefore construed as subpopulations of learners we could realistically expect to 'discover' across a range of diverse online learning contexts. The very same research design was used on a different community of learners in the competing, UK-based platform Futurelearn (Ferguson and Clow, 2015). In this replication, the authors found noticeably different patterns in their data.

ISBN 978-1-86220-324-2

Whilst 'Completing' and 'Sampling' clusters were identified in line with the previous study, more nuanced forms of engagement also emerged: Strong Starters, Mid-way Dropouts, Nearly There, Late Completers and Keen Completers. Ferguson and Clow insightfully suggest that these differences can be explained in light of the different social-constructivist pedagogy that underpins the Futurelearn MOOC platform, which incorporates not just content and assessment but also social interactions. The authors therefore cautiously conclude that 'it is not possible to take a clustering approach from one learning context and apply it in another, even when the contexts are apparently similar. Convincing clusters of learners are most likely to emerge when the approach is informed by the texture of the learning context' (Ferguson and Clow, 2015: 58). Over and above the methodological implications, this unfolding academic discussion in the LA community is particularly interesting from a sociological perspective. The two papers in question point to the differences which are beginning to transpire in the LA epistemic network, with the emergence of centres of expertise that reflect different educational philosophies; one (Stanford's) eager to develop a 'data-driven science of learning'[2] that enthusiastically marries educational research and computer science. The other (the OU's), showing a degree of intellectual alignment with the tradition of 'socially sensitive' British educational research, with its emphasis on conversations, dialogue and contexts (Laurillard, 2002; Wegerif, 2007; Crook, 1996).

Conclusion

The examples briefly discussed in the previous section illustrate the interweaving of interests, choices and technical aspects that become visible when an 'objective' method like Cluster Analysis is examined from a sociomaterial angle. The paper's aim thus far has been to qualify the contention that this method (like similar ones) acts as a 'performative device'. In this situation, 'device' should not be understood literally as a reified artefact that produces certain outputs, but as a networked configuration of expert knowledge, mathematical formalisations, educational philosophies and political-economic interests that operates in a coherent way to produce the same social realities it claims to objectively 'discover'. Following this, what are the implications for social justice in educational technology?

In the first place, there is the suggestion that those concerned with social justice in educational technology need not limit themselves to denouncing structural inequalities and ideological conflicts. At the opposite end of the 'critical spectrum' there is the opportunity to analyse in a more descriptive fashion how hegemonic discourses in education are given authority through techniques and devices. A small example was provided here, meant to describe how the technical and the social become entangled to the point of being inseparable. The choice of cluster analysis was not coincidental either. This method is clearly assuming a certain symbolic quality due to its association with the growing importance of Big Data and the rise of artificial intelligence - two areas whose profound social, economic and cultural significance does not bear repeating.

Above all, the analysis proposed here is meant to challenge the consensus that no alternatives are possible to how things currently stand, because epistemic networks, not matter how big and influential, are never monolithic entities where 'pure' instruments are in the hands of 'pure' agents; but are always open to negotiations e re-interpretations. Like there is no such thing as a 'pure' and inevitable market, there is no such thing as a pure and inevitable 'globalised education'. The awareness of the ontological determinacy at the heart of socio-technical realities opens up spaces for agency and exploration, which are a fundamental prerequisite in any project of social critique that seeks to go beyond simple social commentary and intellectual posturing.

References

Aldenferder and Blashfield (1984). Cluster Analysis. Sage Publications, Beverly Hills
Bonner (1964), R. On some clustering techniques. IBM Journal of Research and Development.
Bowker, G., & Star, S. L. (1999). Sorting things out. Classification and its consequences. Cambridge MA: MIT Press.

