Chen, B., Chen, X., & Xing, W. (2015). "Twitter Archeology" of Learning Analytics and Knowledge Conferences. In Proceedings of the Fifth International Conference on Learning Analytics And Knowledge (pp. 340–349). New York, NY, USA: ACM. doi:10.1145/2723576.2723584

"Twitter Archeology" of Learning Analytics and Knowledge Conferences

Bodong Chen
University of Minnesota
Minneapolis, MN, USA
chenbd@umn.edu

Xin Chen
Purdue University
West Lafayette, IN, US
chen654@purdue.edu

Wanli Xing
University of Missouri
Columbia, MO, US
wxdg5@mail.missouri.edu.edu

ABSTRACT

The goal of the present study was to uncover new insights about the learning analytics community by analyzing Twitter archives from the past four Learning Analytics and Knowledge (LAK) conferences. Through descriptive analysis, interaction network analysis, hashtag analysis, and topic modeling, we found: extended coverage of the community over the years; increasing interactions among its members regardless of peripheral and in-persistent participation; increasingly dense, connected and balanced social networks; and more and more diverse research topics. Detailed inspection of semantic topics uncovered insights complementary to the analysis of LAK publications in previous research.

Categories and Subject Descriptors

K.3.1 [**Content Analysis and Indexing**]: Linguistic processing; H.3.5 [**Online Information Services**]: Pattern analysis; I.2.7 [**Natural Language Processing**]: Text analysis; K.4.3 [**Organizational Impacts**]: Computer-supported collaborative work

General Terms

Algorithms, Human Factors, Measurement

Keywords

Learning Analytics, Twitter, Twitter Analytics, Social Network, Hashtag Analysis, Topic Modeling

1. INTRODUCTION

Learning analytics as a nascent field of scholarship is evolving rapidly and garnering broad interest in both educational research and practice [23]. Since the inaugural Learning Analytics and Knowledge (LAK) conference in 2011, exciting moves have been made during the past four years. The Society of Learning Analytics Research (SoLAR) was launched in 2012; the Learning Analytics Summer Institute (LASI) was first held in 2013; the first issue of the Journal of Learning Analytics was published in 2014. Together these events indicated the establishment of learning analytics as an independent field of research and practice.

Since learning analytics as a field is still in its early stage, efforts have been made to understand the evolution of the field as well as its linkages with others. For instance, researchers have attempted to understand the similarities and distinctions between learning analytics and Educational Data Mining (EDM)—two communities that share similar interests but grew separately in their early years [37]. Colleagues have also attempted to study the roots of learning analytics and its relations with fields such as learning sciences, machine learning, and data-driven analytics [16, 3]. More recently, empirical studies were conducted to understand the field based on the LAK Open Dataset—a dataset which contains structured metadata from research publications in the field of learning analytics and EDM [43, 14]; to this end, topic models [36], ontology [46], visualizations [35], and knowledge systems [30, 20] were built and contributed to the efforts of uncovering key themes of the field and identifying major challenges faced by the community [38].

The present study contributes to the ongoing reflection upon learning analytics by analyzing Twitter archives of the past four LAK conferences from 2011 to 2014. Using tweets posted by the conference participants—who attended LAK either in person or remotely—we are hoping to uncover new insights about the evolution of the field. The significance of this work is two-fold. First, because learning analytics is a relatively new field attracting participation of both academics and practitioners, many community members have not published in conference proceedings or relevant academic journals, or are not intended to publish at all; as a result, the analysis of data from academic publications, as those included in the LAK Open Dataset, falls short in revealing the reach and development of the community. In contrast, Twitter as an information sharing and social networking platform broadly used at LAK conferences affords us with authentic, multimodal data from a richer pool of "participants" besides those who have published in the learning analytics literature. Second, Twitter supports rich, real-time social interactions among conference participants, in the form of retweeting, mentioning, and replying, which could facilitate meaningful exchanges not supported by traditional academic publishing venues. By analyzing social interactions on Twitter—by identifying leaders, characterizing information diffusion patterns, and detecting sub-communities, for instance—we could obtain a more vivid picture of community dynam-

LAK '15, March 16 - 20 2015, Poughkeepsie, New York, USA

http://dx.doi.org/10.1145/2723576.2723584

ics of learning analytics. Therefore, the analysis of Twitter archives, or in our words, "Twitter archeology" of LAK conferences could potentially reveal new insights about the learning analytics community.

In this article, we start with a brief introduction to Twitter and Twitter analytics. Then we introduce the LAK Twitter dataset and the analytic approaches applied in the study. After that, we present and discuss results from our analysis and conclude by discussing limitations and future directions of this study.

2. TWITTER

Twitter is, as we write, an online social networking service that enables users to share short messages known as *tweets*. While Twitter is normally conceptualized as a social network, or a microblogging service, it has essentially grown into an information or news network [27]. Because of the agileness offered by its 140-character limit, it has emerged to become "a personal news-wire," in Twitter's own words [41], on which all types of world events are posted and further spread through *retweeting*. These user behaviors collectively give rise to trending topics and facilitate large-scale social phenomena, such as Haiti earthquake relief efforts [40] and "Arab Spring" [31]. Thanks to Twitter's nature as a personal news platform, it has also contributed to transforming journalism, by changing how people become aware of news [18] or how journalists engage with their profession [28].

Twitter has been widely used at conferences, from its impressive "debut" at the 2007 South by Southwest Interactive (SXSWi) conference to almost every academic conference the authors have attended in recent years. At conferences, Twitter could be used to establish a backchannel to enable richer communication among attendees and extend conversations beyond the conference venue [2]. Within a traditional academic conference setting, the space is normally divided into a "front stage" for the speaker and a large "back area" for the audience [34]. In this context, attention is solely focused to the front and interactions are usually limited to Q&A periods. Because of the constrains posed by time and space, opportunities provided for the audience to interact with each other and to collectively construct understanding of a given speech are usually rare. As a result, the traditional conference model could cause feedback lags, stress for asking questions, and decreasing participation [1, 15]. Twitter could help close the gap of social interactions at conferences, thanks to its simplicity and the ubiquity of Internet connection. Conversationality and collaboration afforded by Twitter, largely through mentioning (`@`) [19] and hashtagging (`#`) [21], could help mitigate the disconnect among conference participants. Not surprisingly, Twitter has become widely used at academic conferences.

