Revealing and Interpreting Crowd Stories in Online Social Environments

Chris Kiefer, Matthew Yee-King and Mark d'Inverno
Department of Computing,
Goldsmiths, University of London
m.yee-king@gold.ac.uk

Abstract

The underlying patterns in large scale social media datasets can reveal valuable information for interaction designers and researchers, both as part of realtime interactive systems and for post-hoc analysis. *Music Circle* is a social media platform aimed at researching the role of community feedback in online learning environments. A large dataset was collected when the platform was used as part of a *Massive Open Online Course* (MOOC). We developed a novel analysis technique for observing global patterns in the behaviour of students. The technique employs network theory techniques to view student activity as an interconnected complex system, and observes the temporal dynamics of network metrics to create timelines which are clustered into groups using unsupervised learning methods. This approach highlighted global trends and groups of outliers that needed further attention or intervention.

1 Introduction

Online social activity has become a fundamental part of many interactive systems, either explicitly as part of their intended design, or implicitly as part of external and pervasive social media networks. With social activity comes large or massive scale data, which describes interactions between individuals mediated through a variety of possible formats. These datasets can reveal stories about individuals and groups that may be of high significance to stakeholders; for interaction designers, this data can show aspects of behaviour that reveal design problems, suggest design solutions and highlight directions for future iterations. For researchers, this data can give us a broad understanding of trends in behaviour within the context of specific technological environments. Large datasets also present significant challenges in analysis, with the scale of raw data often making direct human interpretation an intractable task. However, by bringing computational analysis into the loop, we can attempt to sculpt the raw data into new forms that, while not necessarily giving absolute answers, present data in a suitable format for further (human) interpretation and criticism.

Music Circle is an online social platform, aimed at exploring ways of understanding and enhancing learning through community feedback. For six weeks in the summer of 2014 it was used to support a Coursera MOOC, *'Creative Programming for Digital Media and Mobile Apps'*[1]. A substantial and detailed log of student interactions was collected. While there were specific questions that could be asked of the data, such a large and complex set of interactions could most likely hold some interesting and unexpected results, and it seemed pertinent to follow a bottom-up approach to data analysis, by letting patterns emerge rather than imposing them. To this purpose, a set of techniques was developed that attempted to elucidate the broad patterns and temporal dynamics of crowd behaviour that occurred during the period of the MOOC, to transform the raw data into a format that would give the research and design teams a deeper understanding of student interactions within the Music Circle environment.

A novel approach was developed, which leveraged network analysis and machine learning techniques to cluster temporal data. We outline the development of this technique and present and critique the results. The following sections address the research questions that were encountered during this development process: how can social media data be encoded into a human readable form that describes temporal patterns in actor behaviour? How can network analysis techniques enhance this encoding? What are good ways to present this data for interpretation by stakeholders?

We present this research as a technique for eliciting information from large datasets for analysis by stakeholders and domain experts, rather than as a process which will supply absolute answers concerning student behaviour. In this light, we do not attempt to provide a quantitative evaluation of the effectiveness of the findings, but try to show, through cross-checking of results in a post-hoc analysis of the Music Circle MOOC data, the potential strengths of our method for use in future projects.

2 Related Work

Our approach is rooted in a network theory perspective. Jiawei et. al [Han *et al.*, 2012] review data mining in this context, proposing that we can extract much more valuable

[1] `https://www.coursera.org/course/digitalmedia`

information from a database by viewing it as a heterogeneous information network rather than a homogenous data repository. We draw on techniques highlighted by Holme and Saramäki [Holme and Saramäki, 2012]; they review the emerging field of temporal networks, looking at techniques for analysing how network topology changes over time, and how temporal information flows. There have been varied approaches to social network analysis, for example, Gottron and Pickhardt[Gottron *et al.*, 2013] explore techniques for temporal analysis of social data, Gilbert et. al. use statistical methods to analyse Pinterest [Gilbert *et al.*, 2013], and Diya et. al. [Yang *et al.*, 2013] look for causes of student dropouts in MOOCs using a network theory approach. Rowe et. al. [Rowe *et al.*, 2013] outline their technique for modelling and analysing the behaviour of users in online communities. They focus on defining individual role categories, and look at how the global composition of these roles changes over time. Chao et. al [Chau *et al.*, 2011] look at the intersection between HCI and data mining, investigating interactive machine learning approaches, and sensemaking.

