Towards Understanding the Lifespan and Spread of Ideas: Epidemiological Modeling of Participation on Twitter

Sai Santosh Sasank Peri
LINK Research Lab
University of Texas at Arlington
Arlington, Tx, USA
saisantoshsasan.peri@mavs.uta.edu

Bodong Chen
Department of Curriculum and Instruction
University of Minnesota
Minneapolis, MN, USA
chenbd@umn.edu

Angela Liegey Dougall
Department of Psychology
University of Texas at Arlington
Arlington, Tx, USA
adougall@uta.edu

George Siemens
LINK Research Lab
University of Texas at Arlington
Arlington, Tx, USA
gsiemens@gmail.com

ABSTRACT

How ideas develop and evolve is a topic of interest for educators. By understanding this process, designers and educators are better able to support and guide collaborative learning activities. This paper presents an application of our *Lifespan of an Idea* framework to measure engagement patterns among individuals in communal socio-technical spaces like Twitter. We correlated engagement with social participation, enabling the process of idea expression, spread, and evolution. Social participation leads to transmission of ideas from one individual to another and can be gauged in the same way as evaluating diseases. The temporal dynamics of the social participation can be modeled through the lens of epidemiological modeling. To test the plausibility of this framework, we investigated social participation on Twitter using the tweet posting patterns of individuals in three academic conferences and one long term chat space. We used a basic *SIR* epidemiological model, where the rate parameters were estimated through Euler's solutions to *SIR* model and non-linear least squares optimization technique. We discuss the differences in the social participation among individuals in these spaces based on their transition behavior into different categories of the *SIR* model. We also made inferences on how the total lifetime of these different twitter spaces affects the engagement among individuals. We conclude by discussing implications of this study and planned future research of refining the *Lifespan of an Idea* Framework.

CCS CONCEPTS

• **Information systems~Data analytics** • Information systems~Social networking sites • **Computing methodologies~Modeling methodologies**

LAK '20, March 23–27, 2020, Frankfurt, Germany

ACM ISBN 978-1-4503-7712-6/20/03…$15.00
https://doi.org/10.1145/3375462.3375515

KEYWORDS

Ideas, Epidemiology, Engagement Patterns, Networked Learning, Knowledge Creation, Connectivism

ACM Reference format:

Sai Santosh Sasank Peri, Bodong Chen, Angela Liegey Dougall and George Siemens. 2020 Towards Understanding the Lifespan and Spread of Ideas: Epidemiological Modeling of Participation on Twitter. In *Proceedings of the 10th International Conference on Learning Analytics & Knowledge (LAK'20). ACM, New York, NY, USA, 6 pages.* https://doi.org/10.1145/3375462.3375515

1 Introduction

Learning involves complex processes that are shaped by cultural, cognitive and behavioral factors, enabled by social interactions [1], and supported by technological infrastructures. Starting from this assumption, many learning traditions stress the process of social interaction and participation.

To Lave & Wenger [2], learning is about participating in a community of practice. Other traditions emphasize the mediational and generative nature of learning and focus on the ideas and artifacts produced by learning. To Bereiter and Scardamalia [3], ideas have their ontological existence independent of learners, and deeper learning happens when a community of learners come together to collectively improve their ideas. In a similar vein, connectivist learning is considered as a network forming process with an emphasis on using digital technology for interaction [4, 5]. For these traditions of learning, instead of primarily assessing individual growth of knowledge, tracing the development of ideas in a socio-technical space is of significant interest as the structure of network in which knowledge evolves impacts how people learn. A key question to be answered is: *How does an idea come to exist, spread, and develop in a communal socio-technical space?*

In this paper, we present preliminary work of tackling this challenge by applying epidemiological modeling to the evolution of ideas. As an initial step towards tracing each epistemic agent's cognitive engagement with ideas, we turn to a much simpler construct---*social participation*---that enables the social process of idea exchanges and improvement. By doing so, we aim to show the plausibility of applying epidemiological modeling to this problem and pave the way for future efforts to combine other approaches, such as natural language processing and network analysis [6], to solve the key challenge. Below, we first introduce the "Lifespan of an Idea" framework developed based on epidemiological modeling. We then report an empirical study of applying this framework to four Twitter datasets in which social participation is central. We discuss findings from this analysis and ways in which the Lifespan of an Idea framework could be useful and expanded.

