The Anatomy of Information Cascades in the Classroom: An Observational Study

Luis M. Vaquero, Luis Rodero-Merino, Félix Cuadrado

University of Bristol, Stratio, Queen Mary University of London

Abstract

Online learning platforms offer students the option of sharing content. They have become common tools in many universities over the last 10 years. But there is little information about how content spreads in the classroom, *i.e.* how information cascades appear and evolve and what factors are relevant for the formation of cascades. This work analyses information cascades in the classroom, bringing new insights on student learning: students do not share much content, they prefer to share the content they find themselves as opposed to professor-given content, they share more data towards the end of the course, and they do it in bursts. The paper also reveals different behaviour by high-performing students: their interactions are distributed more evenly over the term, their behaviour is more stable and they tend to share documents faster than low-performing students. Documents with high information tend to be less shared. Documents with fewer well-known entities are also shared fewer times. Paradoxically, high-performing students

exchange more documents with high information, compared to mid- and low-performing students.

Keywords: Collaborative learning, Learning Resources, Networking, Online learning environments

Structured Practitioner notes

What is already known about this topic

- There is little evidence on how students actually interact in the classroom using e-learning tools.
- It has been shown that students form online networks and that certain properties of those networks have an impact on grades.
- Static properties of student networks have been studied. However, the dynamics of content distribution in student networks remain elusive.

What this paper adds

- This paper dissects the anatomy of information exchange in student networks: we analyse how students actively share documents.
- The findings of this paper bring substantial clues about the type and timing of information exchanges.
- The paper introduces the concept of the *information density* applied to educational documents and it uses DBpedia as a knowledge repository to categorise the content of these documents.
- We unveil that exchanged information alone is a predictor of student performance and cascade length
- High-performance students seem to be more prone to use documents containing more esoteric contents and use documents that are more focused (in terms of amount of "topics")

Implications for practice and/or policy

- Student network analysis tools based on the patterns shown in this paper may be useful for early detection of behaviour leading to lower grades, although additional studies are needed. For instance, analysing if our findings hold for other disciplines and students of different ages and, in general, increasing the number of courses using network analysis techniques.
- Such analysis has great potential to identify and select the most effective type of content.
- Sharing patterns bring insights about the class behaviour as a whole.

The Anatomy of Information Cascades in the Classroom: An Observational Study

Biographies

Dr Luis M. Vaquero holds a MSc in Electrical Engineering and PhD in Telematics from the University of Valladolid, and a MSc in Pharmacology and a PhD in Medicine from the University Complutense. He spent some time in some U.S.-based research institutions, later on he contributed to building the cloud of the world's third largest telecom (Telefónica). He is now a principal engineer on large scale distributed systems and patent coordinator at Hewlett Packard Enterprise. luis.vaquero@hpe.com

Dr Luis Rodero-Merino graduated in Computer Science at University of Valladolid and received his Ph.D. degree at University Rey Juan Carlos, where he also worked as Assistant Professor. Previously he had worked in the R&D area of Ericsson Spain. He was a researcher at Telefónica R&D before joining INRIA. Later on he moved to the Technical University of Madrid and Gradiant as postdoc and senior researcher. His research interests include distributed systems and software architectures. l.rodero@acm.org

Dr Félix Cuadrado is a Senior Lecturer in the School of Electronic Engineering and Computer Science, QMUL. He obtained his PhD in 2009 from Technical University of Madrid, for which he received a prize for the best thesis. His research interests include large-scale distributed systems, graph processing, cloud computing, and distributed applications. Dr. Cuadrado is currently the programme coordinator of the MSc in Big Data Science at QMUL. felix.cuadrado@qmul.ac.uk

Introduction

Online learning platforms, like Moodle, have become widespread to complement traditional learning scenarios. These platforms can track student activity, but online social networks (OSNs) such as Twitter and Facebook have dramatically changed the ways students interact and exchange information. Thus, tracking student activity requires combining it with data from these OSNs as well, and understanding how information flows across them.

These online platforms are an excellent vantage point to understand information propagation in the classroom. Previous works have observed the shape of information cascades (*i.e.* data flows created by users posting and forwarding data) and correlated these shapes with academic performance (L. M. Vaquero & M. Cebrian, 2013). In this paper, we study the nature of shared content over a course spanning a full semester. We expand previous work by looking at the nature of information being shared, as well as the temporal patterns of content sharing.

This paper is structured as follows. First, we present the most relevant related work. Then, we introduce the main techniques we used for our study as a preamble to the main results. We end the paper with a discussion of the main findings, conclusions and future work.

Related Work

In (D. Yeager, A. Bryk, J. Muhich, H. Hausman, & L. Morales, 2013) the authors proposed a "practical measurements" framework to enable data-based education. Their goal was to create *"the tools to assess changes, predict which students were at-risk for course failure, and set priorities for improvement work"*.

The advent of OSNs has enabled better understanding of how information is transmitted in a variety of disciplines (P. A. Dow, A. Lada, & A. Friggeri, 2013). In this context, information cascades are useful tools to track the way that content spreads.

Social network analysis is an established technique for education research (A. Martinez, Y. Dimitriadis, B. Rubia, E. Gomez, & P. De La Fuente, 2003) (Stahl, Koschmann, & Suthers, 2006). Many studies have focused on static properties of the network. Other studies have combined structural and content analysis (B. Erlin, N. Yusof, & A. Rahman, 2008). But analysis of the *dynamics* of content propagation is still in its infancy. Vaquero *et al.* analysed the cascades created by students and academics (L. M. Vaquero & M. Cebrian, 2013). They found an interesting pattern: high-performing students tend to produce deeper cascades than low-performing ones. Their analysis focused on the topology of different cascade types, without looking at the specific context of sharing. However, in order to understand these interactions it is necessary to explore the dynamics of information sharing as well as the behaviour for different types of content.