3 Music Circle

Our system (pictured in figure 1) allows students to share and discuss creative work, and acts as a research platform for studying the role of social media in learning [Brenton *et al.*, 2014]. The key feature is the *Social Timeline*, an online environment for annotating and discussing time-based media. The Social Timeline allows students to highlight and comment on sections of time-based media, and to further discuss these comments. The website has been used in a variety of scenarios to explore creative feedback [d'Inverno and Still, 2014] between students, including jazz piano tuition and as a rehearsal support tool for musical ensembles.

Over the course of a MOOC in the summer of 2014, the website was employed for students to discuss, share feedback on and peer assess videos of their coursework pieces. During the six week course, the students were required to submit a piece of coursework every two weeks. Each coursework brief asked them to program a software application, and to submit a video demo of their application to the Music Circle website for peer assessment. Students were also asked to peer assess three other peoples' work for each submission. As part of the peer assessment process, they could discuss other students' work through the website.

To give an overview of the course statistics, during a six week period, 3716 users registered with the website. Of these, 3558 viewed one or more videos, 827 made one or more comments, and 258 made one or more replies to comments. 2898 videos were submitted for three separate assessments, and were viewed a total of 112,189 times. 7370 comments were made, along with 978 replies. Detailed log data was collected, including timestamped records of all discussions and all media viewing activity.

4 Encoding Stories

Having collected the data, we needed to present it in a format that was both interpretable by humans, and cluster-able by a computer. We built networks of data relating to singular concepts, and observed several network metrics as they evolved over time. In this way, we could use network measurements that put an individual's actions into a global interconnected context, rather than observing them in local scope.

Two sets of networks were built, that separately represented commenting and viewing activity. Each set consisted of networks that evolved in two hour windows over the period of the MOOC, giving 503 networks in each set. This two hour period was chosen as a compromise between time resolution and practical limitations in data processing capacity. The networks had directed and weighted connections; each node represented a student, with weighted links representing the numbers of comments or views made from one student to another.

In this analysis, a set of four metrics were chosen for observation from each network. The first two were simple metrics which sum the (a) incoming and (b) outgoing weights of each node. These could also be calculated without using a network. The next was (c) betweenness centrality. This metric was chosen as it provides interesting representations of each user's importance within the global context of the network, based on how much information flows through their node. It shows the extent that the actor is positioned on the shortest path between other pairs of nodes in the network [Leydesdorff, 2007]. Betweenness centrality is calculated based on link direction and weight, thereby using the full information available in the networks we constructed. The last metric was a calculation of each node's (d) *HITS authority*. This was calculated with the HITS algorithm [Kleinberg *et al.*, 1999], which gives a measurement of the importance of a node based on link structure. More specifically the algorithm gives each node two co-dependent scores; a *hub* score based on the authority of nodes that link to it, and an *authority* score based on how the degree to which the nodes that point to it are hubs.

These four metrics were observed for each two hour iteration of the networks, giving each student a set of timelines, one for each metric. The analysis provided a rich data set for further exploration. Network analysis was carried out using *iPython* with the *NetworkX* library.

5 Discovering Themes

Having collected the sets of timelines for each user, the next step was to cluster these timelines into similar groups to reveal underlying patterns. An exploratory approach was taken, searching for interesting features in the data set by creating both clusters of single features and clusters of compound features in order to reveal correlations between groups of metrics. Useful clustering results were obtained using two methods: k-means alone, and k-means with unsupervised pretraining using Restricted Boltzmann Machines. In the latter case, RBMs were used to find sparse, low dimensional representations of the salient features in the data, before k-means clustered these features. We used the Extended RBM from the Oger toolbox [Verstraeten *et al.*, 2012], with gaussian visible units for our continuous valued data. The RBMs were trained with guidance from [Hinton, 2010].