2. The Lifespan of an Idea Framework

2.1 The Lifespan and Spread of an Idea

Ideas are infectious in nature, they can be transmitted through interaction among individuals [7] and follow an epidemic process. Combining theories of connectivism [5], knowledge building [3], and epidemiology [8], we consider learning to be a networked process that involves the individuals to undergo a constant activity of encountering, considering, improving, and also discarding ideas while interacting with others in the network. In the process, ideas are spread from one individual to another through social contact, and the number of individuals fluctuates as people continually encounter and discard ideas in the socio-technical space.

Here, an analogy can be made between holding an idea with being infected with a disease. Previous studies on the spread of scientific ideas have determined that the transmission of an idea from one person to another can be correlated to the spread of infectious diseases [8]. The discipline of epidemiology has a long history of modeling the infection and spread of diseases to assess whether a disease is epidemic and to help inform public health interventions. Borrowing this idea, the process of idea development and transmission can be modelled as a function of time. The infectious nature of ideas can be then gauged in the same way as evaluating diseases. Ideas that are "epidemic" would spread quickly in a network, while some other ideas would fail to generate uptakes.

Based on this conceptual understanding, we apply epidemiological modeling to analyzing the engagement patterns among individuals in a socio-technical space. The temporal dynamics of the idea transmission can be represented in epidemiological terms. In the next section, we introduce the *SIR (Susceptible, Infected, Removed)* model, one of the more basic epidemiological models, adopted in our framework.

2.2 The SIR Epidemiological Model

We consider a homogenous mixture of individuals in a fixed population experiencing interactions with each other. These individuals are divided into categories reflecting their status of infection. Individuals in the population tend to transition from one category to another depending on the intensity of their interaction with the infection/infected people. These categories are defined as *Susceptible* (*S*)---people who are vulnerable to acquire an infection, *Infective* (*I*)---people who have acquired the infection at the beginning from some external source or have been infected by other individuals in the same population, and *Removed* or *Recovered* (R)---people who have attained immunity and are no longer susceptible to the infection. The system of differential equations (1)-(3) represent a basic SIR epidemiological categorical model [9], without considering the births, deaths and growth of the population. This basic model does not include vaccinations and mutations, which are factors in a real-world scenario with biological diseases. The Euler's solution to the *SIR* ordinary differential equations is shown in eq (4)-(7) [10, 11]. Where *n* represents the current population at time *t* and *n+1* represents the population at a time *t+1*. *Δt* is the time difference between *t+1* and *t*. The total population at any time *t* is equal to the sum of individuals in *S*, *I* and *R* categories.

$$\frac{dS}{dt} = -\alpha SI \tag{1}$$

$$\frac{dI}{dt} = \alpha SI - \beta I \tag{2}$$

$$\frac{dR}{dt} = \beta I \tag{3}$$

$$S_{n+1} - S_n = (-\alpha S_n I_n)\Delta t \tag{4}$$

$$I_{n+1} - I_n = (\alpha S_n I_n - \beta I_n)\,\Delta t \tag{5}$$

$$R_{n+1} - R_n = (\beta I_n)\,\Delta t \tag{6}$$

$$N = S_n + I_n + R_n \tag{7}$$

2.3 Applying the SIR Model to Studying the Spread of Ideas

Shifting from disease to idea spread, appropriate definitions are adopted for the different categories of the SIR epidemiological model. The *Susceptible* (*S*) class consists of individuals who are likely to get influenced and adopt an idea when they encounter an "infected" individual in future. The *Infective* (*I*) class are the individuals who are pursuing the idea and actively participate in its transmission to the other individuals by publishing their work in any format. The *Recovered* (*R*) class consists of individuals who have previously adopted and propagated the idea and probably published the work but their influence to infect new individuals is lost due to saturation of their network or they are no longer interested in the idea.

There have been several modifications to the basic SIR model for the spread of ideas to incorporate the changing population dynamics. Some models include the factors like population growth which is a measure of the rate of change of total population. [12]have added extra categories like *idea incubators* (*E*) and *skeptics* (*Z*) to incorporate the changing dynamics in a population and analyzed the spread of scientific ideas using *SEIZ* model. Individuals in an *incubator* class, are exposed to an idea for a brief period before they transition to become infectives. *Skeptics* are susceptibles who do not believe in the idea. Apart from the standard categorical models, researchers also applied weighted network models [13], to study the network structure and properties of individuals involved in the spread of ideas. The addition of new categories and network models help in better analysis of the idea transmission process. For our purposes, as one of the first studies to evaluate idea spread in learning networks, we focus on the traditional *SIR* model.