Our work follows the trend identified by (Eynon, Hjorth, Yasseri, & Gillani, 2016), which states that OSN analysis techniques need to be complemented with other data mining approaches to "shed light on how people are interacting and learning". This paper presents an expanded analysis on information cascades in the classroom, revealing what content gets propagated, as

well as the dynamics of the propagation. The insight on information cascades is combined with content analysis, by classifying content according to its type and information density. These features provide understanding on student content sharing habits.

Student activity in Massive Open Online Courses (MOOC) has gathered substantial interest from education researchers, taking advantage of the potentially enormous volume of data. MOOC participants engage in very different ways with the learning material compared to traditional classroom environments (Seaton, Bergner, Chuang, Mitros, & Pritchard, 2014). All student interactions take place through the course forums (Milligan, 2015), while many interactions cannot be tracked in classroom learning environments (students participate in different courses and may have known each other for years). MOOCs tend to have significantly higher burstiness at the beginning of the course and close to assessment deadlines (Gillani, Yasseri, Eynon, & Hjorth, 2014).

Materials and Methods

Course Details

The data set consists of about 80,000 interactions of 290 students (see details in (L. M. Vaquero & M. Cebrian, 2013)) during two consecutive years of a 12-week long foundational course on “Information and Technology Skills” for freshmen students of journalism. The course syllabus included topics such as content creation, licensing, and image editing.

Every week students attended a theoretical session and a practical session. Students were assessed with weekly coursework, which often required them to spend additional time beyond the practical session. Sharing activities were not taken into account for grading students.

Students were required to hand out the results of a practical assignment that they could develop independently from their own computers or in the Lab with the assistance of the professors. Since the assignments were based on different software packages (e.g. basic image manipulation or advanced text processing) they were essentially independent (not transferable skills from one into the next). Scores for each individual assignment were published in batches every two-three weeks. This was aimed at avoiding low performance students from getting disappointed early in the course.

Student Interactions: Graph Construction

An *interaction* is a communication between two students sharing some documents or messages. The following types of interactions were recorded: 1) conversations in the course Facebook Chat Canvas and the class IRC (see Table 3 in (L. M. Vaquero & M. Cebrian, 2013) for a detailed list of interactions and types), 2) documents shared in Moodle, 3) files appearing in the HOME folder of each student[1], and 4) files shared as URLs by students in their course-specific Twitter and Facebook accounts. The content of text documents was automatically analysed (e.g. references to the same document in different media were detected from titles or URLs). Video contents were processed with HPE's IDOL[2] to obtain a transcription of the video speech. We decided to process contents automatically to speed up the process and used IDOL since it includes a variety of tools and algorithms to be used at scale, especially for video to text transcription.

---

[1] A HOME interaction is based on the first appearance of a document in the HOME directory of the student.

[2] http://www.havenondemand.com/

This information was modelled as a graph, with students as nodes and directed edges representing an interaction between sender and receiver (L. M. Vaquero & M. Cebrian, 2013). Edges were labelled with the time difference δ between the first reference to the document by the source student t_u, and by the destination student $t_v, u \neq v. \delta = t_u - t_v > 0$, since there are no self-edges. If the content is student generated, then the time of first reshare is used to label the edge. If the content was generated by a professor, then t_u is the time when the teacher first shared the content. While other graphs could be used to enrich this graph with further information, this was not included in our analysis as it goes against the privacy of the students and was not included in their consent.

Document-specific subgraphs represent information cascades, where edges represent the information flow.

Content Categorisation

Documents were classified into two categories that we call informal resources and formal resources. This categorisation was motivated by the feedback provided by students of previous years, where they used these words to refer to these two separate groups. Examples of informal resources are blogs, Q&A sites, or online tutorials. Formal documents used in the course were manuals, peer-reviewed papers and presentation slides. While formal resources have been edited by recognised experts (e.g. academics), informal resources tend to have more varied sources, the reputation of the authors cannot always be measured (e.g. h-index of the authors), and they do not normally undergo a careful editing process.

Information Density

A concern is that students may be using materials from unreliable sources. An analysis of each document can provide insights on its suitability for the course.

Sentiment analysis has been used to analyse dropouts and even attempts to crowdsource text analyses, but these would not give an indication of the informativeness of the contents.

It has been suggested that *information density* can be used to measure the "informativeness" of text documents (C. Horn, A. Zhila, A. Gelbukh, R. Kern, & E. Lex, 2013). While information density may not necessarily the best representation of learning value to a student, we chose it as a proxy.

Horn et al. (C. Horn, A. Zhila, A. Gelbukh, R. Kern, & E. Lex, 2013) indicate that the output of automated factual density techniques contains a large portion of uninformative extractions that can lead to overestimating information density in a document. Factual density has become the most widely used mechanism for estimating information density in text documents of all sorts. Unlike the other metrics used to evaluate content based on objectivity classification and stylometric features (E. Lex, A. Juffinger, and M. Granitzer), knowledge maturing (N. Weber, K. Schoefegger, T. Ley, S. Lindstaedt, A. Brown, and S. Barnes), or simple word count (JE Blumenstock), the concept of informativeness incorporates "complex semantic features" (E. Lex, M. Voelske, M. Errecalde, E. Ferretti, L. Cagnina, C. Horn, and M. Granitzer,).

