Analyzing learners' behavior and discourse within large online communities: a Social Learning Analytics Dashboard

Rogério F. da Silva[1], **Itana M. S. Gimenes**[2], **José C. Maldonado**[3]

[1]Campus Avançado em Jandaia do Sul – Universidade Federal do Paraná (UFPR)

[2]Departamento de Informática – Universidade Estadual de Maringá (UEM)

[3]Instituto de Ciências Matemáticas e de Computação – Universidade de São Paulo (USP)

rogerio.ferreira@ufpr.br, imsgimenes@uem.br, jcmaldon@icmc.usp.br

Abstract. *Online Learning Communities (OLC) are nowadays one of the most important producers of Big Data in education. However, the investigation of such environments is underrepresented in educational research. There is a lack of methods and tools that characterize the massive learning associated with the student participation in large OLC. This paper presents a Social Learning Analytics Dashboard (SLAD) to visualize and analyze temporal trend models that outline the evolution of learners behavior over time. Such models suggest that ongoing collaboration and positive emotion have a fundamental role for knowledge sharing in large scale social learning. These findings can be used to take actions in order to regulate social interaction within large OLC.*

1. Introduction

Researchers and policymakers agree that 21st century education must include some important skills, such as critical thinking, collaboration and information literacy [Worsley and Ochoa 2020]. These skills are in line with the recognition of the importance that collaborative practices have in contemporary education [Karoly and Panis 2004]. Increasingly, out-of-school experiences are viewed as critical for arousing and sustaining interest and participation in collaborative problem solving. Thus, it is necessary to understand the educational practices that take place outside institutional settings [Pinkard 2019]. The life-long learning is crucial for acquiring new knowledge and skills in an ever-changing society and it does not necessarily happen within formal education environments [Czerkawski 2016].

In general, formal learning refers to hierarchically structured and chronologically paced educational activities that are facilitated by an instructor [Schreurs and De Laat 2014, Czerkawski 2016]. On the other hand, informal learning refers to unstructured, not teacher-led, and in most cases, spontaneous learning which occurs outside the conventional educational systems, such as instant messaging applications, social networking sites and online communities [Hudgins et al. 2020]. Educational researchers have largely focused on investigating formal learning settings, whilst informal environments have been underrepresented [Hudgins et al. 2020]. Thus, academics describe the lack of methods and tools that assess the learning effectiveness within such environments [Speily et al. 2020]. This paper presents a Social Learning Analytics Dashboard (SLAD) that aims to visually trace learners' behavior in large online communities. It analyzes measures related to interactions and the content of learners' discussion in order to reveal behavioral patterns and discourse styles associated with learning.

DOI: 10.5753/sbie.2021.217468

The main contribution of this paper is addressed to the employment of interactive visual representations and exploratory educational data analysis, in order to amplify cognition and generate insights to the broader understanding of social learning within large Online Learning Communities (OLC). Our conclusions revealed that positive and negative emotion may influence the occurrence of behaviors related to amount of participation, such as reciprocity, simple connectivity and transitivity. The next sections are described as follows: Section 2 introduces the theoretical background and related works; Section 3 describes our methods and tools; Section 4 presents our case study, results and discussion; finally, Section 5 describes our conclusions and future works.

2. Background and Related Work

2.1. Social Learning Analytics

Learning Analytics (LA) refers to the measurement, collection, analysis and reporting of data about learners and their contexts, for purposes of understanding and optimizing learning and the environments in which it occurs [Chatti et al. 2017]. In turn, Social Learning Analytics (SLA) is a distinctive subset of LA strongly grounded in learning theory and focus attention on elements that are relevant in a social and collaborative culture. It aims to demonstrating that new skills and ideas are not solely individual achievements, but are developed through interaction and collaboration [Shum and Ferguson 2012].

SLA utilize data generated by learners' online activity in order to identify behaviors and patterns within the educational environment that signify effective learning. Shum and Ferguson (2012) suggested two analytical methods to perform SLA and investigate sources of data that make sense in a collective context: ***(i)*** **Social Network Analysis (SNA)** is a technique and set of principles based on graph theory for studying relational connections between actors in a network; and ***(ii)*** **Discourse Analysis (DA)** is the collective term for a wide variety of approaches that aims to analyze large amounts of text generated during the online interactions, and potentially provide insights into the quality of students' text and speech posted in online environments.