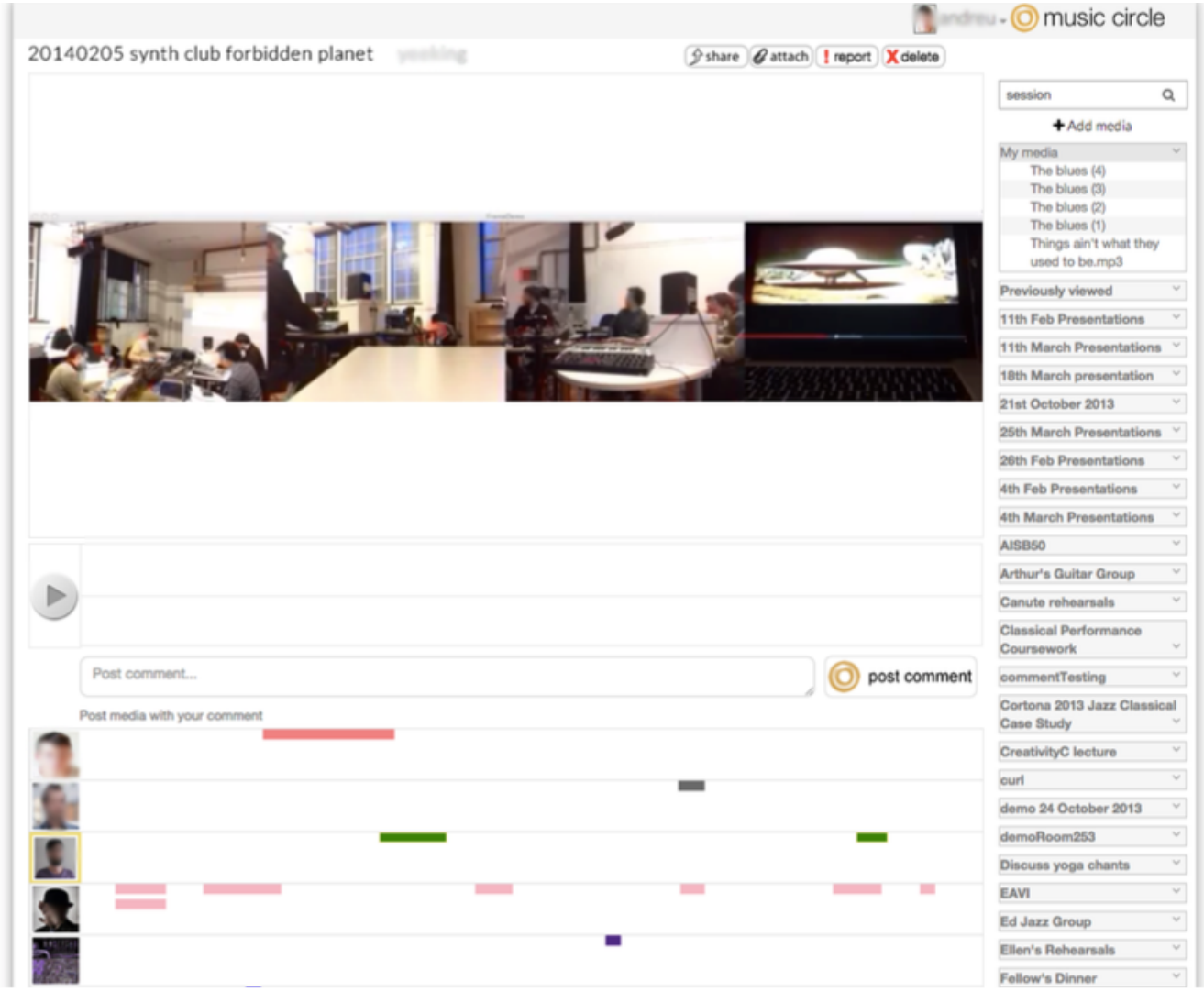

Figure 1: A Screen Shot of the Music Circle Website

6 Visualising Crowd Behaviour

Having calculated clusters, we visualised them in two ways. Simple graphs of means and variances for each cluster group (e.g. figure 2) gave an easily interpretable summary. A more complex view showing members of individual clusters was presented in sets of polar graphs, where each graph displayed every individual timeline in a cluster, superimposed with semi-transparent colouring to show patterns of density (e.g. figure 3). Upon identifying a cluster of interest, a set of graphs was generated to show the cluster mean for each metric, compared to the global means (e.g. figure 4).