3. TWITTER ANALYTICS

Because the extensive use of Twitter in various social sectors, the analysis of Twitter data carries the potential to offer actionable knowledge for stakeholders, or to help us discover information diffusion paradigms on social media. For example, sentiment analysis of tweets has also been broadly applied to understand customer perception of certain products or brands [8, 11], or to characterize presidential debates by combining tweets with live television programs [12]. New information diffusion mechanisms could also be discovered from various aspects of Twitter usage such as social linkages [27] and retweeting behaviors [42]. In education, Twitter and other social media platforms are increasingly used in classrooms to facilitate communication between teachers and students, as well as among students [17, 25, 26]. Combining data-driven approaches and ethnographic approaches, researchers use Twitter data to investigate the unique online culture among the digital natives [22, 7], students identity performance on social media [24], and college students' learning experiences [10].

3.1 Twitter Analytics in the Academic Conference Context

Twitter usage at academic conferences has also attracted some research attention. Previous studies have mainly focused on three aspects. The first aspect centers on users and usage of Twitter at academic conferences. For example, some studies seek to understand who use Twitter at conferences, why they use it, and how. Using a survey, colleagues identify attendees, online attendees, speakers, and organizers of conferences as the main user groups of Twitter at a conference [15]. Through content analysis of tweets, researchers also distinguish seven main purposes of using Twitter during conferences, including: comments on presentations, sharing resources, discussions and conversations, jotting down notes, establishing an online presence, and asking organizational questions [34].

The second category of research focuses on interaction on Twitter during conferences. For instance, Social Network Analysis of online interactions of Twitter users identified different types of users, characterized by different levels of participation and influence [9]. Visual analytics platforms are designed to facilitate Twitter users' interaction during conferences [13].

The last cluster of research centers on the effect of using Twitter for academic conferences. For example, one study explores whether the use of Twitter enhances conference experience, collaboration, and collective construction of knowledge [34]. Others incorporate timeline analysis and Social Network Analysis together to study whether the use of Twitter could help reach a broader audience [29].

The present study has a unique agenda different from all the three categories. It is the very first study, as far as we know, attempting to track the evolution of an academic field by mining Twitter data from its annual conferences. To achieve this goal, we attempt to answer the following major research questions through in-depth analysis of the LAK Twitter archive:

- To which extent did Twitter enable participation and conversation in each year's LAK conference, characterized by the occurrences of tweets, retweets, and replies?
- Supposing Twitter participants could approximately represent the learning analytics community, how did the composition of Twitter participants change over the years?
- What does the social networks of LAK Twitter participants look like? Who are the influential figures in the community? To which extent did the community dynamics change over the years?

- What are the underlying topics in LAK Twitter participation? To which extent did the topics evolve over the years?

4. METHODS

4.1 Dataset

The dataset was aggregated through the official LAK conference hashtags, i.e., `#LAK11`, `#LAK12`, `#LAK13`, and `#LAK14`, and archived using the Twitter Archiving Google Spreadsheet (TAGS).[1] A total of 10,736 tweets by 1,217 unique Twitter users were archived in this dataset. An overview of Twitter participants[2] and tweets is provided in Table 1.

Conference	Participants	Tweets
LAK11	215	1358
LAK12	606	4050
LAK13	280	2223
LAK14	362	3105

Table 1: Overview of Dataset

4.2 Preprocessing

Before any actual data analysis, substantial efforts were put to clean the dataset. Because TAGS and the Twitter API have been evolving over the years, inconsistencies were evident in the multi-year dataset. For example, it was until 2012 when the Twitter API would return an `entities_str` that encapsulates all information related to a tweet; in the 2011 dataset, to whom a tweet was addressed was not provided, whereas such information was stored in later archives in a `to_user` field. More importantly, the biggest challenge we faced when cleaning the data was a systematic mistake of user ids in the 2011 archive. To fix this issue, we replaced one's user id in 2011 if the user could be found in later archives; otherwise, we used the current Twitter API to retrieve the user id. In addition, we paid special attention to track users who changed their screen names over the years, by tracking their user ids and replacing the obsolete names with the newest ones.

After data cleaning, further parsing was conducted at the tweet level. Specifically, if a tweet was identified as a retweet (i.e., starting with "RT @user: ..."), the user from whom this tweet was retweeted was extracted; if a tweet was identified as a reply, the user(s) to whom this tweet was addressed to were also parsed.

The cleaned data was saved into a set of comma-separated values (CSV) files each containing tweets, users, retweets, replies, mentions, and hashtags for later analysis. Other formats such as `.Rdata` and `JSON` were also created to meet needs within our team.

4.3 Data Analysis

To answer these research questions, we conducted a range of analysis on the LAK Twitter dataset.

[1]Martin Hawksey's Twitter Archiving Google Spreadsheet (TAGS), version 3 and 5.

[2]In the following sections of this paper, we are using "participants" to denote Twitter participants of LAK conferences, unless it is specified otherwise.

4.3.1 Desriptive Analysis

To get a basic understanding of Twitter participation at LAK, we first conducted descriptive analysis on the dataset. In addition to the summary presented in Table 1, we produced summarizing statistics with regards to retweets and replies in each year's conference. We also conducted descriptive analysis at the user level, computing the means of tweets, retweets and replies sent by each user, as well as the average numbers of times each user got retweeted or was replied to. Comparisons were made across years to uncover possible changes.

4.3.2 The "Flow" of Twitter Participants at LAK

To understand the composition of Twitter participants at LAK, we tracked new and returning participants across four conferences. For each participant, we identified the year(s) he or she participated. A Sankey diagram, which is commonly used to visualize energy or material flow [33], is produced to visualize the flow of Twitter participants at LAK. Further descriptive statistics were conducted to understand different types of participants defined by their appearance over the years.