2.2. Related work and research gaps

This section describes some observable research gaps of available studies, in order to point out key aspects of our approach.

Lack of combination of analytical methods: the connection between SNA and DA is well established in numerous sociological and sociolinguistic studies. However, the combination of these methods in a holistic approach is notably an educational research gap [Joksimović et al. 2015]. Majumdar et al. 2019 developed a Learning Analytics Dashboard (LAD) that analyzes student logs and gather evidence of learning. The authors defined 9 indicators of learner engagement that can be visualized in different graphs. The LAD provides an email widget that enables teachers send feedback to selected cohorts of students. However, the authors did not investigate the content produced by participants.

Small sample sizes: the sample size of majority of the studies that investigates online learning settings is small. Some authors explicitly state this as a limitation [Jan 2019]. We have analyzed two systematic literature reviews in order to support this claim in the LAD domain. Matcha et al. 2019 presented a systematic review of LAD research that reports empirical findings to assess their impact on learning and teaching. Based

on 29 papers, the number of participants involved in each study varied from one to 500 in 27 papers; only two papers involved more than 500 participants. Valle et al. 2021 conducted a systematic review to report trends and opportunities regarding the design of LAD, contexts of implementation, as well as types and features of 28 papers. The mean sample size of participants per publication was 211; there was one paper with 1,406 and 27 papers under 1,000 participants.

The need to investigate dynamics over time: a pressing need for investigating informal learning settings is a move away from static analyses that observe an OLC at one point in time to pursue instead systematic accounts of how such communities change over time [Schreurs and De Laat 2014]. Becheru et al. 2018 proposed a SNA-based platform for visualizing students' collaboration patterns that integrates several social media tools, such as blogs and wikis. The authors implemented a list of visualization needs outlined by teachers, such as exhibit the general status of collaboration and the status of collaboration for each learner. However, they did not provide more details about the behavior trends to explore the temporal dynamics of interactions.

The SLAD described in this paper aims to bridge the gaps above described. By analyzing a large time frame (more than one year), we have combined SNA and DA in order to provide details about temporal trends related to learners' behavior and their discourse style. We have used the dashboard to support the analysis of two large OLC with more than one million registered members. Such communities are nowadays one of the most important producers of Big Data in education. However, there is a lack of research related to their investigation [Nistor et al. 2015].

3. Methods and tools

Generally, informal learning environments have no compulsory assessment procedures. For that reason, OLC commonly provide a peer assessment process performed by participants when interacting with each other. It is based on a reward system and displays for all community members a point scheme that recompenses the frequency and quality of individual participation [Hudgins et al. 2020]. Our SLAD has been applied to data obtained from OLC within online news sharing site Reddit[1]. Nowadays, Reddit comprises approximately 52 million daily active users, 303.4 million posts and 2 billion comments per year[2]. Participants, known as *redditors*, can evaluate (positively or negatively) the discussion topics, creating their score. In general, the positive votes associated with a particular discussion indicate the community's opinion about it. Thus, the topics with the best answers will likely be rated with higher scores [Hudgins et al. 2020]. *Redditors* are also able to assign points to each other responses. These points, named *karma*, indicate the members' expertise and reflect their popularity [Silva et al. 2020]. Discussion score and *karma* points comprise the Reddit peer assessment data.

Our method has fitted a set of machine learning models that identified a group of relevant measures correlated to the peer assessment data. Such models help to recognize the learners behavioral patterns and discourse styles. Fig. 1 depicts our method. It has three stages, described in next subsections.