7 Results

An exploratory approach was taken to analysing the data, visualising features separately and in combination to find clusters of possible interest. The following examples describe salient outcomes of this process.

7.1 Example A: Betweenness Centrality

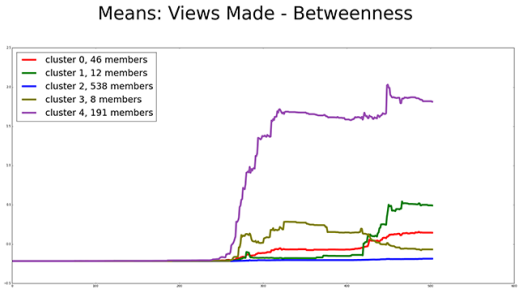

Figure 2: Mean View Betweenness Timelines, in 5 Clusters

Figure 2 shows the means of 5 clusters of betweenness centrality timelines from the *views* network, generated by training an RBM with 503 visible units and 5 hidden units, the output of which was clustered with k-means. In this context, we could consider betweenness centrality to indicate the extent to which a student is engaged in a community of other students who are active in viewing each others' work. A large group of low activity users is shown in cluster 2, which is what would be expected in a social network dataset. A smaller group of high activity users is highlighted in cluster 4 (shown in more detail in figure 3). This timing of this higher viewing activity correlates with an incentive being offered to students for engaging in forum activity. Cluster 3's value is dropping while other clusters are rising, indicating that this cluster may include students who need attention in some way. Further analysis shows that the number of comments received by this group is below the global average (see figure 4), strengthening the case that this group may need some sort of help or motivation.

7.2 Example B: HITS Authority

Clusters of the HITS authority timelines for the last 40% of the course were clustered using k-means (shown in figure 5). The graph shows that the students in cluster 4 have a steadily

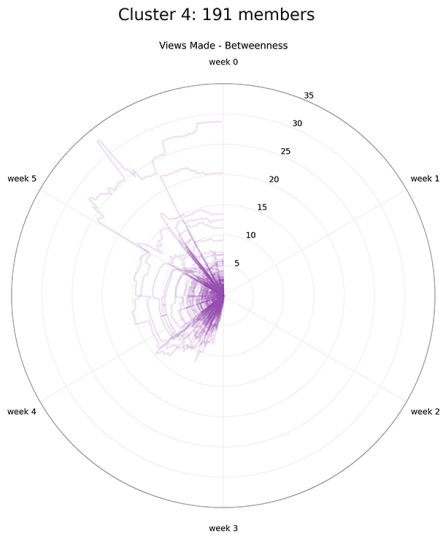

Figure 3: A polar plot of superimposed timelines from a single cluster

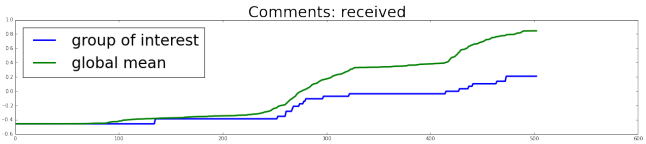

Figure 4: The cluster mean vs global mean for 'comments made'

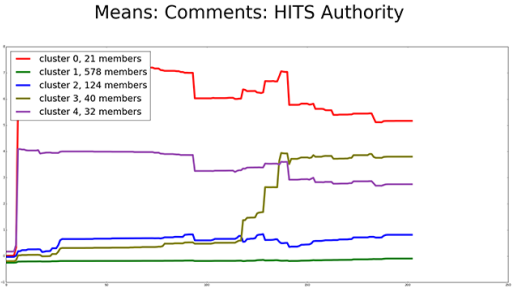

Figure 5: Means of HITS authority for the comments network

declining authority, which may indicate the students in this group have low activity and are therefore becoming less important community members. The concern is verified when looking at their viewing timelines; they are significantly below the global mean for betweenness centrality of views.