4.3.3 Interaction Social Networks

Interaction network graphs were generated based on retweet, reply, and mention interactions. These network graphs were directional. For example, if user A retweeted user B, a directional edge would be drawn from user A to user B. We initially focused on *reciprocate rate*, an important network measure for a directional network. This concept is defined as the percentage of the pairs that have edges pointing to each other among all connected pairs of nodes. A higher reciprocate rate would imply a more egalitarian network. We also analyzed other characteristics of each year's interaction network, including *average degree*, *network diameter*, *average path length*, and the proportion of the *largest connected component*.

Because retweeting is an essential mechanism for information propagation in Twitter [42], we further studied retweets to determine who is influential in the network or how many people a piece of information reaches [45]. In particular, community detection was performed to identify the influential figures and explore the online community development over the four years. This analysis was composed of the following steps: (1) Because clustering algorithm is sensitive to outliers, extreme outliers have to be removed [44]. In this study, giant component filter algorithm is applied to the data in each year and expected to eliminate the outliers vertices (users); (2) after getting rid of the outliers, we performed the community detection (clustering) analysis using the fast unfolding algorithm [6]. According to [6], this algorithm starts with assigning each node into a different community in the network. Then the algorithm evaluates the gain of modularity by placing a node into another community. The node will stays in the original community if no positive modularity is obtained. The previous process is repeated until no further improvement of modularity is possible. On the other hand, to affirm that consistent clusters (communities) across different runs, the fast unfolding community detection algorithm is executed 20 times on each year's network and the highest frequency number of communities appearing in those networks are chosen as the resulting communities.

4.3.4 Evolution of Topics in the Community

In addition to analyzing participants and social interactions among them, another important aspect was to understand the topics discussed on Twitter and their changes over time. We first did a hashtag analysis. Hashtag is an important Twitter mechanism that users use to signal the central topics expressed in their tweets. We took the occurrences of hashtags and generated a hashtag cloud for each year. The hashtags `#LAK11` to `#LAK14` were removed because they appeared in all tweets from respective years. Then we inspected the hashtag clouds for popular topics in each year and their changes over the years.

Second, topic modeling was applied to uncover underlying topics in the tweets. In particular, Latent Dirichlet Allocation (LDA) [5] was used. The analytical process included the following steps:

(1) *Text sanitizing.* First of all, text that was semantically less meaningful or irrelevant was removed. Such text included Twitter users' screen names, links, Twitter-specific syntax (e.g., "RT", "via"), and special character encodings (e.g., `&`). We also decided to remove hashtags because they tended to distort the semantic space because of their high frequencies.

(2) *LDA and visual exploration.* Second, we adopted the `topicmodels` R package to model the topics on the sanitized text. To identify the optimal number of topics for topic modeling, we tested with a sequence of numbers from 1 to 100, which are suitable for the size of dataset based on our experience. The model selection was made based on the harmonic mean of the estimated log-likelihood of each model [32]. After choosing the optimal topic model, we then used `LDAvis` to assist visual exploration and interpretation of extracted topics. Compared to the turbo topics [4], which has been applied on LAK literature data [36], LDAvis enabled interactive exploration and clustering of topics. Using `LDAvis`, we were able to interactively make sense of the topics, assign names to meaningful topics, and cluster them into several categories.

(3) *Tracking selected topics.* Finally, we chose to track the development of certain research topics of learning analytics over the four years. Note that LDA would assign a most probable topic to each tweet [5]. We were then able to count how many times each topic has appeared in each conference. Data were further visualized for interpretation.

5. RESULTS AND DISCUSSION

5.1 Descriptive Analysis of Twitter Participation and Conversation

Table 2 presents descriptive statistics of each year's tweets. With the exception of LAK12, the numbers of participants and tweets have been increasing over the years. The counts and percentages of retweets and replies also increased overall, with the exception of LAK12 again.

To understand the reason why LAK12 had more participants and tweets, we specially consulted with the conference organizers. The explanation was that a substantial amount of tweets during LAK12 was about the technologies adopted for live video streaming. Tweets were posted to illustrate how the streaming technologies could be used in such a context. As a result, folks who may not be interested in Learning Analytics per se but more in video streaming and recording were attracted to the conversation.

Conference	Tweets	Retweets	Replies
lak11	1358	450 (33.1%)	230 (16.9%)
lak12	4050	1207 (29.8%)	430 (10.6%)
lak13	2223	570 (25.6%)	363 (16.3%)
lak14	3105	1255 (40.4%)	570 (18.4%)

Table 2: Descriptive Analysis of Tweets at LAK

Descriptive analysis of Twitter participants is presented in Table 3. Overall, the number of tweets, retweets, and replies have been improving over the years, except for LAK12 probably because of its broader participation. This result indicated growing participation and interactivity within the learning analytics community. The standard deviation of each measure also increased over the years, showing increasing disparity of participation. This could be partially accounted by the increase of participants over the years (see Table 1), most of whom were peripheral on Twitter discussion. Even though further tests failed to confirm a power law distribution on these measures, the distribution of them was extremely positively skewed.

Overall, descriptive analysis at the conference and participant levels indicated extending reach of the LAK community and increasing interactions among its members over the years.

Conf	Tweets	RTs	Replies	RT-ed	Replied
lak11	6.3 (14.1)	2.1 (4.7)	1.1 (3.5)	2.0 (7.5)	0.9 (3.4)
lak12	6.7 (23.8)	2.0 (5.0)	0.7 (2.9)	1.9 (13.4)	0.6 (3.7)
lak13	7.9 (31.5)	2.0 (4.5)	1.3 (7.4)	2.0 (7.7)	1.1 (4.2)
lak14	8.6 (30.9)	3.5 (10.4)	1.6 (7.5)	3.5 (13.5)	1.5 (6.0)

Table 3: Means and Standard Deviations of Tweets, Retweets (RTs), Replies, Times Being Retweeted (RT-ed), and Times Being Replied for Participants

5.2 The Flow of the LAK Community

To understand the composition of the LAK community as reflected by Twitter participation, we tracked participation of all Twitter participants over the years, focusing on new and returning participants each year. The results are visualized as a Sankey diagram presented in Figure 1. In this diagram, the horizontal axis represents the time dimension, with each column representing one conference. The fifth (or last) column represents participants who stopped participating at some point. Within each column, new and returning participants are presented separately. All lines are drawn from left to right, representing the flow of participants from one section to another between two columns; the width of a line denotes the volume of its flow. For example, for participants who participated in LAK11 and never returned, the line will be drawn from LAK11 directly to "Leave" in the last column; for those who participated in both LAK11 and LAK12, they are represented by the line from LAK11-New to LAK12-Back. Using this visualization, we can easily inspect the flow of Twitter participants across the years.