[1]http://www.reddit.com
[2]http://redditblog.com/2020/12/08/reddits-2020-year-in-review/

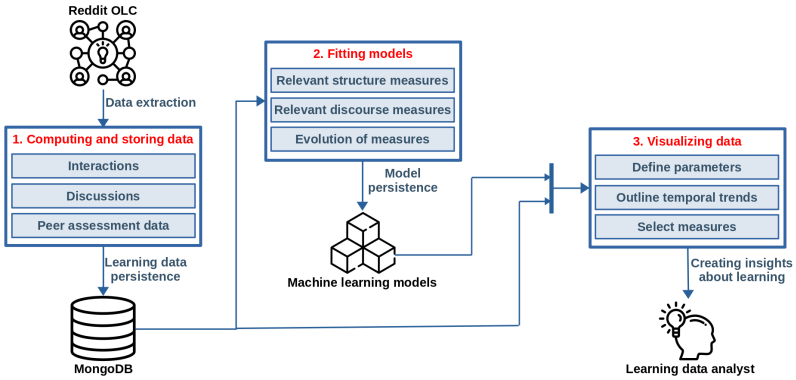

Figure 1. Representation of our method stages

3.1. Stage 1: Extraction, computing and storing learning data

We have extracted data from OLC by using two Python packages for Reddit official Application Programming Interface (API): PSAW[3] and PRAW[4]. Stage 1 has extracted data related to community participants, their interactions, discussions and peer assessment information. These data have been used to compute the measures that assess user behavior and discourse, described as follow:

- Structured measures: they refer to SNA measures that helps to understand how participants are connected and how they interact with each other. In order to analyze the hierarchical structure of messages posted by participants, we have computed in this stage additional measures, such as number of participants in the discussion, discussion size and time of first reply.
- Discourse measures: we have used the well-known linguistic framework Linguistic Inquiry Word Count (LIWC) [Pennebaker et al. 2015] to extract the main word categories of discussions. LIWC extracts 93 measures divided into the categories as summary of language variables, linguistic dimensions, grammar, social processes, affective processes, and others.
- Peer assessment data: they refer to discussion score and participant *karma* points. They are used to fit the models described in next subsection.

These learning data and measures have been stored in non-relational database MongoDB[5].

3.2. Stage 2: Fitting models

Stage 2 has fitted three models created by Silva et al. (2020) that identified the significant structured and discourse measures associated with the best rated discussions. These models are described as follow:

- Relevant structured measures: a linear regression model has identified that the most significant structured measures are the ones related to the amount of participation: *(i)* number of participants in the discussion; *(ii)* discussion size; *(iii)* discussion width; *(iv)* number of sub-communities; *(v)* number of bottlenecks; and *(vi)* number of triangles (or triads).

[3]http://github.com/dmarx/psaw
[4]http://praw.readthedocs.io
[5]http://www.mongodb.com

- Relevant discourse measures: the clustering algorithm *Kmeans* have grouped the discussion topics in order to identify the most significant discourse measures, analyzing which of them are more strongly associated with the best rated discussions. The results have pointed out eight LIWC measures: *(i)* positive emotion; *(ii)* affective processes; *(iii)* drive words; *(iv)* perceptual processes; *(v)* assent words; *(vi)* affiliation words; *(vii)* social words; and *(viii)* negative emotion.
- Evolution of measures: it refers to multiple time series models to explore temporal dynamics of structured and discourse measures, in order to reveal the evolution of learners' behavior and discourse in the period under investigation.

3.3. Stage 3: Visualizing data and creating insights

Stage 3 aims to visualize and create insights about learning data. Data analysts can perform this stage to realize exploratory data analysis in order to create insights about social learning. In addition, informing learners of their level of interaction and increasing awareness of the status of collaboration with their peers, may lead to enhanced self-regulation of social interaction and knowledge sharing in online communities [Joksimović et al. 2015]. Figure 2 shows the aspect of our SLAD. The main characteristics are described as follow:

- Define parameters (see Fig. 2-**A**) - it refers to parameters that configure the data visualization: ***(i)*** Select OLC - it allows to choose one or more OLC data, with the purpose of comparing their similarities and differences; ***(ii)*** Select trend scale - it applies a method to standardize the time series, in order to present the measures at the same scale; and ***(iii)*** Select analysis type - it allows to exhibit data of structured or discourse measures.
- Outline temporal trends (see Fig. 2-**B**) - it shows the behavior over time of the most relevant structured and discourse measures according to the models fitted in Stage 2. The data viewing period can be shortened in order to investigate specific time intervals.
- Select measures (see Fig. 2-**C**) - it allows to disable some measures, with the purpose of emphasizing the most important ones.