7.3 Example C: A Compound Feature

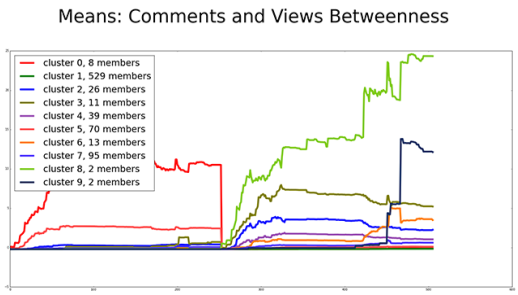

Figure 6:

Figure 6 shows 10 clusters of a compound feature, made with k-means. The first half is a betweenness centrality from the views network, and the second half is the same metric from the comments network. By joining features in this way, the clustering process can pick out potentially interesting correlations or contrasts between the sources. In this example, we can see that the students in cluster 5 have a relatively high value for views but a very low value for comments. This could indicate students who are potentially active, but are part of low activity peer communities, and would perhaps benefit from being introduced to new peers.

8 Discussion

By exploring and clustering the sets of timelines describing behaviour on Music Circle, it was possible to reveal global patterns of crowd behaviour in the specific context of the network metrics we observed. The clusters also highlighted groups of outliers; understanding these small groups can be of great value in trying to understand a complex social system, both in terms of isolating problems, and understanding positive deviance [Ramalingam, 2013]. The technique can be used both post-hoc and for live analysis. Post-hoc analysis can help us to understand crowd behaviour in order to improve the design of future iterations. Live analysis has a number of possible uses in the context of our platform; outlying groups may predict students who are likely to drop out or disengage, and need some contact, reward or support from peers and teachers. Outliers may also highlight successful groups who we may try to link with new peers to strengthen the overall community of learning. A further use is to give students live feedback of their status in terms of these metrics, in order to aid their learning or motivate them. For example, a live timeline of centrality in the comments network, together with a summary based on cluster membership, could provide a good motivator to increase commenting activities.

The first two examples in particular demonstrate the potential strengths of our analysis technique. Both highlight groups of interest, which are validated by patterns in other metrics on other domains. e.g. in example A, the betweenness centrality clusters for viewing behaviour highlight a group that may need attention. Further investigation of the commenting network reveals that this group is less active at making comments, compared to the global mean. Examples A and B also demonstrate the values of analysing our data from a network perspective, viewing user activity as a complex evolving system of interconnected nodes. In example A, the clusters highlight a group whose betweenness centrality is dropping progressively. Observing non-network based metrics for this group, i.e. the number of views made and received, the timelines for this group do not differ greatly from the global average, so this group would not show up in any clusters. However, the highlighting of the group is validated by their inactivity in commenting. In example B, a group is revealed whose HITS authority for comments is dropping. Again, this would be difficult to spot from this group's number of comments made and received, which are close to the global average, but the choice is validated by revealing their low viewing activity. Overall, network analysis algorithms such as betweenness centrality and HITS evaluate each node in the wider context of a complex system. This means these metrics are much more sensitive to global events in the network, and can reveal dynamics that locally scoped measurements may fail to. The results show the merits of temporal analysis of these network metrics; the clusters of interest were highlighted by discovering anomalies in temporal dynamics, and reveal more detailed information compared to instantaneous analysis.

A challenge of using this system is in interpretation. To fully interpret a graph of clusters, it is necessary to understand the network analysis metric being presented, along with its meaning in the context of the network and in the wider context of the source domain of the data. For example, to understand betweenness centrality of commenting activity, we need to understand the concept of this measurement along with the network theory that supports it, and we also need to understand how people are connected by comments on Music Circle and the affordances of the interface that allow activity to propagate though the network of students. It's also a challenge to present the clusters in an optimal way. The means and variance give a good idea of general trends but miss some details. The graphs of superimposed timelines can become dense and difficult to compare, but do give much more detail. Conducting analysis with both of these perspectives seems a good compromise, but ideally an interactive tool would be very useful.

9 Conclusions

The motivation for this project was to reveal patterns of global crowd behaviour based on a large scale database of social and educational activity. Our approach was to look at simple information through the perspective of network analysis. We observed how a variety of network analysis measurements vary over time, and then undertook an exploratory analysis of these timelines though clustering. The clusters highlighted interesting global behaviours of groups of users, and also re-

vealed smaller groups of outliers that may need some sort of intervention or attention. The strength of this technique is demonstrated in examples where the highlighted clusters were shown to need attention though cross checking with other data sources. The possibilities of our technique were demonstrated through post-hoc analysis of forum data. The next step would be to apply this technique on a live forum, and observe the effects of any pedagogical interventions that are made based on the analysis of the resulting data.