To our surprise, only a small fraction of participants have been returning to LAK conferences' Twitter discussion, indicated by the thinner lines towards the "Back" sections of LAK12 to LAK14. By looking at the returning participants

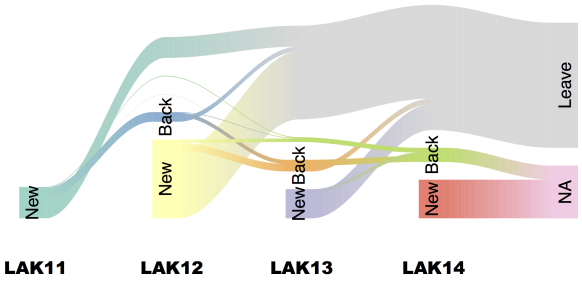


Figure 1: Flow of participants from LAK11 to LAK14.

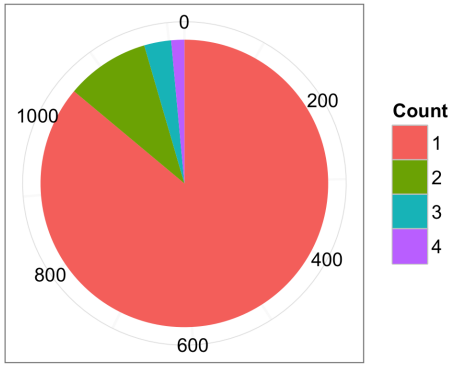


Figure 2: Twitter participants by the count of conferences.

Participants	LAK11	LAK12	LAK13	LAK14
sheilmcn	1	224	202	19
gsiemens	68	142	33	61
houshuang	106	1	10	115
sbskmi	52	51	19	58
dan_suthers	2	56	99	6
dougclow	35	43	46	29
bodong_c	6	2	12	90
dgasevic	7	66	9	9
R3beccaF	2	23	11	51
cab938	18	47	14	3
shaned07	5	40	19	13
mhawksey	5	23	26	13
abelardopardo	16	27	8	7
ErikDuval	8	24	21	1
aneesha	19	1	1	1
helinur	4	9	1	2
cteplovs	4	1	5	2
georgekroner	1	8	2	1

Table 4: Twitter Users Participating in All LAK Conferences, Sorted by the Total of Tweets

more closely, we found that only 18 colleagues (among all 1,217 unique participants) participated in Twitter discussion during all LAK conferences (see Table 4). This number only increased to 55 when we counted users participating at least three conferences. Overall, as indicated in Figure 2, the majority of participants only participated in one conference and never returned. Thus, although the learning analytics community is having a broader reach shown in the previous section, many participants remain peripheral.

In particular, the action of leaving was especially popular for new participants each year. As shown by the outbounding lines from "New" participants each year, most of them direct to the "Leave" category, indicating these new participants never participated again. In contrast, returning participants were much more likely to return, indicated by the more balanced divide between leaving and returning within participants in the "Back" category in each column.

Based on these observations, the LAK community seems to be "fluid" reflected by Twitter participation during its annual conferences. While it keeps attracting interested colleagues from various domains, the community is still more or less unstable as an emerging field of research and practice.

5.3 Interaction Network Analysis

Interaction networks were generated based on retweet, reply, and mention actions. As shown in Table 5, the reciprocate rates were increasing (except LAK12), implying increased interactions among the participants. The increased average degree, decreased average path length and network diameter, and overall increased percentage of nodes contained by the largest connected network component indicated that the network was becoming denser and more connected.

In the network graphs in Figure 3, the node size and color are based on betweenness centrality. Betweenness centrality is a centrality measure of a node within a network. It denotes a node's position within a network in terms of its ability to bridge the connection between other node pairs or groups in the network. Hence, the nodes with larger betweenness centrality are more influential and function as bridges connecting the community. As shown in the network visualizations, from LAK11 to LAK12, the network became much larger; however, due to the reasons explained earlier, a large portion of nodes in the LAK12 network are in peripheral positions, only loosely connected through one influential figure. There are also less nodes in the center, implying the discussion was led by only a few key figures. In LAK13 and LAK14, an increasing number of nodes gain higher betweenness centrality and emerged to connect the community more tightly. Hence, when the LAK Twitter community had an increasing reach and interaction, a larger group of "leaders" (at least in terms of Twitter participation) have also been emerging.

5.4 Community Detection in the Retweet Networks

Community detection in the retweet networks generated 3, 6, 5, 6 communities respectively for LAK to LAK14. The retweeting network graphs are visualized in Figure 4. In these graphs, different communities are coded in different colors and the different levels of influence of participants within a network, measured by betweenness centrality, is scaled to the node size.