Defining parameters, outlining temporal trends and selecting measures can be performed iteratively. Thus, analysts can explore learning data on a large time frame and apply filters to analyze details on demand.

4. Supporting analytics with SLAD: results and discussion

4.1. Data Collection

Our SLAD has been applied to data obtained from two Reddit OLC, named *subreddits*, *learnprogramming*[6] and *MachineLearning*[7], in order to assess their similarities and differences. The *subreddit learnprogramming* was created in September 2009, and had 3,440,477 members enrolled at the evaluation snapshot in June 2021. The *subreddit MachineLearning* was created in July 2009, and had 1,935,702 members enrolled at our evaluation snapshot. In both *subreddits* we have extracted all discussion topics, replies and peer assessment data posted between 2019-Jan-01 and 2020-Dec-31. These *subreddits* were chosen because they are domain oriented OLC, where the members are focused on learning a specific domain (how to learn computer programming and discuss about machine learning, respectively). Table 1 shows the details of data extracted and analyzed.

[6]http://www.reddit.com/r/learnprogramming
[7]http://www.reddit.com/r/MachineLearning

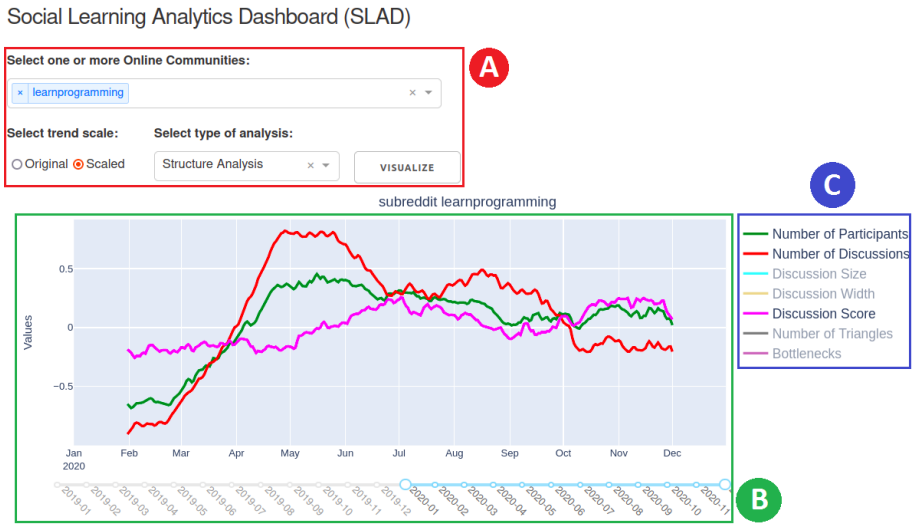

Figure 2. Design of our SLAD

Table 1. Data extracted and analyzed (from 2019-Jan-01 to 2020-Dec-31)

	learningprogramming	***MachineLearning***
Discussion topics	69,447	22,124
Unique active users	95,335	35,702
Replies or interactions	442,243	152,625

4.2. Visualizing learning data

After defining the parameters and selecting measures, we have analyzed the results. Fig. 3 shows the temporal trend models that represent the behavior of structured measures over time in both *subreddits*. The *subreddit learnprogramming* has presented an increasing trend of measures number of participants and number of discussions. This scenario has produced a growth trend of measures related to amount of participation (size, width and score of discussions), although they were less intense. High levels of activity and participation in OLC are the key to the success of such environments. A learner as a member of these communities is both a producer and consumer of information. Thus, they have an important role in creating knowledge artifacts and sharing them to the their peers [Speily et al. 2020]. The *subreddit MachineLearning* has presented a similar growth trend of measures number of participants and number of discussions. However, these increasing trends have not produced a greater amount of participation, because the measures size, width and score of discussions have presented consistent decreasing trends over time. The Figure 4, described as follow, could help to clarify this scenario.