3. Case Study: Applying the Framework to Twitter Conversations

3.1 Datasets

We performed a comparative study to evaluate the use of the framework to understand the engagement patterns in different types of spaces on Twitter. For this purpose, we chose datasets from three academic conferences lasting less than one week and one weekly Twitter chat space that had a lifetime of 27 months. Among the conferences, two datasets were from different years of a same conference named Learning Analytics and Knowledge (LAK): LAK 2017 and LAK 2019. The other conference dataset was from International Literary Association (ILA) 2017.

The Twitter chat space was TXEDUCHAT, a state-specific chat space about education for the state of Texas, USA. The twitter data was collected using the official hashtags of the respective events. For the LAK conference, data was collected using Twitter Archiving Google Spreadsheet (TAGS) as previously reported [14]. For the ILA and TXEDUCHAT, tweets were collected using the official hashtags through a Twitter scraping algorithm developed in Python. An overview of the number of participants, tweets, and timeline is provided in Table 1.

Table 1: An overview of the duration, number of authors and tweets for all the Twitter datasets.

Twitter Hashtag	Time Period	Total number of unique authors (*N*)	Total number of tweets
LAK 2017	Mar 12-18, 2017	539	3814
LAK 2019	Mar 3-9, 2019	393	2695
ILA 2017	Jul 13-18, 2017	6070	24328
TXEDUCHAT	Dec 2016-Mar 2019	15839	65117

3.2 Adaptation of SIR Model to Twitter Data

The individuals in a Twitter environment are divided into SIR categories based on their activity of posting tweets. We considered only those tweets that used the official hashtags of the conferences and chat space. As an initial study to establish plausibility of the framework, we modeled participation in Twitter dialogues instead of ideas expressed in tweets. Specifically, current authors posting a tweet are in the ***infective*** (*I*) class. After a certain amount of time, these posts lose their influence and the authors are moved into the ***removed*** (*R*) class. All the other individuals on the Twitter group who would be future authors belong to the ***susceptible*** (*S*) class. Authors who re-tweet or reply to an existing post are also part of the infective class. For tweets from conferences, we assumed the influence of a tweet is lost after one day, so we build the removed class daily from the previous day's authors. For the chat space, we assumed that tweets would lose infectivity at the end of a month.

3.3 Preprocessing

Data was preprocessed in preparation for the SIR modeling. The TXEDUCHAT and ILA conference tweets were collected in a json format. We extracted all the Twitter variables of interest such as author user ID, text of tweet, timestamp, hashtags from json files and saved into a comma separated values (CSV) file. The LAK tweets from TAGS were also preprocessed for the same variables and stored as CSV format files. The timestamp was converted into a Month, Day, and Year format for our convenience.

3.4 Data Analysis

We followed a step-by-step process to compute the rate parameters of our datasets and analyzed the temporal dynamics of the population when individuals moved from one category to other. We considered two sets of SIR categories, one consisting the measured values, $SIR_{measured}$, (S_{n+1}-S_n, I_{n+1}-I_n, R_{n+1}-R_n), which were the original values from the twitter dataset. The other set consisted of estimated values, $SIR_{estimated}$, ($-\alpha S_n I_n \Delta t$, $[\alpha S_n I_n - \beta I_n]\Delta t$, $\beta I_n \Delta t$) which were obtained from the Euler's solutions (eqs (4)-(6). Timepoint n is a cumulative value of days (for conference datasets) or months (for chat space dataset) depending on the dataset. Δt is the time difference between $n+1$ and n. The $SIR_{estimated}$ values were calculated using a guess value for rate parameters α and β. The objective was to estimate the true values of rate parameters by minimizing the difference between the measured and estimated variables for each category of the *SIR* model. Section *3.4.1* describes how we computed the values for the *I* category. The methods for calculating number of individuals in the SIR categories of the measured variables and estimated variables is discussed in the sections *3.4.2* and *3.4.3* respectively. Section *3.4.4* discusses the procedure to perform the parameter estimation through error minimization.