Overall, except for LAK 12, the numbers of communities increase steadily from 3 to 6 in LAK14, along with the

Conf	Nodes	Edges	Reciprocate Rate %	Avg. Degree	Diameter	Avg. Path Length	Largest Component %
lak11	215	569	13.1	2.65	7	2.95	90.70
lak12	606	1521	12.8	2.51	9	3.10	83.33
lak13	280	736	14.9	2.63	8	2.99	87.50
lak14	362	1369	19.5	3.78	6	2.73	91.71

Table 5: Interaction Network Characteristics

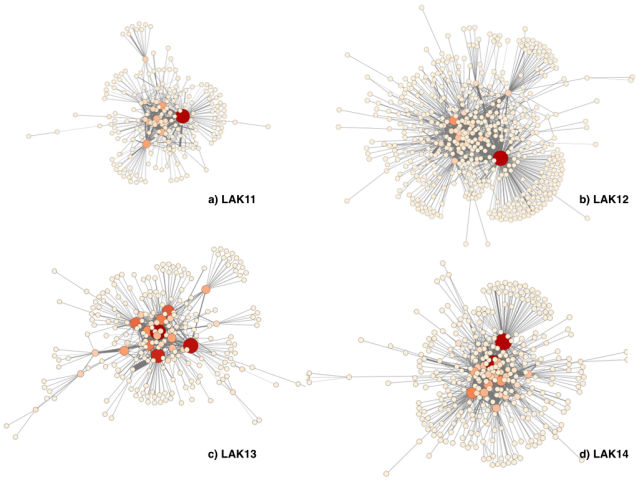


Figure 3: The largest connected interaction network components from LAK11 to LAK14. *Note*: Node size and color are based on betweenness centrality.

increasing numbers retweets from 450 to 1255. The complexity of the network structure was also enhanced. In particular, because every detected community in the networks was usually dominated or centered on an influential figure, the increased numbers of communities means the emergence of new hubs in Twitter retweeting networks. This finding is consistent with the observations of emerging leaders in the previous section.

Second, we also found the connections among members inter or intra communities were not that extensive, except for more dominant participants. Given a community is normally centered around a leading figure, the following situation might often take place: a community leader generates a tweet, and then a leader from another community retweet it and further diffuse the idea to his or her entire community.

Furthermore, the sizes of various communities are more balanced in LAK13 and LAK14, reflecting more participants from various disciplines contributing to the LAK conferences. The connections within and between communities in LAK13 and LAK14 were significantly improved as well. More exchanges among different scholars can contribute to the development for long-term viability for the learning analytics community. These findings are also consistent with the interaction network analysis in the previous section.

5.5 Hashtag Analysis

In the hashtag analysis, the hashtags `#LAK11` – `#LAK14` and `#LearningAnalytics` took dominant proportions and were therefore removed from the hashtag cloud visualizations. However, we also noticed that the dominance of `#LAK1*` became weaker over the years, implying the emergence of more

Conf	Hashtags	Freq	#lak1* %	#LearningAnalytics %
lak11	113	1817	74.7	0
lak12	281	5820	69.5	24.0
lak13	177	2900	76.7	13.9
lak14	197	3890	71.0	19.6

Table 6: Descriptive Statistics of Hashtags

hashtags. This finding indicated that the research topics in the community have become more diverse, connected with an increasingly deepening inquiry in the community.

Interesting trends could be observed from the hashtag clouds in Figure 5. Some hashtags were popular in a specific year but eventually faded away later. For example, `#edchat`, `#edtech20`, and `#edtools` were very dominant in LAK11, but did not appear in the successive years. The fade of these more general hashtags and the rise of `#LearningAnalytics` (see Table 6) implied the formation of a collective community identity.

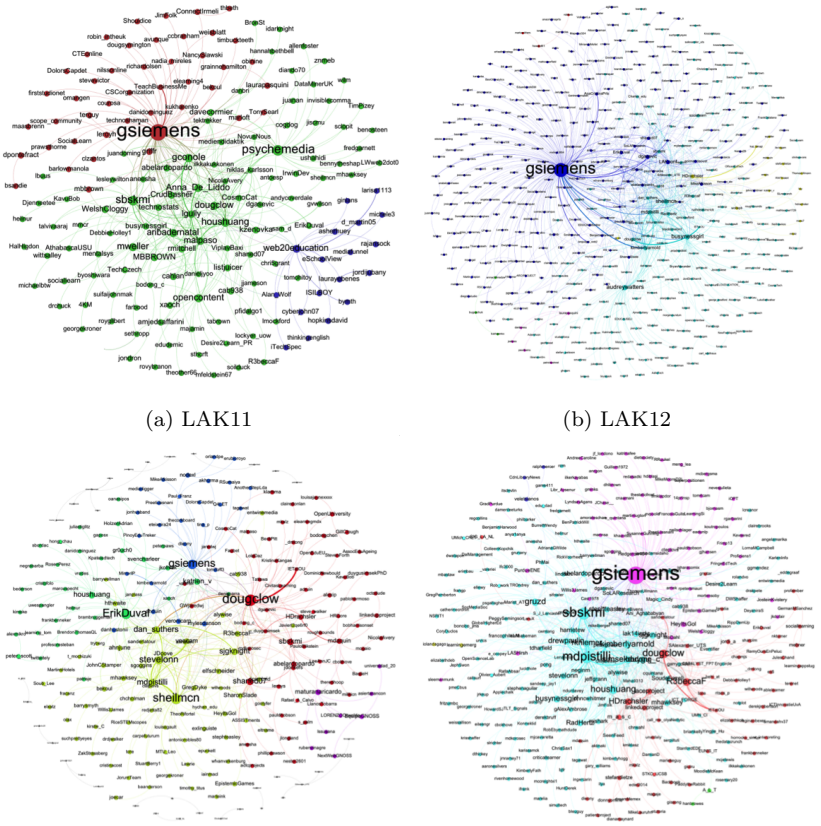
In addition, the popularity of some hashtags in a specific year was related to promotion efforts of certain workshops, projects or technologies. For instance, the hashtags `#elifocus` and `#EduLive` relevant to the video streaming technologies were popular in LAK12, but did not appear again in LAK13 and LAK14. `#Linkedupproject` and `#plasma`, which were related to two learning analytics projects, had some momentum in LAK13 and LAK14 respectively. `#DCLA13` and `#lakdata14` were respectively related to the Discourse-Centric Learning Analytics workshop in 2013 and the LAK data challenge in 2014.

Other than these promotion hashtags, a few hashtags emerged and persisted over the years. For example, data-related hashtags and `#MOOC(s)` appeared from the first year and represented long-standing interests within the community. `#EDM`, `#DataMining`, and `#BigData` started to become more pervasive in 2012, indicating the bridging between the learning analytics and EDM communities. Though these terms may be overshadowed by some of the most promoted hashtags, they were the most persistent topics in the community.