Fig. 4 shows the temporal trend models that represent the behavior of discourse measures. In *subreddit learnprogramming* we have emphasized the measures related to emotions, affective and perceptual processes, because the other measures have not exhibited significant increasing or decreasing trends. The measures positive emotion (words like love, good and nice) and affective processes (words like admire, interesting and laugh) have presented increasing trends, whilst the measures perceptual processes (words like look, hear, feeling) and negative emotion (words like angry, bad and nasty) have exhibited smooth decreasing trends over time. On the other hand, in *subreddit Machine-*

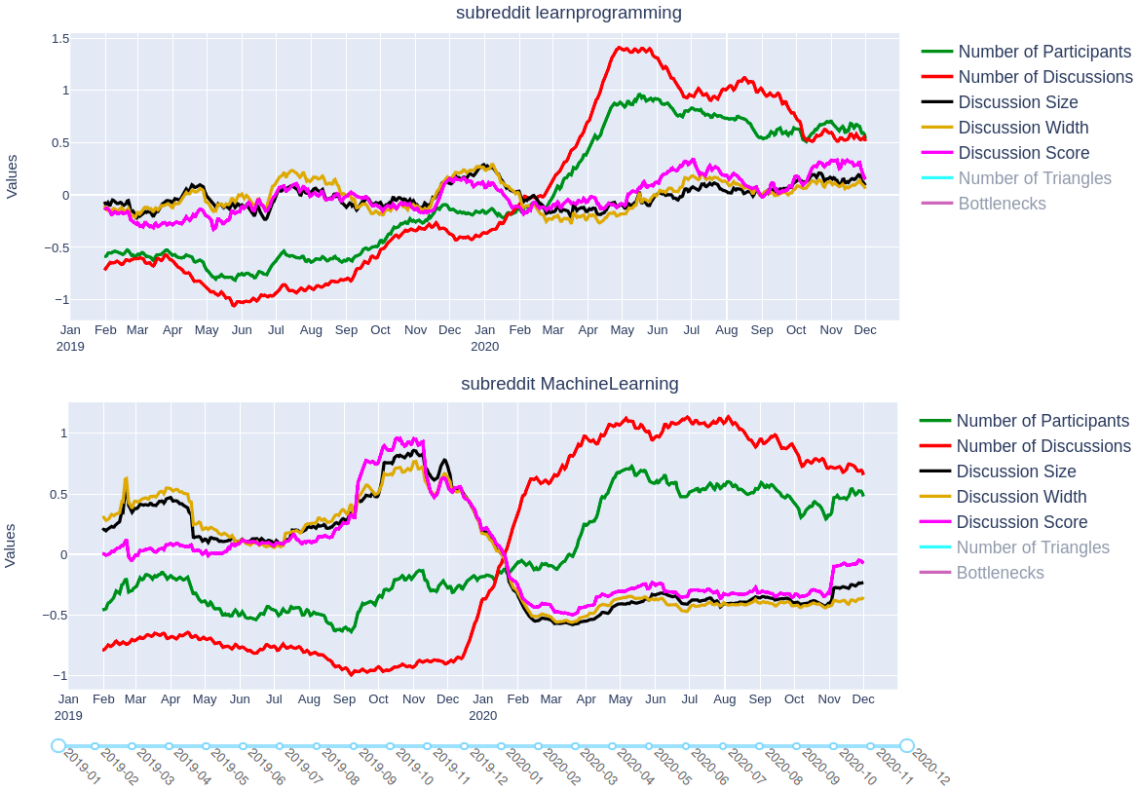

Figure 3. Temporal trend models of structured measures

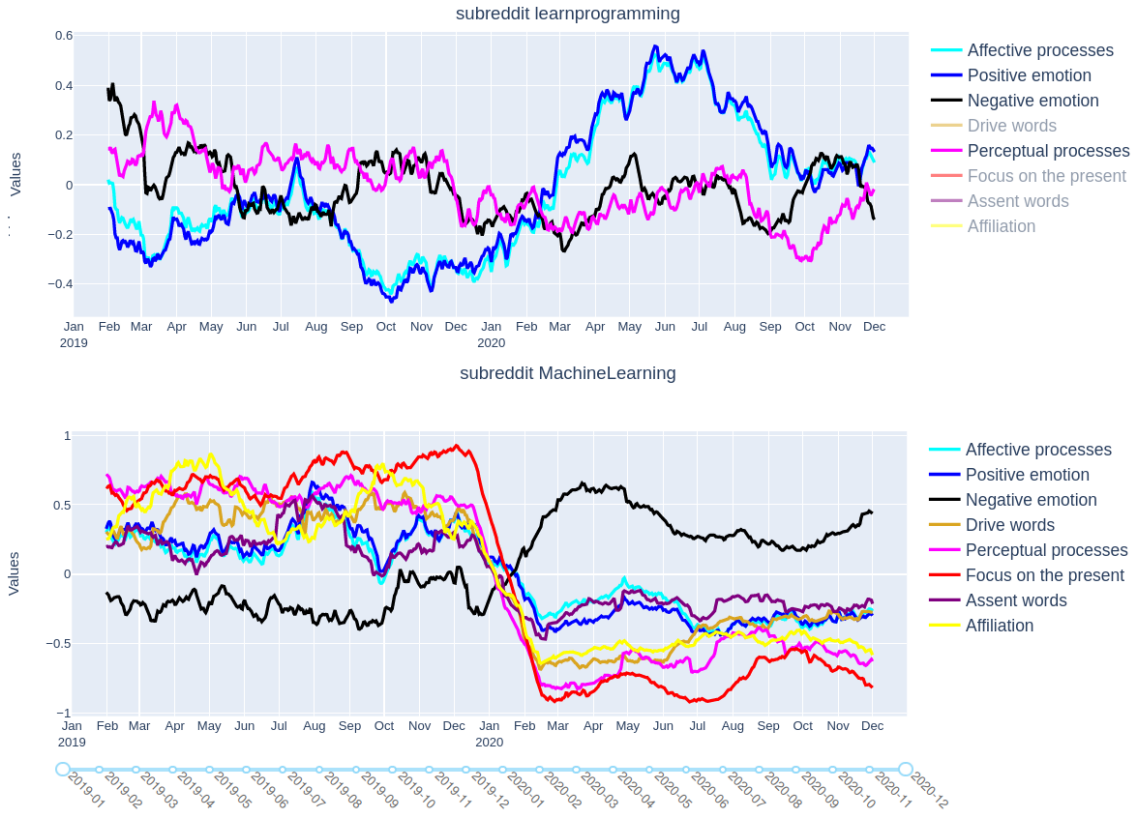

Figure 4. Temporal trend models of discourse measures

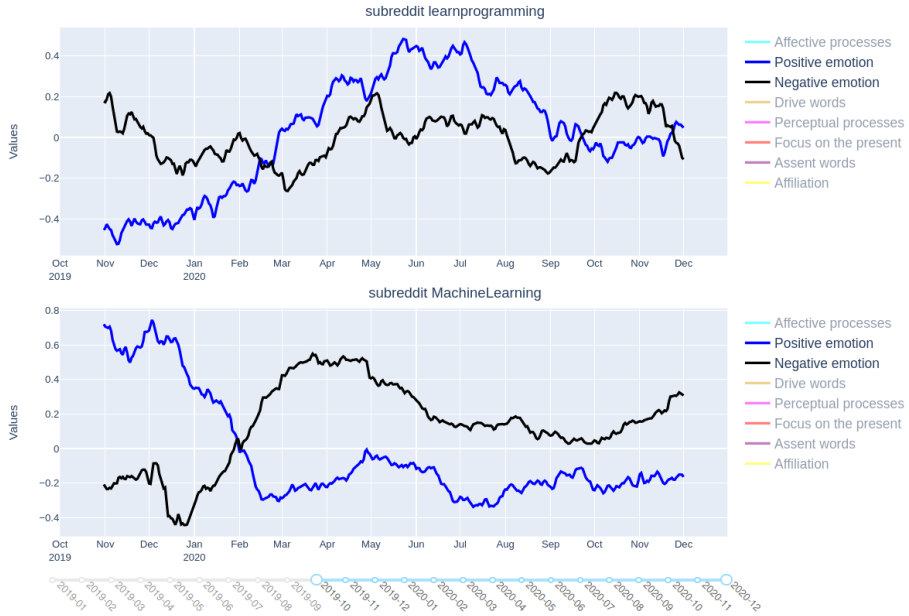

Figure 5. Behavior of positive and negative emotion in a shortened time interval

Learning all discourse measures have presented decreasing trends, except measure negative emotion which has exhibited increasing trend over time. Emotions play a critical role during learning and problem solving with learning technologies. Experimental research has assumed that positive emotions facilitate the use of flexible and creative learning strategies; whilst activating negative emotions leads to more rigid strategies like simple rehearsal and superficial ways of processing information [Pekrun 2006].