3.4.1 Distribution of individuals into I category of SIR. Based on the assumption that the authors of a tweet belong to the *I* category, we calculated the number of authors per day (for conference datasets) and per month (for chat space dataset) to compute the total number of individuals in the *I* category of our Twitter datasets at every *n*. This was achieved by creating a pivot table from the datasets using the variables, author user ID, text of the tweet and timestamp of the tweet. From this pivot table we computed the total population (*N*) which is the total number of unique author user IDs during lifetime of the dataset (refer to Table 1). The pivot table also provided information about the total number of tweets for every hashtag for the duration of the conference and chat space considered in this analysis (refer to table 1).

3.4.2 *Calculating the measured values Sn, I_n, and R_n.* The SIR category models are time dependent, so we represent time as n. We considered n to be an incremental value of each day of the conference with n=0 being one day prior to the official start date of the conference and last value of n corresponds to one day after the official end date of the conference. For the chat space, we assumed n=0 to be Dec 2016 with n value increasing in increments of 1 month until March 2019. We started the measurement of S_1, I_1, R_1 ($n=1$) based on values at $n=0$, where the value of R_0 is null, I_0 is the initial number of authors, and S_0 is rest of the population. The measured values of *infective* category (I_{n+1}-I_n) are the total number of authors at any time n in the dataset derived from the pivot table (as described in *3.4.1*). The measured values of *removed* category (R_{n+1}-R_n) are calculated based on our assumption that posts lose infectivity with time, so $R_n = I_{n-1}$. We computed the number of individuals in the *susceptible* category (S_{n+1}-S_n), where S_n is calculated by substituting the values of N, I_n, and R_n into eq (7). Therefore, we created a set of values for S, I, and R categories at every time point n.

3.4.3 *Calculating the estimated values of SIR variables using Euler's solution.* We used the Euler's solutions for *SIR* ordinary differential equations represented in equations (4)-(6) and the values of S_n, I_n, and R_n to calculate the estimated values of the categories. While solving the Euler's equations we used a guess value for the two rate parameters, infection rate (α) and removal rate (β). These values determine the rate at which individuals enter and exit each category at every time point of the dataset which is directly proportional to the participation behavior of these individuals. We selected a guess value for the parameters such that the categories $SIR_{measured}$ had a closer match to $SIR_{estimated}$. Estimating the appropriate parameter values was not possible through random guess work, so, we utilized a computational algorithm to perform the parameter estimation through minimizing the error between the $SIR_{measured}$ (S_{n+1}-S_n, I_{n+1}-I_n, R_{n+1}-R_n) and $SIR_{estimated}$ values ($-\alpha S_n I_n \Delta t$, $[\alpha S_n I_n - \beta I_n]\Delta t$, $\beta I_n \Delta t$) of the SIR category models.

3.4.4 *Parameter estimation, best fit optimization and basic reproductive ratio.* The optimum values of the parameters α and β were estimated using the non-linear least squares optimization technique. The sum of squared error (SSE) between the measured ($SIR_{measured}$) and estimated values ($SIR_{estimated}$) of SIR categories was minimized. The rate parameters at the minimum SSE value are the optimum values that best represent the dynamics in the dataset. Infection rate (α) determines the rate at which individuals are meeting the individuals who hold an idea and participating in the twitter conversations to further propagate that idea. Recovery rate (β) is the rate at which the individuals are removed after participation in idea propagation. The basic reproductive ratio (R_0) which is the ratio of infection rate to recovery rate, provides information whether the population will experience an epidemic. If the R_0 >1, then one infective can transmit the idea to multiple susceptibles and the population will experience an epidemic. On the contrary, if R_0 < 1, then the idea will eventually die out since one infective is not able to transmit the idea to at least one susceptible. The R_0 values of various populations can be compared to understand which one had a higher infection rate and greater participation among the infected individuals. We calculated the best fit estimates, mean squared error (MSE) and R-squared (R^2) values for the curve fitting of $I_{estimated}$ values to the $I_{measured}$ values of all the four datasets in Table 2. These best fit estimates show the power of optimization algorithms in minimizing the error between a measured and estimated value with respect to a dataset.

4. Results

Figures 1(a)-1(d) represent the number of original authors (I_n, black dots) and the estimated authors ($-\alpha S_n I_n \Delta t$, grey curve fit line) as a function of time for LAK 2017, LAK 2019, ILA 2017 and TXEDUCHAT respectively. Optimized rate parameters and their best fit estimate values are presented in Table 2.