5.6 Topics Modeling of Twitter Discussion

Going beyond hashtag analysis, we applied LDA to uncover underlying topics in Twitter discussion during LAK conferences. The harmonic mean of the log-likelihood per number of topics is plotted in Figure 6. The maximum is reached when the LDA model was trained with 34 topics. The 34-topic model that best accounted for the corpus was chosen accordingly.

LDAvis [39] was then used to help interpreting the topic model. Figure 7 to 9 illustrate three screenshots of LDAvis during our exploration of the topic space. To enable visualizations of topics, LDAvis first projected the 34-dimensional space to 2D using multidimensional scaling. Each topic



(a) LAK11 (b) LAK12

(c) LAK13 (d) LAK14

Figure 4: Largest connected components in retweet networks.

(a) LAK11 (b) LAK12

(c) LAK13 (d) LAK14

Figure 5: Hashtag clouds. *Note*: Hashtags #LAK1* and #LearningAnalytics were removed for each year.

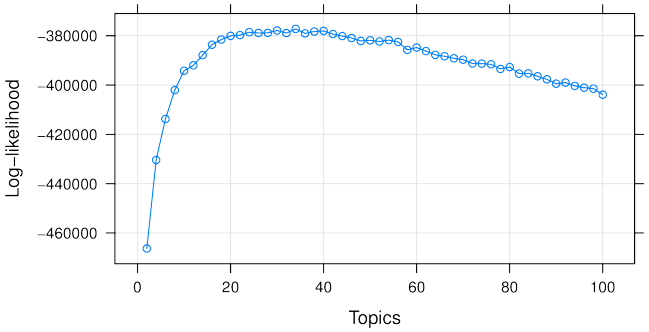


Figure 6: The harmonic mean of estimated log-likelihoods per number of topics. *Note*: The optimal number of topics is 34 when the maximum log-likelihoods is observed.

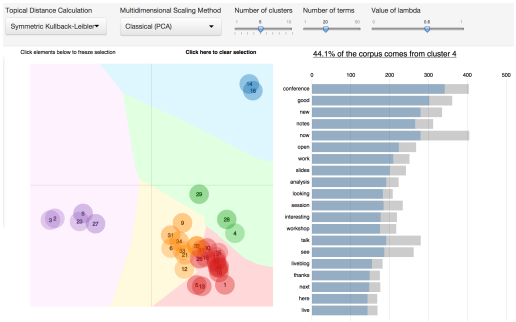


Figure 7: Red cluster is chosen, and corresponding top terms are displayed.

could then be represented by a circle in the 2D space. Using certain distance calculation algorithms, LDAvis could further cluster topics based on their distance among each other. After some exploration, 5 clusters appeared to be the most interpretable choice. By clicking on each cluster in the left panel, the top terms corresponding to a cluster would be updated in the right panel (Figure 7). The red cluster corresponded to terms such as {conference, good, notes, work, slides, session} and appeared to be related to "meta-information" of conferences; the green cluster featured terms including {like, best, great, interested, cool} and was related to positive sentiments about the conferences; the blue and yellow clusters, being associated with {data, big, mining, educational, challenge} and {student(s), social, research, use, course} respectively, were related to specific research topics; the violet cluster was linked to terms like {learning, analytics, knowledge, paper(s), conf, journal} that were more general in the LAK context.

We further explored each individual topic, by clicking on its corresponding circle (Figure 8). The distribution of one specific term among topics could also be inspected, by hovering over the term in the right panel (Figure 9). Results indicated that the trained LDA model was meaningful, in that most topics were interpretable based on their terms and topics under a same cluster were semantically closer. Similar to the findings from topic clusters, we found tweet topics generally fell into a few distinguishable categories, including (1) information-sharing related to conferences and the community, (2) experience-sharing and comments, and (3) more specific research topics (such as MOOC, assessment,

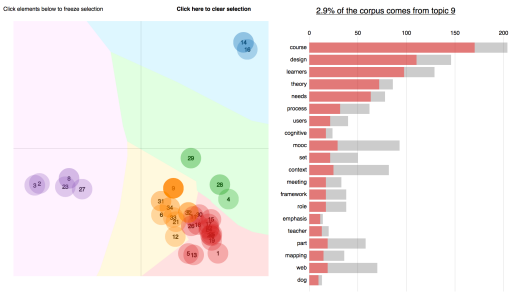


Figure 8: Topic 9 is chosen, top terms displayed.

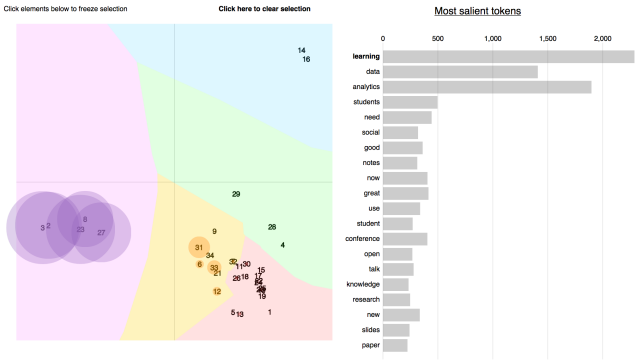


Figure 9: Term "learning" is chosen, relevant topics enlarged.

students, course design).

Based on the exploration, we tagged each topic with its pertinent terms and then focused on topics most relevant to learning analytics research and tracked their evolution over the years. In Figure 10, the change of eleven topics is illustrated. The y-axis represents the percentage of a topic in a year's tweets. As a validity check, two topics popular in LAK14 are included: Topic 10 {social use media win networks share survey ipad} related to a promoted project and Topic 11 {graesser systems predictive agents model tutors} related to a keynote speech. These two topics both peeked in LAK14, showing a certain level of validity of LDA.