In order to investigate the behavior of positive and negative emotion with more details, we have shortened the analysis interval from Oct-2019 to Dec-2020. The result is shown in Fig. 5. The *subreddit learnprogramming* has presented a positive emotion increasing trend most of the time. However, *subreddit MachineLearning* has exhibited higher negative emotion increasing trend near from Feb-2020 to the end of period under analysis. Thus, our SLAD has helped to identify a specific context that could lead to lower levels of activity and participation in *subreddit MachineLearning*. Consequently, we have investigated whether the negative emotion prevalence could influence the occurrence of network effects related to amount of participation, as described in next subsection.

4.3. Creating insights about participation and negative emotion

We have created an Exponential Random Graph (ERG) model [Xiong et al. 2020] to investigate whether negative emotion could influence the occurrence of the following network effects: ***(i)*** **reciprocity** reflects learners' tendency to form reciprocal ties and cluster together, that is, it indicates continuity, collaboration and negotiation of meaning; ***(ii)*** **simple connectivity** reveals the propensity to participate, that is, it means that users who receive messages are more likely to send them and vice versa; and ***(iii)*** **transitivity** indicates the creation of alternative paths that facilitate the information flow in the interaction network.

We have considered only interactions that took place between Jan-2020 and Dec-2020. The result is shown in Table 2. The *subreddit learnprogramming* has presented

Table 2. Results of ERG model for network effects

subreddit learnprogramming			*subreddit MachineLearning*		
Estimates	**Coefficient**	**SE**	**Estimates**	**Coefficient**	**SE**
Baseline (edges)	-12.1109***	0.0092	Baseline (edges)	-11.1476***	0.0087
Reciprocity	8.0950***	0.3342	Reciprocity	6.2750***	0.0805
S. connectivity	0.2756***	0.0077	S. connectivity	0.0367*	0.0001