Systems like OpenIE (C. Horn, A. Zhila, A. Gelbukh, R. Kern, and E. Lex) and ReVerb (E. Lex, A. Juffinger, and M. Granitzer) have made extraction for tuples very straightforward, even for non-experts. We adapted the techniques described in ReVerb[3] and built on part-of

[3] https://github.com/knowitall/reverb

speech (POS) analysis techniques to build triplets better adapted to the specific grammatical details of the language (only Spanish and English were used, which corresponded to 99% of the documents). For instance, a single verb V has to be followed by a preposition P and followed by sequence of words (noun, adjective, adverb, pronoun or determinant) and ends with a P. We also used well-known heuristics such as: prefer longest matches, merge adjacent sequences, and the ones included in ReVerb. IDOL was applied to extract relational *tuples* from text (and transcripted video) documents. Tuples represent *facts* from text without requiring a predefined vocabulary or manually tagged training corpora[4].

Triplet quality was assessed by manual inspection of the triplets created for a 10% random sample of video transcripts and text documents shared by students. This manual curation process was used to refine some heuristics (e.g. changes in regexes) and applied to all the triplets. After curation, all accepted triplets were used.

The *fact* count is determined for each document as the number of triplets found[5]. Information density is calculated as the *fact* count for that document divided by its length in characters. In order to study how *fact* density influences information diffusion, we specifically looked at extreme categories in information density We refer to documents with information densities in the 10^{th}/90^{th} percentiles as low/high information density, respectively. As observed in Table 2, the distribution of density varies greatly across different classes of documents. 10^{th} percentile documents mainly correspond to presentations and blog articles; 90^{th} percentile documents include papers, tech sites and class notes. ~~The documents do not follow a uniform~~

[4] Videos were also automatically transcribed, the analysis revealing they follow the same pattern of information density as regular text files.

[5] *fact* count can be seen as an indirect measure of information of the contents of a document.

~~distribution, which is skewed towards documents with an above average information density (average is ~ 11e-2, median ~ 13e-2).~~

~~These percentiles are representative of more extreme patterns. These measures can be used to explore how content influences information diffusion.~~

Content Analysis of Shared Documents

All documents (including videos or power point presentations) are converted to text.

In order to categorise the content included in the documents shared among students, we utilised an external ontology. Since students used many external sources, we selected the DBpedia[6] ontology, derived from well-known Wikipedia entities. Wikipedia is a basic source of knowledge for any discipline and curated by the community, which eventually includes a representative indication of accepted well-known entities and relationships.

The DBpedia ontology groups entities in different hierarchical categories[7] of the extracted facts. With the graph of entities extracted from the tuples, we calculated the percentage of facts that contain well-known entities and if these entities belong to the same set of classes[8]. 3rd level DBpedia entities can also be seen as a form of categorisation of the contents of the document, for they are broad enough but not too general. For instance, *Work -> Document -> Image* are examples of DBpedia classes at these three levels. Hence, we assumed 3rd level DBpedia entities indicate the "focus" of the contents of the document.

[6] http://wiki.dbpedia.org/
[7] Based on the DBpedia ontology http://wiki.dbpedia.org/services-resources/ontology
[8] http://mappings.dbpedia.org/server/ontology/classes/

Burstiness Analysis

Burstiness can be understood as a measure of how closely in time events occur. (D. L. Jagerman & B. Melamed, 1994) defined the index of dispersion in a computer network as a measurement of its *burstiness*. Given a time series of random variables X_n, where $n = 0, \ldots, \infty$, the index of dispersion I is defined as follows: $I = SCV(1 + 2\sum_{k=1}^{\infty} \rho_x(k))$ where the squared-coefficient of variation (SCV) quantifies the variability in this series and the lag-k autocorrelation coefficient $\rho_x(k)$ expresses the relationship between consecutive occurrences of the random variable with respect to its mean. Please see (D. L. Jagerman & B. Melamed, 1994) for a detailed explanation about *I*. The index of dispersion *I* jointly captures variability and burstiness in a single index, with the summation of autocorrelations measuring the strength of burstiness. The computed time series in this work is X_n, where $n = 0, \ldots, K-1, K = arrivalRate * C$. C is a constant set by trial and error to 2000 in order to smooth cross-batch variations in the autocorrelation coefficient. *C* depends on the dataset and the size of the batch. Tuning it means making the autocorrelation function, which is computed in discrete intervals, look as smooth as possible. For that, we applied a brute force exploration procedure that systematically changed batch size and *C* and chose the two values that rendered a minimum squared sum of time-contiguous inter batch autocorrelations. In order to focus on only short-term changes, we truncate the ∞ limit to *batchsize* / 10.

Large changes in the Index of Dispersion I within successive batches show a significant change in the arrival rate for a given batch; that information coupled with the current and previous arrival rates can show whether a burst has begun or ended. The indices of dispersion on their own are not really informative, and must be considered collectively, in sequence, to understand changes in burstiness.

To account for random fluctuations, we fixed the time interval distribution, shuffled the time series to remove autocorrelation and then calculated the Index of Dispersion I' for the shuffled series. The difference $I - I'$ offers a more accurate account of burstiness. In this data set, however, we calculated the value of I'. It has a negligible value, so we report I as a measurement of burstiness.