2 http://lytics.stanford.edu/report/

ISBN 978-1-86220-324-2

Callon, M., (2007),'What does it mean to say that Economics is Performative?', in Do Economists make Markets? On the Performativity of Economics, edited by MacKenzie, D., Muniesa, F. and Siu, L. Princeton, NJ: Princeton University Press.
Callon, M., Millo, Y., & Muniesa, F. (2007). Market devices. London: Wiley.
Campbell, John P., Peter B. De Blois, and Diana G. Oblinger. 2007. ''Academic Analytics: A New Tool for a New Era.'' Educause Review 42 (4): 40_57.
Cattell, R.B. (1945). The principal trait clusters for describing personality; Personality bulleting, 42: 129-161.
Clow, D. (2013). An overview of learning analytics. Teaching in Higher Education, 18(6), 683-695.
Crook, C. (1996). Computers and the collaborative experience of learning. Psychology Press. London: Routledge.
Del Valle, R., & Duffy, T. M. (2009). Online learning: Learner characteristics and their approaches to managing learning. Instructional Science, 37(2), 129-149.
Estivill-Castro, V. (2002). Why so many clustering algorithms: a position paper. ACM SIGKDD explorations newsletter, 4(1), 65-75.
Ferguson, R., & Clow, D. (2015, March). Examining engagement: analysing learner subpopulations in massive open online courses (MOOCs). In Proceedings of the Fifth International Conference on Learning Analytics And Knowledge (pp. 51-58). ACM.
Fuchs, C. (2010). Labor in Informational Capitalism and on the Internet. The Information Society, 26(3), 179-196.
Halkidi, M., Batistakis, Y., & Vazirgiannis, M. (2001). On clustering validation techniques. Journal of Intelligent Information Systems, 17(2), 107-145.
Hung, J. L., & Zhang, K. (2008). Revealing online learning behaviors and activity patterns and making predictions with data mining techniques in online teaching. MERLOT Journal of Online Learning and Teaching.
Kizilcec, R. F., Piech, C., & Schneider, E. (2013, April). Deconstructing disengagement: analyzing learner subpopulations in massive open online courses. In Proceedings of the third international conference on learning analytics and knowledge (pp. 170-179). ACM.
Latour, B., & Woolgar, S. (1986). Laboratory life: The construction of scientific facts. Princeton University Press.
Latour, B. (2004). "Why Has Critique Run Out of Steam? From Matters of Fact to Matters of Concern." Critical Inquiry 30, no. 2. (Winter 2004).
Laurillard, D., 2002. Rethinking University Teaching: a Framework for the Effective Use of Learning Technologies. London: Routledge.
Muniesa, F., Millo, Y., & Callon, M. (2007). An introduction to market devices. The sociological review, 55(s2), 1-12.
Ozga, J. (2009). Governing education through data in England: From regulation to self-evaluation. Journal of education policy, 24(2), 149-162.
Peters, M. A., & Bulut, E. (Eds.). (2011). Cognitive capitalism, education, and digital labor. New York: Peter Lang.
Ruppert, E., Law, J., & Savage, M. (2013). Reassembling social science methods: The challenge of digital devices. Theory, Culture & Society, 30(4), 22-46.
Savage, M. (2013). The 'social life of methods': A critical introduction. Theory, Culture & Society, 30(4), 3-21.
Savage, M., & Burrows, R. (2007). The coming crisis of empirical sociology. Sociology, 41(5), 885-899.
Selwyn, N. (2010). Looking beyond learning: Notes towards the critical study of educational technology. Journal of Computer Assisted Learning, 26(1), 65-73.
Serçe, F. C., Swigger, K., Alpaslan, F. N., Brazile, R., Dafoulas, G., & Lopez, V. (2011). Online collaboration: Collaborative behavior patterns and factors affecting globally distributed team performance. Computers in human behavior, 27(1), 490-503.
Siemens, G. (2013). Learning analytics: The emergence of a discipline. American Behavioral Scientist, American Behavioral Scientist October 2013 57 (10): 1380-1400.
Wegerif, R. (2007). Dialogic education and technology: Expanding the space of learning (Vol. 7). Springer Science & Business Media.