Transitivity	14.8768***	0.8773	Transitivity	7.3114***	0.2025

Notes: *** means *p-value*<0.001; * means *p-value*<0.05; SE: Standard Error.

higher significant estimates for all network effects. This means that such effects occurred less frequently in *subreddit MachineLearning*. Thus, we argue that the prevalence of negative emotion has produced less network effects associated with amount of participation than expected by chance.

5. Conclusions and future work

This paper presented a SLAD that combines structured and discourse analyses, with the aim of providing to academics, users and community moderators valuable information about learner participation and discourse style in large OLC, an underrepresented learning environment in the educational research. Our SLAD supported the analysis of *subreddits learnprogramming* and *MachineLearning* from the online news sharing site Reddit. The combination of trend models visualization and exploratory educational data analysis was able to point out that the prevalence of negative emotion could explain the decreasing participation in online communities. We confirmed this claim by fitting an ERG model that evaluated network effects associated with the amount of participation. The results showed that the period with negative emotion increasing trend produced such effects less frequently than expected by chance. In future works, we intend to investigate more deeply how the expression of positive or negative sentiment may affect the level of participation in informal learning settings.

References

Becheru, A., Calota, A., and Popescu, E. (2018). Analyzing students' collaboration patterns in a social learning environment using studentviz platform. *Smart Learning Environments*, 5(1):1–18.