Results

Share Analysis

Only a fraction of the total set of documents uploaded in educational portals, user accounts and social platforms was propagated to the students[9]. As Figure 1 shows 80.5% of all the content is posted but never accessed again by any student, and content that is reshared more than twice is rare.

[9] Content in educational portals like Moodle is not actually reshared, but mentioned in comments. We counted this as a reshare in a Twitter/Facebook platform.

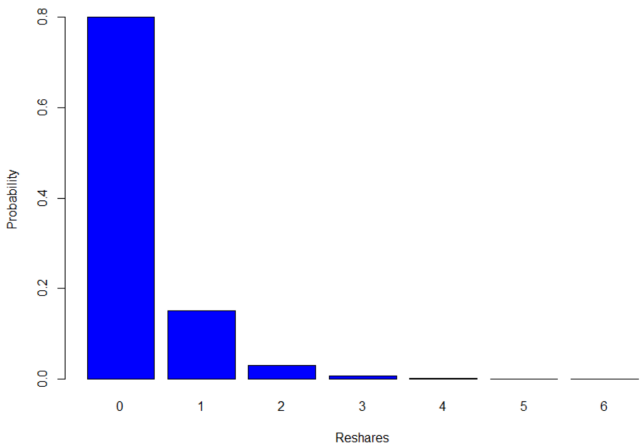

Figure 1 Probability of reshares

Table 1 shows that professors and students generated the same amount of content, but student content was reshared more frequently[10].

Table 1 Content generated and shared. (indicates P~~p~~ < 0.01 compared to student sharing rates using ~~Student's~~ a Mann–Whitney U test ~~t-test~~)*

Source	% Generated	% Reshared
Professor	49.6%	9.3%*
Student	50.4%	21.1%

Formatted: English (United Kingdom)

Table 2 shows an overview of the metrics analysed. First, blog entries, tutorials/examples, and technical sites are the most disseminated in contrast to papers, presentations (power point and video presentations) and manuals. Second, technical sites, tutorials/examples, and manuals are disseminated in long cascades while other sources have a shorter reach. Third, there are no

[10] Professors may leave a doc in Moodle, but students may still refer to it (URL detected in content analysis) or forward to other students.

important variations in the mean time to reshare among the different sources (except in the case of blog entries).

Table 2 Sharing statistics depending on the source/type of document. # Cascade lengths were normalized by the date the content was initially created/shared.

Source	% of docs	Length (pages)[11]	% shared by student	Cascade length#	Information density (x 10^{-2})	Mean time to reshare(h)[12]
Blog	30%	1.4 ± 0.2	83%	1.4 ± 0.8	1.5 ± 0.3	21.6 ± 0.7
Manual	6%	11.3 ± 6.6	12%	1.9 ± 0.2	13.7 ±1.2	32.4 ± 3.5
Tutorials/Examples	10%	2.1 ± 0.7	62%	3.6 ± 1.7	18.1 ± 3.1	29.1 ± 0.5
Presentations	8%	8.9 ± 4.7	15%	0.5 ± 0.5	0.7 ± 0.1	28.9 ± 2.0
Papers	4%	13.5 ± 2.8	7%	1.5 ± 0.9	29.5 ± 4.8	38.1 ± 4.1
Tech. sites	18%	0.8 ± 0.1	95%	5.4 ± 2.4	23.5 ± 2.9	31.2 ± 1.4
Class notes	24%	3.5 ± 1.0	1%	0.8 ± 0.2	14 ± 8.9	37.6 ± 5.0

Table 2 also shows details on the percentage of shared items for each category of documents. When looking together at the percentage of content shared by students, and the average length of information cascades, the categories shared mostly by professors have shorter cascades. Longer cascades contain contents suggested by students (*e.g.* almost 90% of the documents that were shared more than 6 times were posted by students). While the source of a document is a relevant factor, document length did not seem as important (correlation is -0.52 with standard error of 0.38, $p > 0.05$), see top panel in Figure 2. Also, statistical differences could not be found on the weighted average length of informal documents vs. formal documents.

[11] 10pt A4 font size, no images. Calculated using Linux wc command on the text version of the file.

[12] Time lapsed from the first appearance of the document until the last reshare or mention to it.

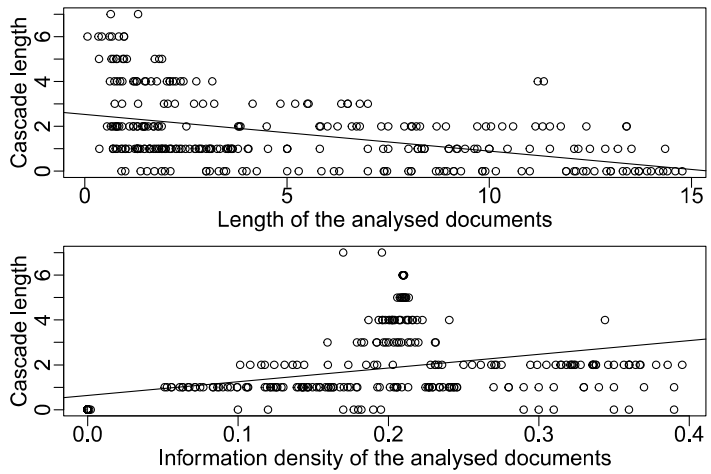

Figure 2 Cascade length analysis of the reshared content. ***(top)*** *Cascade length vs. document length;* ***(bottom)*** *Cascade length vs. information density of the documents.*

In our dataset, we could not identify significant temporal drifts in the sharing patterns of any given content type. Different content types were consistently shared (or not shared) throughout the full duration of the course.