Further inspection of the topics identified an increasing emphasis on students, need, assessment, and feedback in the community, indicated by the growing popularity of Topics 1-4. In addition, each year's conference presented unique hot topics: topics relevant to ethics and social media were relatively popular at LAK11; big data, linked data, and educational data mining were trending at LAK12; course design, ethical issues, discourse, and measurement were popular during LAK13; topics involving assessment, student needs, intelligent tutors, and educational data mining were rising at LAK14. Comparing with previous analysis of learning analytics publications, the popularity of student-related topics appeared to be shared by LAK academic literature and tweets [46]; however, the evolution of topics in tweets tended to not agree with the analysis of publications [36].

6. CONCLUSIONS

The present study built on previous research that set to understand the field of learning analytics from a variety of angles. Using a unique set of Twitter data from previous LAK conferences, we aimed to uncover new insights about

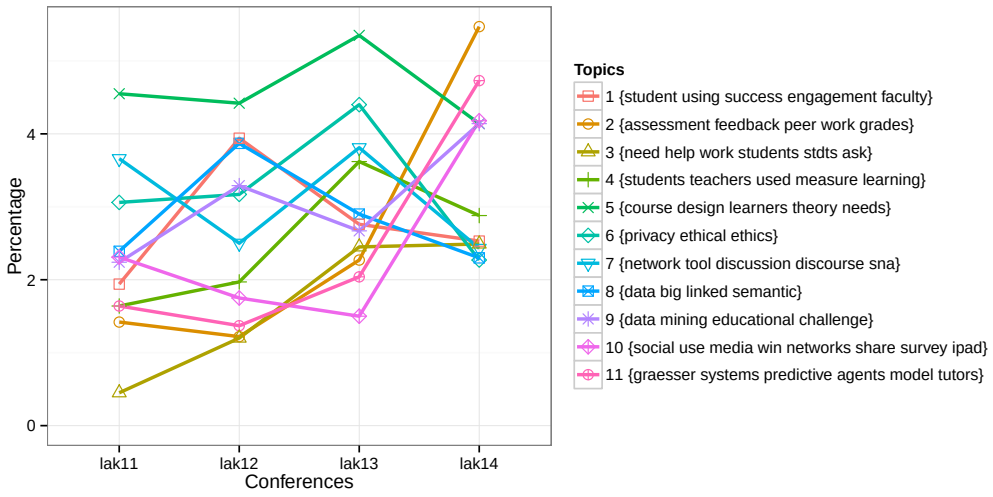


Figure 10: Tracking changes of selected research topics.

the community combining multiple analysis attending to different aspects of Twitter participation. Through descriptive analysis, interaction network analysis, hashtag analysis, and topic modeling, we found an extended reach of the community and increasing interactions among its members; increasingly dense, connected, and balanced social networks; peripheral and in-persistent participations; and more and more diverse research topics. In particular, detailed inspection of semantic topics identified a rising emphasis on students over the years as well as distinctive hot topics in each year's conference.

We would like to mention a few limitations or potential risks of the present study. First, we were not able to test the comprehensiveness of the Twitter archive. Based on our experience with TAGS, some tweets could have got lost for various reasons (e.g., the time an archive was created). Second, because of the 140-character limit, tweets might be less suitable for in-depth semantic analysis. While most topics we identified were meaningful, some topics were hard to interpret. Third, since not all conference participants or community members use Twitter, the analysis of tweets could only reconstruct parts of the dialogues and would run the risk of missing important messages, events, or figures. For future directions, we would like to connect tweets and academic publications, to further construct a more integrated picture of the learning analytics community.

7. REFERENCES

[1] C. Anderson. *The long tail: Why the future of business is selling less of more*. Hyperion, 2008.

[2] C. Atkinson. *The backchannel: how audiences are using Twitter and social media and changing presentations forever*. New Riders, 2009.

[3] N. Balacheff and K. Lund. Multidisciplinarity vs. Multivocality, the Case of Learning Analytics. In *Proceedings of the Third International Conference on Learning Analytics and Knowledge*, LAK '13, pages 5–13, New York, NY, USA, 2013. ACM.

[4] D. M. Blei and J. D. Lafferty. Visualizing topics with multi-word expressions. *arXiv preprint arXiv:0907.1013*, 2009.

[5] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3:993–1022, Mar. 2003.

[6] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, 2008.

[7] D. Boyd. *It's Complicated: the social lives of networked teens*. Yale University Press, 2014.

[8] W. Chamlertwat, P. Bhattarakosol, T. Rungkasiri, and C. Haruechaiyasak. Discovering Consumer Insight from Twitter via Sentiment Analysis. *Journal of Universal Computer Science*, 18(8):973–992, 2012.

[9] B. Chen. Is the Backchannel Enabled? Using Twitter at Academic Conferences. *2011 Annual Meeting of American Educational Research Association*, 2011.

[10] X. Chen, M. Vorvoreanu, and K. Madhavan. Mining Social Media Data for Understanding Students' Learning Experiences. *IEEE Transactions On Learning Technologies*, 7(3):246–259, 2014.

[11] D. Davidov, O. Tsur, and A. Rappoport. Enhanced sentiment learning using twitter hashtags and smileys. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pages 241–249. Association for Computational Linguistics, 2010.

[12] N. A. Diakopoulos and D. A. Shamma. Characterizing debate performance via aggregated twitter sentiment. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1195–1198. ACM, 2010.

[13] M. Dork, D. Gruen, C. Williamson, and S. Carpendale. A visual backchannel for large-scale events. *Visualization and Computer Graphics, IEEE Transactions on*, 16(6):1129–1138, 2010.

[14] H. Drachsler, S. Dietze, E. Herder, M. d'Aquin, and D. Taibi, editors. *Proceedings of the LAK Data Challenge 2014, held at LAK 2014, the 4th Conference on Learning Analytics and Knowledge (LAK2014), CEUR Workshop Proceedings, Vol. 1137*, 2014.

[15] M. Ebner, G. Beham, C. Costa, and W. Reinhardt. How people are using Twitter during conferences. *Creativity and innovation Competencies on the Web*, page 145, 2009.