Chatti, M. A., Muslim, A., and Schroeder, U. (2017). Toward an open learning analytics ecosystem. In Daniel, B. K., editor, *Big data and learning analytics in higher education*, pages 195–219. Springer International Publishing.

Czerkawski, B. C. (2016). Blending Formal and Informal Learning Networks for Online Learning. *The Int. Review of Research in Open and Distributed Learning*, 17(3).

Hudgins, W., Lynch, M., Schmal, A., Sikka, H., Swenson, M., and Joyner, D. A. (2020). Informal Learning Communities: The Other Massive Open Online 'C'. *L@S 2020 - Proceedings of the 7th ACM Conference on Learning @ Scale*, pages 91–101.

Jan, S. K. (2019). Investigating virtual cops with social network analysis: guidelines from a systematic review of research. *Int. J. of Web Based Communities*, 15(1):25–43.

Joksimović, S., Gašević, D., Kovanović, V., Riecke, B. E., and Hatala, M. (2015). Social presence in online discussions as a process predictor of academic performance. *Journal of Computer Assisted Learning*, 31(6):638–654.

Karoly, L. A. and Panis, C. (2004). *The 21st century at work: Forces shaping the future workforce and workplace in the United States*. RAND Corp., Pittsburgh, 1 edition.

Majumdar, R., Akçapınar, A., Akçapınar, G., Ogata, H., and Flanagan, B. (2019). Laview: Learning analytics dashboard towards evidence-based education. In *Companion Proceedings of the 9th International Conference on Learning Analytics and Knowledge*.

Matcha, W., Gašević, D., Pardo, A., et al. (2019). A systematic review of empirical studies on learning analytics dashboards: A self-regulated learning perspective. *IEEE Transactions on Learning Technologies*, 13(2):226–245.

Nistor, N., Derntl, M., and Klamma, R. (2015). Learning Analytics : Trends and Issues of the Empirical Research of the Years 2011 – 2014. *Design for Teaching and Learning in a Networked World*, 4:453–459.

Pekrun, R. (2006). The control-value theory of achievement emotions: Assumptions, corollaries, and implications for educational research and practice. *Educational psychology review*, 18(4):315–341.

Pennebaker, J. W., Boyd, R. L., Jordan, K., and Blackburn, K. (2015). The development and psychometric properties of liwc2015. Technical report, Univ. of Texas at Austin.

Pinkard, N. (2019). Freedom of movement: Defining, researching, and designing the components of a healthy learning ecosystem. *Human Development*, 62(1-2):40–65.

Schreurs, B. and De Laat, M. (2014). The Network Awareness Tool: A web 2.0 tool to visualize informal networked learning in organizations. *Computers in Human Behavior*, 37(1):385–394.

Shum, S. B. and Ferguson, R. (2012). Social learning analytics. *Journal of educational technology & society*, 15(3):3–26.

Silva, R. F., Gimenes, I. M. S., and Maldonado, J. C. (2020). An Approach for Assessing Large Online Communities in Informal Learning Environments. In *Anais do XXXI Simpósio Brasileiro de Informática na Educação*, pages 642–651.

Speily, O. R. B., Rezvanian, A., Ghasemzadeh, A., Saghiri, A. M., and Vahidipour, S. M. (2020). Lurkers versus posters: Investigation of the participation behaviors in online learning communities. In *Educational Networking*, pages 269–298. Springer.

Valle, N., Antonenko, P., Dawson, K., and Huggins-Manley, A. C. (2021). Staying on target: A systematic literature review on learner-facing learning analytics dashboards. *British Journal of Educational Technology*.

Worsley, M. and Ochoa, X. (2020). Towards collaboration literacy development through multimodal learning analytics. In *Proceedings 10th International Conference on Learning Analytics and Knowledge (LAK20)*, volume 2610, pages 53–63.

Xiong, J., Feng, X., and Tang, Z. (2020). Understanding user-to-user interaction on government microblogs: An exponential random graph model with the homophily and emotional effect. *Information Processing & Management*, 57(4):102229.