Information of Posted Content

Figure 2 (bottom) shows results on the length of sharing cascades depending on the information density. Measuring information as *fact* count brings similar results, indicating that student behaviour is potentially affected by the amount of information students are exposed to. The most shared documents are clustered in a region with intermediate information values.

Images and Video Content

Students also exchanged images and videos. These elements amounted to less than 10% of the total amount of content, with the majority of them (80%) shared during the first three

weeks of the course. When compared with text files we did not find any statistical differences on how they propagate. Approximately 20% of the tutorials/examples and almost all presentations were videos. The percentage of images shared was negligible (except for those included in documents).

Course Dynamics

Temporal analysis revealed changes in the way documents are exchanged and their information density as the course progresses. Students exchanged mostly low density content over the first 3 weeks. They started sharing higher density documents around week 4 and continued doing that during the rest of the course. This can be observed in Figure 3. x

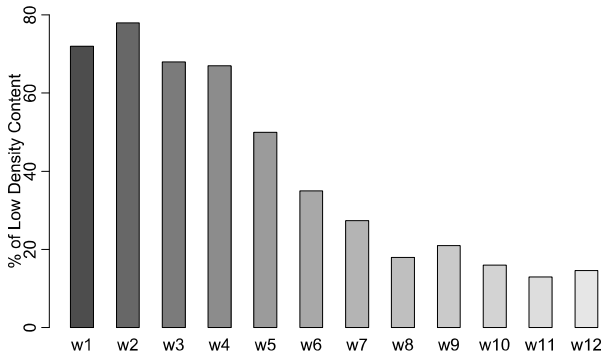

Figure 3 Percentage of low density documents as a fraction of the total reshared documents in that time period.

Looking at students' activity within a week, Figure 4 (top) shows that most sharing actions occur towards the end of each week (deadlines were on Friday). On the other hand, Figure 4 (bottom) shows how long the time is to re-share a document during the first few weeks ($\cong 40h$ as a 90_{th} percentile). As the course progresses, documents are shared earlier.

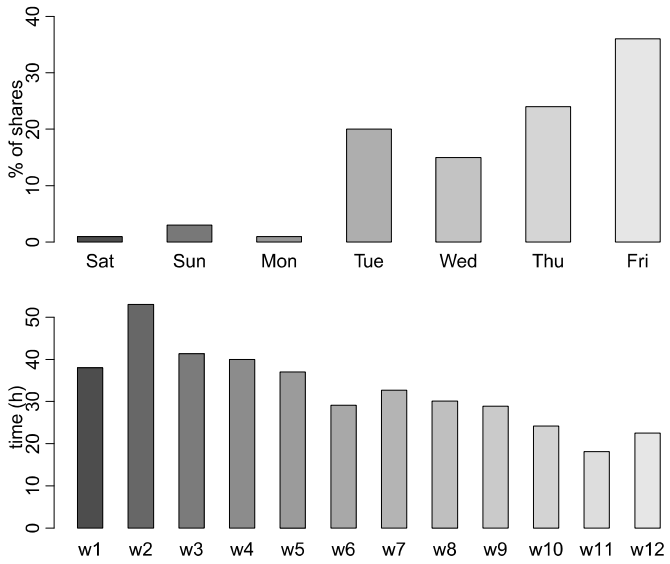

Figure 4 ***(top)*** *Fraction of total reshares for week 8 happening each day.* ***(bottom)*** *90th percentile re-share time based on the week of course.*

Regarding *burstiness*, Figure 5 shows the index of dispersion per 3h sliding window as a function of elapsed time. We show the results from the last assignment, as it required knowledge and techniques from all previous assignments and had a higher weight in the final score. The resulting high index of dispersion, shows that document sharing occurs in short-term bursts.

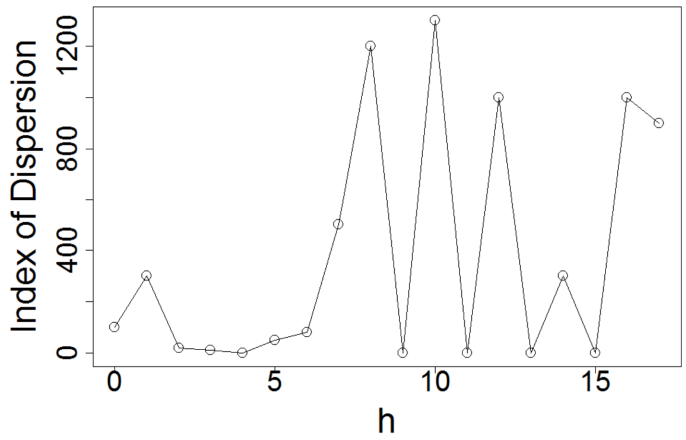

Figure 5 Index of Dispersion for the sharing of events happening on the day of the last deadline of the course. These data correspond to an arrival rate of a few tens of events every two hours.

Finally, Figure 6 shows how the index of dispersion sharply increased with time and reached its maximum value during the second half of the course. As the course proceeds, the assignments grow in complexity and require revisiting skills acquired in previous lectures/assignments. A similar observation is apparent within a single week: the index of dispersion increased towards the end of the week, in this case the explanation may be that ~~as~~ the assignment deadline got closer. The intra-week increase was calculated by normalising the value obtained before the deadline with the value obtained on the day right after the previous deadline. On average, a 310% increase in the intraweek index of dispersion can be observed.