10 Future Work

The analysis highlights two areas which could benefit from further development. Firstly, the development of presentation tools to aid human analysis of the clusters. Secondly, the network analysis algorithms employed here have been successful in highlighting clusters but also add an extra layer of interpretation. It would be interesting to investigate the development of domain specific networks measurements, whose output is closely matched to the semantics of the forum behaviour being observed.

Acknowledgments

The work reported in this paper is part of the PRAISE (Practice and Performance Analysis Inspiring Social Education) project which is funded under the EU FP7 Technology Enhanced Learning programme, grant agreement number 318770.

References

[Brenton *et al.*, 2014] Harry Brenton, Matthew Yee-King, Andreu Grimalt-Reynes, Marco Gilles, Maria Krivenski, and Mark d'Inverno. A social timeline for exchanging feedback about musical performances. In *Proceedings of the 28th International BCS Human Computer Interaction Conference on HCI 2014-Sand, Sea and Sky-Holiday HCI*, pages 281–286. BCS, 2014.

[Chau *et al.*, 2011] Duen Horng Chau, Aniket Kittur, Jason I Hong, and Christos Faloutsos. Apolo: making sense of large network data by combining rich user interaction and machine learning. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 167–176. ACM, 2011.

[d'Inverno and Still, 2014] Mark d'Inverno and Arthur Still. Creative feedback: a manifesto for social learning. In *Proceedings of the Workshops held at Educational Data Mining 2014 conference*, 2014.

[Gilbert *et al.*, 2013] Eric Gilbert, Saeideh Bakhshi, Shuo Chang, and Loren Terveen. "i need to try this"?: A statistical overview of pinterest. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '13, pages 2427–2436, New York, NY, USA, 2013. ACM.

[Gottron *et al.*, 2013] Thomas Gottron, Olaf Radcke, and Rene Pickhardt. On the temporal dynamics of influence on the social semantic web. In *Semantic Web and Web Science*, pages 75–87. Springer, 2013.

[Han *et al.*, 2012] Jiawei Han, Yizhou Sun, Xifeng Yan, and Philip S Yu. Mining knowledge from data: An information network analysis approach. In *Data Engineering (ICDE), 2012 IEEE 28th International Conference on*, pages 1214–1217. IEEE, 2012.

[Hinton, 2010] Geoffrey Hinton. A practical guide to training restricted boltzmann machines. Technical Report 1, University of Toronto, August 2010.

[Holme and Saramäki, 2012] Petter Holme and Jari Saramäki. Temporal networks. *Physics reports*, 519(3):97–125, 2012.

[Kleinberg *et al.*, 1999] Jon M Kleinberg, Ravi Kumar, Prabhakar Raghavan, Sridhar Rajagopalan, and Andrew S Tomkins. The web as a graph: measurements, models, and methods. In *Computing and combinatorics*, pages 1–17. Springer, 1999.

[Leydesdorff, 2007] Loet Leydesdorff. Betweenness centrality as an indicator of the interdisciplinarity of scientific journals. *Journal of the American Society for Information Science and Technology*, 58(9):1303–1319, 2007.

[Ramalingam, 2013] Ben Ramalingam. *Aid on the edge of chaos: rethinking international cooperation in a complex world*. Oxford University Press, 2013.

[Rowe *et al.*, 2013] Mattew Rowe, Miriam Fernandez, and Harith Alani. Modelling and analysis of user behaviour in online communities: Ieee computer society special technical community on social networking e-letter. *IEEE Computer Society Special Technical Community on Social Networking E-Letter*, 1(2), May 2013.

[Verstraeten *et al.*, 2012] David Verstraeten, Benjamin Schrauwen, Sander Dieleman, Philemon Brakel, Pieter Buteneers, and Dejan Pecevski. Oger: modular learning architectures for large-scale sequential processing. *The Journal of Machine Learning Research*, 13(1):2995–2998, 2012.

[Yang *et al.*, 2013] Diyi Yang, Tanmay Sinha, David Adamson, and Carolyn Penstein Rosé. Turn on, tune in, drop out: Anticipating student dropouts in massive open online courses. In *Proceedings of the 2013 NIPS Data-Driven Education Workshop*, 2013.