An LDA Topic Model and Social Network Analysis of a School Blogging Platform

Xiaoting Kuang
EdLab
Dept. of Human Development
Teachers College
Columbia University
xk2120@columbia.edu

Hui Soo Chae
EdLab
Teachers College
Columbia University
hsc2001@columbia.edu

Brian Hughes
EdLab
Teachers College
Columbia University
bsh2001@columbia.edu

Gary Natriello
EdLab
Teachers College
Columbia University
gjn6@columbia.edu

ABSTRACT
Pressible is a school blogging and content management system developed by EdLab at Teachers College Columbia University. In this paper, social network analysis and natural language processing with Latent Dirichlet Allocation topic model approaches were utilized to gain insights into Pressible, to explore four developmental stages of a college-wide social network and their associations with blog content. The results showed that professors who developed courses became the most influential persons in the network. Students extended the online discussion topics beyond the scope of course topic set by professors.

Keywords
SNA, NLP, Topic Model, LDA

1. INTRODUCTION
EdLab adapted the Wordpress Content management systems (CMS) framework and developed Pressible for the Teachers College (TC) community in 2008. It was designed for fast content delivery, minimization of users' time spent managing technology, and developing connections between users (Zhou, 2013). From the perspective of social constructivist theory, people communicate, contribute and acquire knowledge through social engagement and discussion of topics (Vygotsky, 1978). People also gain knowledge online via connecting information (Siemens, 2004). Massive Open Online Courses (MOOCs) provide more opportunities for people to study for personal intellectual growth (Kizilcec et al., 2017). Social factors from online discussion forums (Rose, et al., 2014) and engaging in higher order thinking behaviors enhanced learning in MOOCs (Wang, et al., 2016). Higher Education utilizes academic blogging to facilitate social networking, self-directed learning, and collaboration. Simulation studies on the blogosphere indicate that improved management facilities on course blogs positively affect the density and connectedness in learning networks (Wild & Sigurðarson, 2011). This study utilized social network analysis (SNA) to investigate human-human interaction and the development of social connections on this blogging platform. Next, Latent Dirichlet Allocation (LDA) topic model method was applied to understand human-information interaction during different developmental stages of Pressible. This study provides an exploratory examination of four developmental stages of an online learning community in a school blogging system.

2. METHODOLOGY
2.1. Participants and Data Collection
The data were collected from the entire Pressible database and contained 3598 users and 594 sites, with 50422 posts in total. The specific aim of this study was to explore the social network and its association with content creation. Only the interactions between registered IDs were counted as valid connections. After the reconstruction of the database for SNA, there were 172 blogs with data on a total of 11146 connections and 429 interactive users.

2.2. Social Network Analysis
SNA is a method to analyze the connections, relationships, and interactions between individuals and communities in the collaborative social network, expressed as the node and edge diagrams (Wild, 2016; Slater et al., 2017). In this study, R package igraph (Csardi & Nepusz, 2006) constructs, modifies and calculates the social networks. Density measures the proportion of contacts observed between pairs of nodes in the network; Eigen centrality measures the importance of a node's network by weighting its top connecting nodes' indegree and outdegree centrality (Daniel, et al., 2010).

2.3. Latent Dirichlet Allocation Topic Modeling
To analyze the content of comments and posts in the blogs, LDA topic modeling was utilized to discover and infer the general topics by scanning the words and their distribution probabilities within documents (Blei, et al., 2003). The R package *tm* was used to construct the corpus for text mining. The *tm* package removes spaces, stop words, numbers, spaces, and punctuation, converting the words to lower case and roots to construct a term-document matrix, which allows analysis of individual words in the corpus (Feinerer & Hornik, 2015; Lang, 2017). The R packages *topicmodels and tidytext* were utilized to calculate the term frequency, construct the inverse document matrix, remove the uncommon terms, find the most common words for individual topics and group the documents by generated topics (Grün & Hornik, 2011; Lang, 2017; Silge & Robinson, 2017).

3. RESULTS AND DISCUSSIONS
3.1. Social Network Development
Descriptive statistics analysis on yearly data was conducted to show the general social network activity in Pressible by developmental stages (Tables 1). The results indicate that this blogging system shifted from a development stage (beginning to 2010 Summer), to a stable growth stage (2010 Fall to 2012 Summer), a rapid growth stage (2012 Fall to 2015 Summer), into a decline stage (2015 Fall until now). The active member numbers increased from the development stage to rapid growth stage and decreased in the decline stage. Their engagement rates as average connection numbers increased from development to the rapid growth stage, which also dropped at the decline stage. Therefore, the number of active members and their engagement rate determine the growth of this online social learning community. The density of the social network among active members decreased while the network was growing from 2011 to 2015 (Fig. 1), indicating that the network became decentralized as more active members joined. Most of the participants were students. They became less active in interactions on Pressible after graduation. New students joined the social network and formed new social centers. Thereby, the global social density decreased because of the dynamic student community (Fig. 1). As more professors built their courses on Pressible, more active students joined this online learning community for discussions and made meaningful connections. Recruiting more professors to take

advantages of Pressible for its online course creation features is a key to maintaining the rapid growth of this social network.