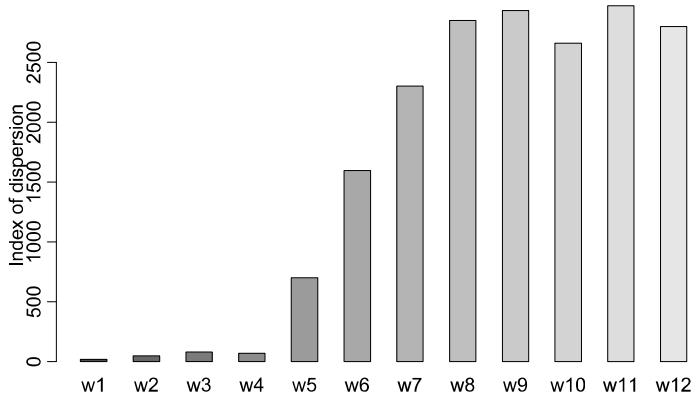

Figure 6 Evolution of the Index of Dispersion with Time.

Cascades and Performance

To further analyse the effects on score, students were grouped into high (> 6.5), mid (between 6.5 and 3.5) and low (< 3.5) scoring groups (scores were given in a 0-10 scale). These boundaries were selected based on a bell curve shape to the scores obtained by students on the same course through the years. The central group (between 3.5 and 6.5) corresponds to more than 60% of the students of the course, while the other two groups represent more extreme scores. If we moved the boundaries between groups10% down (2.5 and 5.5) the end result would be nearly half of the students would be top performing (score above 5.5), which does not correspond to the bell-shaped scoring curve observed through the years.

The same occurs when moving the scoring boundaries up (4.5 and 7.5): nearly half of the students would be low performing.

All the properties reported in our analysis hold when moving the lower score boundary down and the top boundary score up by 5% though, which supports the robustness of this classification. In order to further validate this, we also ran an automated clustering (k-means

testing with $0< k < 6$) and minimising the mean square error of the distance between scores. The value that rendered minimum error was that for k=3.

Vaquero *et al.* reported that most cascades are "trivial" (single source and destination), with high-performing students taking part in deeper cascades. The re-sharing time for high performance students was significantly lower than that of low performance ones, and the index of dispersion for high-performing students (see Table 3) was lower than the one for low-performing ones, despite these students interacting more often. This suggests high-performing students share information at a more regular pace and they may be sharing more information (see analysis below).

Table 3 Normalised dispersion indices for different groups of students. Students were ranked as low, mid and high-performing.

Dispersion Index	High	Mid	Low
Intraweek (Tue to Fri)	0.11	0.45	0.68
Interweek (Weeks 6 to 12)	0.15	0.29	0.63

Content Analysis/Categorisation

Table 4 shows an overview of the contents of the analysed documents. As can be observed, information density is not a determinant factor ($correlation \cong 0.2$) for how many well-known entities existing in DBpedia the document contains. Content shared more frequently by students tends to have more well-known DBpedia entities and be less focused (higher number or distinct 3rd level DBpedia classes).

Table 4 Content analysis of the shared documents

Information source	% well known DBpedia entities	Distinct 3rd level DBpedia classes
Blog	25%	4.4
Manual	0.9%	0.4
Tutorials/Examples	19.9%	3.75
Presentations	7.5%	1.12

Information source	% well known DBpedia entities	Distinct 3rd level DBpedia classes
Papers	1.9%	0.78
Tech. sites	32.8%	5.56
Class notes	0.7%	1.23

Table 5 shows how high-performing students shared more documents with less density of well-known entities than mid- and low-performing class mates. Also, the well-known DBpedia entities contained in the documents high-performing students tend to share fall in fewer categories.

Table 5 Normalised average information shared, % of DBpedia entities found per doc, and 3rd level DBpedia entities depending on student performance. Students were ranked as low, mid and high-performing.

Student performance	Normalised average fact count per doc	Normalised average % of DBpedia entities per document	Normalised average number of 3rd level DBpedia entities
Low	1	6.21	3.38
Mid	3.74	2.96	2.06
High	9.37	1	1

Discussion

In this paper, we expand the work of (L. M. Vaquero & M. Cebrian, 2013) by evaluating the nature of content shared by students on information cascades, as well as the temporal patterns behind information sharing.

While the information density or information of documents related to an assignment may provide an indication on how complex solving the assignment may be, our results indicate that documents containing well-known DBpedia entities are preferred by students and are more often reshared, independently of their information (or information density). This suggests that the ability to access more information from external sources and find more links and explanations

may be making documents more popular. This phenomenon may apply strongly to foundational courses like this one as opposed to more advanced ones.

Mean time to reshare is very similar for all information sources. Although there are statistically significant differences (e.g. between the time it takes to share a blog and a paper), there seems to be no correlation between cascade length or information (or information density) and the time to reshare. Classroom educational cascades seem to be radically different from more generic cascades, such as those in large OSNs and MOOCs (P. A. Dow, A. Lada, & A. Friggeri, 2013) (Anderson, Huttenlocher, Kleinberg, & Leskovec, 2014). Educational cascades tend to be much shorter (never more than 5-6 hops in the cascade). It could be argued this is an effect of a lower average distance between users (4.7 for Facebook, see (J. Ugander, B. Karrer, L. Backstrom, & C. Marlow, 2011), vs. 2.6 in the graph generated from this paper). However, these differences in network structure cannot explain by themselves the substantial differences in the length of the cascades (hundreds vs. 5-6). Smaller number of users may be behind this effect.