[16] R. Ferguson. Learning analytics: drivers, developments and challenges. *International Journal of Technology Enhanced Learning*, 4(5):304–317, 2012.

[17] G. Grosseck and C. Holotescu. Can we use Twitter for educational activities. In *4th international scientific conference, eLearning and software for education, Bucharest, Romania*, 2008.

[18] A. Hermida. Twittering the news: The emergence of ambient journalism. *Journalism Practice*, 4(3):297–308, 2010.

[19] C. Honey and S. C. Herring. Beyond microblogging: Conversation and collaboration via Twitter. In *System Sciences, 2009. HICSS'09. 42nd Hawaii International Conference on*, pages 1–10. IEEE, 2009.

[20] Y. Hu, G. McKenzie, J.-A. Yang, S. Gao, A. Abdalla, and K. Janowicz. A Linked-Data-Driven Web Portal for Learning Analytics: Data Enrichment, Interactive Visualization, and Knowledge Discovery. In *LAK Workshops*, 2014.

[21] J. Huang, K. M. Thornton, and E. N. Efthimiadis. Conversational tagging in twitter. In *Proceedings of the 21st ACM conference on Hypertext and hypermedia*, pages 173–178. ACM, 2010.

[22] M. Ito, S. Baumer, M. Bittanti, D. Boyd, R. Cody, B. Herr, H. Horst, P. Lange, D. Mahendran, K. Martinez, et al. Hanging out, messing around, geeking out: Living and learning with new media, 2009.

[23] L. Johnson, R. Smith, H. Willis, A. Levine, and K. Haywood. The 2011 horizon report. *The New Media Consortium, Austin, Texas*, 2011.

[24] R. Junco. *Engaging Students through Social Media: Evidence-Based Practices for Use in Student Affairs*. John Wiley & Sons, 2014.

[25] R. Junco, C. M. Elavsky, and G. Heiberger. Putting twitter to the test: Assessing outcomes for student collaboration, engagement and success. *British Journal of Educational Technology*, 44(2):273–287, 2013.

[26] R. Junco, G. Heiberger, and E. Loken. The effect of Twitter on college student engagement and grades. *Journal of Computer Assisted Learning*, 27(2):119–132, 2011.

[27] H. Kwak, C. Lee, H. Park, and S. Moon. What is Twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web*, pages 591–600. ACM, 2010.

[28] D. L. Lasorsa, S. C. Lewis, and A. E. Holton. Normalizing Twitter: Journalism practice in an emerging communication space. *Journalism Studies*, 13(1):19–36, 2012.

[29] J. Letierce, A. Passant, J. G. Breslin, and S. Decker. Using Twitter During an Academic Conference: The# iswc2009 Use-Case. In *ICWSM*, 2010.

[30] G. R. Lopes, L. A. P. P. Leme, B. P. Nunes, and M. A. Casanova. RecLAK: Analysis and Recommendation of Interlinking Datasets. In *LAK Workshops*, 2014.

[31] G. Lotan, E. Graeff, M. Ananny, D. Gaffney, I. Pearce, et al. The revolutions were tweeted: Information flows during the 2011 Tunisian and Egyptian revolutions. *International Journal of Communication*, 5:31, 2011.

[32] M. Ponweiser. Latent Dirichlet Allocation in R, 2012.

[33] P. Riehmann, M. Hanfler, and B. Froehlich. Interactive sankey diagrams. In *Information Visualization, 2005. INFOVIS 2005. IEEE Symposium on*, pages 233–240. IEEE, 2005.

[34] C. Ross, M. Terras, C. Warwick, and A. Welsh. Pointless babble or enabled backchannel: conference use of twitter by digital humanists. *Digital Humanities*, 2010.

[35] M. Scheffel, K. Niemann, S. L. Rojas, H. Drachsler, and M. Specht. Spiral me to the core: Getting a visual grasp on text corpora through clusters and keywords. In *LAK Workshops*, 2014.

[36] M. Sharkey and M. Ansari. Deconstruct and Reconstruct: Using Topic Modeling on an Analytics Corpus. In *LAK Workshops*, 2014.

[37] G. Siemens. Learning analytics: envisioning a research discipline and a domain of practice. In *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge*, pages 4–8. ACM, 2012.

[38] G. Siemens and R. S. d Baker. Learning analytics and educational data mining: towards communication and collaboration. In *Proceedings of the 2nd international conference on learning analytics and knowledge*, pages 252–254. ACM, 2012.

[39] C. Sievert and K. Shirley. Ldavis: A method for visualizing and interpreting topics. In *2014 ACL Workshop on Interactive Language Learning, Visualization, and Interfaces*, Baltimore, June 2014.

[40] B. G. Smith. Socially distributing public relations: Twitter Haiti, and interactivity in social media. *Public Relations Review*, 36(4):329–335, nov 2010.

[41] B. Stone. Twitter As News-wire, July 2008.

[42] B. Suh, L. Hong, P. Pirolli, and E. H. Chi. Want to be Retweeted? Large Scale Analytics on Factors Impacting Retweet in Twitter Network. In *Social Computing (SocialCom), 2010 IEEE Second International Conference on*, pages 177–184, Aug 2010.

[43] D. Taibi and S. Dietze. Fostering analytics on learning analytics research: the LAK dataset. In *CEUR WS Proceedings Vol. 974, Proceedings of the LAK Data Challenge*, 2013.

[44] W. Xing, B. Wadholm, and S. Goggins. Learning analytics in CSCL with a focus on assessment: an exploratory study of activity theory-informed cluster analysis. In *Proceedins of the Fourth International Conference on Learning Analytics And Knowledge*, pages 59–67. ACM, 2014.

[45] T. R. Zaman, R. Herbrich, J. Van Gael, and D. Stern. Predicting information spreading in twitter. In *Workshop on Computational Social Science and the Wisdom of Crowds, NIPS*, volume 104, pages 17599–601. Citeseer, 2010.

[46] A. Zouaq, S. Joksimovic, and D. Gasevic. Ontology Learning to Analyze Research Trends in Learning Analytics Publications. In *LAK (Data Challenge)*, 2013.