Table 1. Descriptive Statistics by Developmental Stage

Stage	Ave. conn. (/year)	Ave. active IDs (/year)	Ave. conn. per IDs (/year)	Top popular topic of the stage
Development	215.5	36.5	5.9	video game
Stable Growth	1333	89.5	14.9	teach and learn
Rapid Growth	2021	118.7	17.0	think and know
Decline	993	104	9.6	music performance

3.2. Most Influential Members and Topic Interaction Analysis by Developmental Stage

To determine the optimal number of topics of the whole Pressible database, the perplexity values of models were calculated. The LDA topic training model was constructed based on 10000 documents with the range of 2 to 50 topic numbers. The other 1146 documents are used to test the model with the calculation of perplexity and entropy. Based on the perplexity of testing data, 30 is the optimal topic number for this dataset.

During the developmental stage, the library staff was the most active members in the network. Their online discussions focused on the topics: "video game, education", indicating that library staff was using Pressible as a communication tool to share thoughts and discuss education media.

During the stable growth stage, a TC professor (ID: 1490) from the music education program built his courses on Pressible for three years (2011 to 2013), and he continuously received the highest eigen centrality score for three years. During the stable growth stage, the popular topics became focused on education. People who talked about "think and know" were also interested in "video game" at this stage.

During the rapid growth stage, the professor with ID 3132 brought new students into this blogging system though his courses *Creativity & Problem Solving in Music Education*. It was a course extended from the materials developed by the professor with ID 1490, with the same topic "read" and high-frequency words "music, read" for most of the posts. This was the pedagogy course to meet the New York State and national teacher preparation standards.

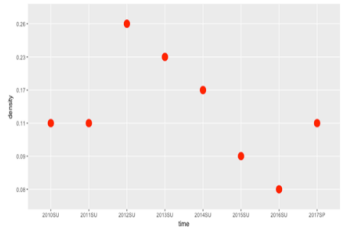

Figure 1. Social Network Density by Year.

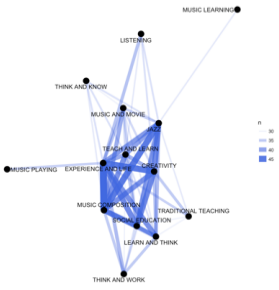

Figure 2. Topic Co-occurrence frequency in the rapid growth stage

Individuals' topic co-occurrence indicated a robust network in the rapid growth stage (Fig. 2). People talked about the topics of "creativity", "music composition", "Jazz", "social education", "learn and think", "experience and life" and "teach and learn" at high co-occurrence frequencies (above 30). In the decline stage, the topic co-occurrence network dropped in topic connection intensity which might be due to less active members in the overall network (Table 1). This finding indicated that more active members encouraged online discussions with more diverse topics. In course blogs, students extended discussion topics to the perspectives that they care about: "music learning, music playing, social education, creativity and experience and life", beyond the scope of the professor's set topic "read".

4. IMPLICATIONS

This study identifies and explores four developmental stages of the social network: development, stable growth, rapid growth, and decline. The SNA and topic model analysis results imply that the influential people will bring new communities into the social network by sharing the content of the hottest topics. Deliberately recruiting more influential people into the social network would accelerate its transition from the stable growth stage to the rapid growth stage.

5. REFERENCES

[1] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. Journal of Machine Learning Research, 3(Jan), 993-1022

[2] Csardi, G., & Nepusz, T. (2006) The igraph software package for complex network research, InterJournal, Complex Systems 1695. 2006.

[3] Daniel, M., Messing, S., Nowak, M., & Westwood, S. J. (2010) Social Network Analysis Labs in R. Stanford University

[4] Feinerer, I., & Hornik, K. (2015). tm: Text Mining Package. R package version 0.6-2.

[5] Grün, B., & Hornik, K. (2011). "topicmodels: An R Package for Fitting Topic Models." Journal of Statistical Software, 40(13), pp. 1-30

[6] Kizilcec, R. F., Pérez-Sanagustín, M., & Maldonado, J. J. (2017). Self-regulated learning strategies predict learner behavior and goal attainment in Massive Open Online Courses. *Computers & Education*, *104*, 18-33.

[7] Lang, C. (2017) HUDK 4051: Learning Analytics: Process and Theory. Columbia University. New York

[8] Rosé, C. P., Carlson, R., Yang, D., Wen, M., Resnick, L., Goldman, P., & Sherer, J. (2014). Social factors that contribute to attrition in MOOCs. In Proceedings of the first ACM conference on Learning (pp. 197-198). ACM

[9] Siemens, G. (2004). Connectivism: A learning theory for the digital age. elearnspace. Retrieved December 12, 2007, CHI '00. ACM, New York, NY, 526-531

[10] Silge, J., and Robinson, D. (2017) "Text Mining with R: A Tidy Approach" O'Reilly Media

[11] Slater, S., Joksimovic, S., Kovanovic, V., Baker, R., & Gasevic, D. (2017) Tools for Educational Data Mining: A Review. Journal of Educational and Behavioral Statistics. 2017, Vol. 42, No. 1 p85-106

[12] Vygotsky, L. (1978). Mind and society: The development of higher psychological processes. Cambridge, MA: Harvard University Press.

[13] Wang, X., Wen, M., & Rosé, C. P. (2016, April). Towards triggering higher-order thinking behaviors in MOOCs. In Proceedings of the Sixth International Conference on Learning Analytics & Knowledge (pp. 398-407). ACM

[14] Wild, F., & Sigurðarson, S. E. (2011). Simulating learning networks in a higher education blogosphere–at scale. In European Conference on Technology Enhanced Learning (pp. 412-423). Springer Berlin Heidelberg

[15] Wild, F. (2016). Learning analytics in R with SNA, LSA, and MPIA. Springer.

[16] Zhou Z. (2013) Connecting Teacher Bloggers: Unleashing the Educational Power of Wordpress