Figures 1(a) and 1(b) show similar curve fitting lines for LAK conferences with two peaks on day 2 and day 4 of the conferences in both the years and share similar best fit estimate values (refer to Table 2). This shows that the population of authors in 2017 show strikingly similar social interactions or engagement patterns to the authors in 2019. One of the reasons for such similar patterns in both the iterations of the LAK conferences could be that most of the individuals would have been repeated in 2019 from the 2017 conference. These best fit curves yielded the optimum rate parameters α and β. From Table 2, we can observe LAK2019 had 4 times higher infection rate (α) and 3 times higher removal rate (β) than LAK2017. Though LAK2017 had higher number of initial susceptibles S (0), the initial number of infectives I (0) were similar in both the years.

The LAK conferences had reproductive ratio R0 < 1 (refer to Table 2), which means the tweet was no longer epidemic in nature at the end of its lifetime. When compared to the R0 values of other datasets, LAK conferences had significantly larger R_0, which shows that there was higher infection rate or social participation and therefore greater engagement in the population with respect to the discussion over the idea of LAK conference. Within the LAK conferences, 2017 conference had a R_0 value smaller than 2019, indicating a possibility that if authors continued to tweet about LAK, they could infect a few more future authors before reaching saturation. Figures 2 (a) and 2(b) represent the standard SIR epidemic curves for the LAK conferences. These curves represent the estimated values of each category of $SIR_{estimated}$ based on the optimum rate parameters for the complete lifetime of the dataset. These graphs provided information about the S and R categories representing the future authors and removed authors respectively. In LAK2019 by day 4, most of the population was distributed in I and R categories with the S category reaching saturation. The R category had greater number of individuals than I. This behavior indicates that the number of future authors had reduced significantly, and majority of the individuals were removed authors whose tweets lost the influence and the remaining were current authors who actively posted tweets using the LAK hashtag.

Whereas in 2017, the S curve did not reach saturation and had individuals who were not yet active in posting tweets, leaving the possibility for more social participation for a few more days after the end of conference. These results show that LAK2019 conference had higher amount of social participation and therefore higher engagement in the population in less than a week's time.

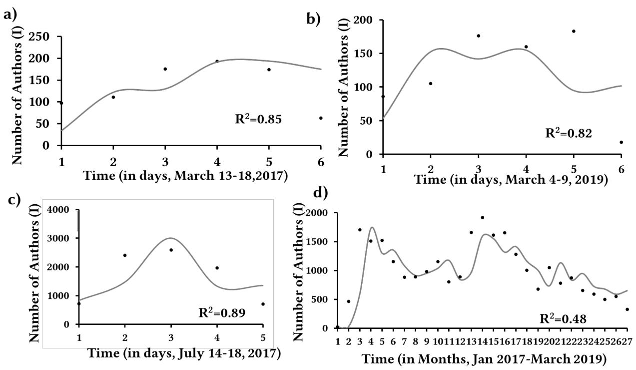

Figure1: Observed authors (black dots) as a function of time with the best fit curve (grey line) of *dI/dt* for the datasets a) LAK 2017, b) LAK 2019, c) ILA 2017 and d) TXEDUCHAT.

ILA 2017 conference had significantly larger population of current authors (*I*) and future authors (*S*) than LAK conferences due to the larger number of attendees. The original authors and estimated authors curve fit graph for ILA conference is shown in Figure 1(c). The best fit line has its highest peak on day 3 compared to any other day. The rate parameters and best fit estimates are presented in Table 2. The α value of ILA 2017 conference is one order of magnitude smaller than both the LAK conferences. The β value was higher than LAK 2019, which might be the reason that the conference reached its highest number of infectives slowly by day 3 and immediately had a diminishing curve (Figure 1(c)) when compared to the two peaks in LAK conferences (Figures 1(a) and 1(b)). Though the population of individuals was higher in each category, the rate at which the individuals in the *I* category were propagating the idea to the *S* category was not sufficient to utilize the advantage of having a larger population. In Figure 2(c), the I curve is reaching a peak slowly after 2 days of the conference, confirming the lower infection rates. There is an increase in the removal curve too due to our assumption that infectives are removed every day in a conference. While the S curve is diminishing rapidly, it reaches saturation on day 3, where the *I* and *R* categories are overlapping which confirms with the higher rates of removal in population. The R_0 was one order of magnitude lower than the LAK conferences, which indicates that the epidemic has passed and there are no future authors of the idea.