Similar analyses on educational interactions have been done in the Massive Open Online Courses (MOOC) arena (Gillani, Yasseri, Eynon, & Hjorth, 2014) (Milligan, 2015) (Seaton, Bergner, Chuang, Mitros, & Pritchard, 2014), but the existence of long lived offline and online interactions (e.g. students who always work together in assignments) and stable groups (study groups) in the classroom suggest the usage of technology and, therefore, course dynamics may be different in MOOC and classroom courses. This may, in part, be explained by the fact that students in University know each other in person and interact in different courses, being a closer community than students in MOOCs, which tend to be geographically dispersed. There is potential in using analysis of information cascades analysis guiding pedagogic decisions.

(Milligan, 2015) *et al*. hypothesised that that theory-informed descriptions of MOOC data access can be used to assess an individual MOOC participant. Monitoring course-related online systems and merging data from different sources may prove useful for early detection of low performance students, enabling corrective actions. Alerting systems based on metrics such as reshare time and burstiness may help in this regard. Also, information cascades may be used as an indicator of student engagement, which is a key factor in student retention (Kuh, Cruce, Shoup, Kinzie, & Gonyea, 2008). Further studies would be needed to validate this point in different courses or different disciplines.

Online learning systems have recently been reported as essential for MOOCs (Eynon, Hjorth, Yasseri, & Gillani, 2016). This type of learning system can enable quick reporting about content quality, so that teachers can get early feedback on content usage and explore interventions to support students in a more agile manner. Enyon and cols typify learners and blending different information sources, including qualitative ones such as questionnaires and surveys; they also reached a point where they could use email as a form of intervention to promote social engagement. In these regards their study is more complete. Our study brings a new set of metrics that can be useful in educational settings. We also show that these metrics have a statistical correlation with student scores. We are not aware they did a detailed factual analysis or used DBpedia as a source of ground truth. They did not study cascades of content reshare.

Students tend to share documents with contents they know something about or where it is easy to find external resources. High-preforming students seem to be more daring with more esoteric resources. They also shared contents faster, in more complex cascades and more regularly than mid- and low-performing students. High-performing students share more content

(forming longer cascades) and the contents in the shared documents tend to be more focused than the contents shared by mid- and low-performing students. We cannot rule out these students had more interactions in person with the professor (e.g. office hours where students can ask questions to the professor outside lecture hours) and these interactions may have been fundamental in guiding them and driving their behaviour.

Qualitative evaluations (e.g. a questionnaire to students) would also be helpful to better-understand which offline or out-of-course factors (e.g. other courses students' are taking, interactions with other students or with the professor) could help explain some of the observed variations.

This work presents an isolated study on a specific population of students with a well-defined background. More similar single case studies are needed in order to generalise our findings to other courses and populations. In this regard, the input from other researchers with their groups and experiences with MOOCs may help to triangulate several case studies confirming or refuting whether or not our findings are broadly generalisable. Indeed, an interventional study would be needed to gather information about group dynamics. However, the knowledge of how a student is involved with documents and an analysis on document content quality (information density, well-known entities, etc.) may be used by a teacher in the classroom.

~~The combination of tools and systems we used are open source (or have an open source equivalent) that can be used by any institution to build weekly reports for teachers to use these metrics.~~

Such studies would not be free of ethical implications that are beyond the scope of this work. We are also considering the use of chatbots to respond to student questions in particularly bursty periods. As future work, similar studies are needed to rule out dependencies on the influence of the specific professor and contents of this course. For instance, it would be interesting to find out whether the same results can be obtained for a completely different subject or degree course.

In addition to offline studies, it would also be interesting to study content sharing while courses are taking place. Online monitoring of student activity could be integrated into existing Online Learning Environments and university student folders. The techniques we present in this paper can be replicated with widely available open source software (except our manual classification of content types). This way, content sharing behaviour could be added to the existing student engagement indicators, potentially enabling individualised actions on these students. Additionally, online monitoring of content sharing would allow teachers to analyse how effective are the learning resources that are offered to the students, potentially discovering alternative information sources that students prefer to engage with.

Statements

On Ethics

All students signed an informed consent agreeing to have their online activities tracked. The student identifiers and contents of the communications were anonymised before data analysis.

On Conflict of Interest

No conflict of interest.

On Open Data

The university regulations on data protection don't allow the release of the collected data, as explicit consent was not provided by students.

References

A. Martinez, Y. Dimitriadis, B. Rubia, E. Gomez, & P. De La Fuente. (2003, December). Combining qualitative evaluation and social network analysis for the study of classroom social interactions. *Computers \& Education, 41*, 353-368.

Anderson, A., Huttenlocher, D., Kleinberg, J., & Leskovec, J. (2014). Engaging with Massive Online Courses. *Proceedings of the 23rd International Conference on World Wide Web* (pp. 687-698). Seoul: ACM.

B. Erlin, N. Yusof, & A. Rahman. (2008, August). Integrating Content Analysis and Social Network Analysis for Analyzing Asynchronous Discussion Forum. *Proceedings of the International Symposium on Information Technology* (pp. 1-8). IEEE.

B. Golub, & M. O. Jackson. (2012, August). How Homophily Affects the Speed of Learning and Best-Response Dynamics. *The Quarterly Journal of Economics, 127*, 1287-1338.