Taken together, all the three education-related conferences followed a similar temporal dynamic, showing active and increasing engagement patterns among the individuals after day 1 and subsequently reducing towards the end of the conference. Therefore, individuals were interacting on the Twitter spaces of conferences only during the most active period of the conference, and completely losing interest towards the end. Finally, the TXEDUCHAT chat space demonstrated completely distinct patterns compared to the conferences even though there were significantly higher number of infectives over a period of 27 months.

Table 2: Mean Square Error (MSE), R-squared (R2) value, Initial infectives I (0) and initial susceptibles S (0), rate of infection (α), rate of recovery or removal ( ), reproductive ratio number (R0) of all the twitter datasets.

Twitter hashtag	MSE	R^2	*I (0)*	*S (0)*	α	β	R_0
LAK 2017	7.11E +02	0.85	25	514	1E-03	0.17	6E-03
LAK 2019	7.15E +02	0.82	24	369	5E-03	0.71	7E-03
ILA 2017	6.93E +04	0.89	356	5714	3E-04	0.91	3E-04
TXEDU CHAT	1.15E +05	0.48	17	15823	2E-04	1.98	1E-04

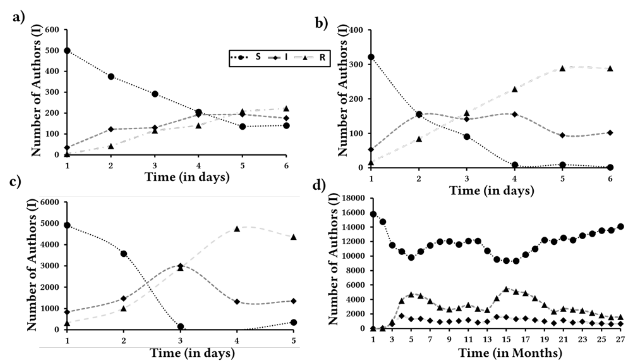

Figure 2: SIR curves as a function of time the datasets. a) LAK 2017, b) LAK 2019, c) ILA 2017 and d) TXEDUCHAT. SIR representation for every timepoint is given as: S curve (circles), I curve (diamonds), R curve (triangles).

Figure 1(d) shows the best fit line for the estimated authors in comparison to the original authors in the dataset. This dataset had the lowest infection rate α and highest removal rate β among all the Twitter hashtags (refer to Table 2) which did not allow the infectives to transmit the idea to significant number of susceptibles. The current authors' tweets lost their infectivity sooner than it could make an impact on the future authors. The total population of TXEDUCHAT was approximately 30 times greater than LAK and 3 times higher than ILA. Though the chat space had highest number of authors and longest duration of interaction time, no significant social interaction was observed. We confirmed this inference about TXEDUCHAT dataset through the SIR estimated curves in Figure 2(d) where there is no interaction between the categories. The *I* curve is almost saturated, with almost no increase in the infective population. The *R* curve has a higher amplitude than *I*, which is in sync with the very high β value. The susceptible population has the highest amplitude at any point of time, which indicates their lowest amount of interaction with the infective group. Additionally, we performed a test to divide the TXEDUCHAT dataset into three equal portions and repeat the error minimization and parameter estimation procedures on these individual portions to see if there was any period of successful infectives spread. This test confirmed that the engagement among the individuals to create new infectives was significantly lower than the total susceptible population at any given time. These long-term online chat spaces do not attract the amount of engagement when compared to a short lifetime event such as a conference.

5. General Discussion and Conclusions

In this paper, we propose to model the spread of ideas in socio-technical spaces using epidemiological methods. A premise of this work is that social encounters and connections play essential roles in many learning contexts; learning is a function of social participation and information diffusion. Therefore, instead of focusing on individual knowledge acquisition, we argue for attending to the social process of idea expression, spread, and evolution. To this end, the Lifespan of an Idea framework is proposed.

To test the plausibility of this framework, we investigated social participation---an enabler of idea spread---in four different Twitter spaces. We modeled their tweet posting patterns through the lens of *SIR* epidemiological modeling. We divided the population into different categories and measured the rate parameters for all the four datasets through minimization of the error between measured and estimated values of *SIR* categories and represented their trends as standard *SIR* curves. These curves represent the engagement patterns of individuals in each category. We discussed the differences among these datasets and made inferences on how their total lifetime affects the overall interest in participation among the individuals. Our results show that long term chat spaces need to have greater infection rates than recovery rates, to sustain the participation in the group. Though the chat space has new discussion questions every week, the chat organizers need to focus on discussing topics that might gather interest from larger population of the group and naturally improve the participation rates. Meanwhile, conferences could improve the participation rates of individuals on the first and last day through dedicated online virtual discussion or presentation sessions on Twitter. Finally, the rate of infection can be improved to increase the social participation by deliberately attracting new susceptibles using the network connections and influence of highly infective individuals.

A further investigation into the historical patterns of all the 9 years of LAK conferences along with the underlying network structure can give an insight into how to improve these engagement patterns by infecting new individuals. This is possible by understanding the topics of interest in the field and attracting new researchers to present their work in the next iterations of the conference.

Findings of the comparative case studies show early promise of the *Lifespan of an Idea* framework. Future work has been planned to further advance this framework. We will use the networked epidemiological models [13, 15]to analyze the origin, spread and saturation of ideas with the infectives or authors as nodes and susceptible or future authors as connections. Finally, to move from modeling participation to idea spread, we plan to integrate text mining and natural language processing techniques to represent ideas and topics in these conversations.

ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under Award Number 1546393.

REFERENCES

[1] Vygotsky, L. S. *Mind in society: The development of higher psychological processes*. Harvard university press, 1980.
[2] Lave, J. and Wenger, E. Situated learning: Legitimate peripheral participation. Cambridge university press, 1991.
[3] Scardamalia, M. and Bereiter, C. Knowledge building: Theory, pedagogy, and technology Cambridge University Press., 2006.
[4] Downes, S. Recent Work in Connectivism. European Journal of Open, Distance and E-learning, 22, 2 (2019).
[5] Siemens, G. Connectivism: A Learning Theory for the Digital Age 2(1) (2005).
[6] Halatchliyski, I., Hecking, T., Goehnert, T. and Hoppe, H. U. Analyzing the main paths of knowledge evolution and contributor roles in an open learning community. Journal of Learning Analytics, 1, 2 (2014), 72-93.
[7] Goffman, W. and Newill, V. A. Generalization of epidemic theory: An application to the transmission of ideas. Nature, 204, 4955 (1964), 225-228.
[8] Goffman, W. Mathematical approach to the spread of scientific ideas—the history of mast cell research. Nature, 212, 5061 (1966), 449-452.
[9] Kermack, W. O. and McKendrick, A. G. A contribution to the mathematical theory of epidemics. Proceedings of the royal society of london. Series A, Containing papers of a mathematical and physical character, 115, 772 (1927), 700-721.
[10] Milner, F. and Pugliese, A. Periodic solutions: a robust numerical method for an SIR model of epidemics. Journal of mathematical biology, 39, 6 (1999), 471-492.
[11] Hossain, M. T., Miah, M. M. and Hossain, M. B. Numerical Study of Kermack-Mckendrik SIR Model to Predict the Outbreak of Ebola Virus Diseases Using Euler and Fourth Order Runge-Kutta Methods. American Scientific Research Journal for Engineering, Technology, and Sciences (ASRJETS), 37, 1 (2017), 1-21.
[12] Bettencourt, L. M., Cintrón-Arias, A., Kaiser, D. I. and Castillo-Chávez, C. The power of a good idea: Quantitative modeling of the spread of ideas from epidemiological models. Physica A: Statistical Mechanics and its Applications, 364 (2006), 513-536.
[13] Kiss, I. Z., Broom, M., Craze, P. G. and Rafols, I. Can epidemic models describe the diffusion of topics across disciplines? Journal of Informetrics, 4, 1 (2010), 74-82.
[14] Chen, B., Chen, X. and Xing, W. Twitter archeology of learning analytics and knowledge conferences. ACM, City, 2015.
[15] Chavalarias, D. and Cointet, J.-P. Phylomemetic patterns in science evolution—the rise and fall of scientific fields. *PloS one*, 8, 2 (2013), e54847.