C. Horn, A. Zhila, A. Gelbukh, R. Kern, & E. Lex. (2013, May). Using Factual Density to Measure Informativeness of Web Documents. *Proceedings of the 19th Nordic Conference on Computational Linguistics* (pp. 227-238). Linkoping University Electronic Press.

C. Ullrich, K. Borau, & K. Stepanyan. (2010, October). Who Students Interact With? A Social Network Analysis Perspective on the Use of Twitter in Language Learning. *Proceedings of the 5th European Conference on Technology Enhanced Learning Conference* (pp. 432-437). Springer-Verlag.

D. Bhattacharya, & S. Ram. (2012, August). Sharing News Articles Using 140 Characters: A Diffusion Analysis on Twitter. *Proceedings of the 2012 IEEE/ACM International Conference on Advances in Social Network Analysis and Mining* (pp. 966-971). IEEE.

D. L. Jagerman, & B. Melamed. (1994). Burstiness Descriptors of Traffic Streams: Indices of Dispersion and Peakedness. *Proceedings of the Twenty Eighth Conference on Information Sciences and Systems*, *1*, pp. 24-28.

D. Liben-Nowell, & J.Kleinberg. (2008, March). Tracing Information Flow on a Global Scale Using Internet Chain-letter Data. *Proceedings of the National Academy of Sciences of the United States of America (PNAS), 105*, 4633-4638.

D. W. Johnson, & R. T. Johnson. (2009, June). An Educational Psychology Success Story: Social Interdependence Theory and Cooperative Learning. *Educational Researcher, 38*, 365-379.

D. Yeager, A. Bryk, J. Muhich, H. Hausman, & L. Morales. (2013, December). *Practical Measurement.* Tech. rep., Carnegie Foundation for the Advancement of Teaching, Texas University.

E. Lex, A. Juffinger, and M. Granitzer, "Objectivity classification in online media," in Proceedings of the 21st ACM conference on Hypertext and hypermedia - HT '10, 2010, pp. 293–294.

E. Lex, M. Voelske, M. Errecalde, E. Ferretti, L. Cagnina, C. Horn, and M. Granitzer, "Measuring the Quality of Web Content using Factual Information," in Proceedings of the 2nd Joint WICOW/AIRWeb Workshop on Web Quality, 2012, no. iii, pp. 7–10.

Eynon, R., Hjorth, I., Yasseri, T., & Gillani, N. (2016). Understanding Communication Patterns in MOOCs: Combining Data Mining and qualitative methods. In *Data Mining and Learning Analytics: Applications in Educational Research.* Wiley. Retrieved from http://arxiv.org/abs/1607.07495

Gillani, N., Yasseri, T., Eynon, R., & Hjorth, I. (2014). Structural limitations of learning in a crowd: communication vulnerability and information diffusion in MOOCs. *Scientific Reports*.

C. Horn, A. Zhila, A. Gelbukh, R. Kern, and E. Lex, "Using Factual Density to Measure Informativeness of Web Documents," in In Proceedings of the 19th Nordic Conference on Computational Linguistics (NODALIDA), 2013.

J. E. Blumenstock, "Size Matters : Word Count as a Measure of Quality on Wikipedia," in Proceedings of the 17th international conference on World Wide Web, 2008, pp. 1095–1096.

J. Leskovec, L. A. Adamic, & B. A. Huberman. (2007, May). The Dynamics of Viral Marketing. *ACM Transactions on the Web, 1*.

J. Ugander, B. Karrer, L. Backstrom, & C. Marlow. (2011). The Anatomy of the Facebook Social Graph. *Computing Research Repository*.

Kuh, G. D., Cruce, T. M., Shoup, R., Kinzie, J., & Gonyea, R. M. (2008). Unmasking the effects of student engagement on first-year college grades and persistence. *The Journal of Higher Education, 79*, 540-563.

E. Lex, A. Juffinger, and M. Granitzer, "Objectivity classification in online media," in Proceedings of the 21st ACM conference on Hypertext and hypermedia - HT '10, 2010, pp. 293–294.

L. M. Vaquero, & M. Cebrian. (2013, January). The ``Rich Club'' Phenomenon in the Classroom. *Nature (Scientific Reports), 3*.

Milligan, S. (2015). Crowd-sourced Learning in MOOCs: Learning Analytics Meets Measurement Theory. *Proceedings of the Fifth International Conference on Learning*

Analytics And Knowledge (pp. 151-155). New York, NY, USA: ACM. doi:10.1145/2723576.2723596

N. Gillani. (2013, September). *Learner Communications in Massively Open Online Courses.* Diploma Thesis, Oxford University, Department of Education.

N. Weber, K. Schoefegger, T. Ley, S. Lindstaedt, A. Brown, and S. Barnes, "Knowledge Maturing in the Semantic MediaWiki : A Design Study in Career Guidance," Lect. Notes Comput. Sci., vol. 5794, pp. 700–705, 2009.

P. A. Dow, A. Lada, & A. Friggeri. (2013, July). The anatomy of large facebook cascades. *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media* (pp. 145-154). The AAAI Press.

Seaton, D. T., Bergner, Y., Chuang, I., Mitros, P., & Pritchard, D. E. (2014, #apr#). Who Does What in a Massive Open Online Course? *Commun. ACM, 57*, 58-65. doi:10.1145/2500876

Stahl, G., Koschmann, T., & Suthers, D. (2006). Computer-supported collaborative learning: An Historical Perspective. In R. Sawyer (Ed.), *Cambridge handbook of the learning sciences* (pp. 409-426). Cambridge, U, K: Cambridge University Press. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.63.1418