Received: 15 October 2018 | Revised: 4 February 2019 | Accepted: 9 April 2019
DOI: 10.1002/cpe.5321

WILEY

SPECIAL ISSUE PAPER

Dynamic social network analysis: A novel approach using agent-based model, author-topic model, and pretopology

Thi Kim Thoa Ho[1,3] | Quang Vu Bui[2] | Marc Bui[3]

[1]Hue University of Education, Hue University, Hue City, Vietnam
[2]Hue University of Sciences, Hue University, Hue City, Vietnam
[3]CHArt Laboratory EA 4004, EPHE, PSL Research University, Paris, France

Correspondence
Thi Kim Thoa Ho, Hue University of Education, Hue University, Hue City 530000, Vietnam.
Email: thi-kim-thoa.ho@etu.ephe.psl.eu

Quang Vu Bui, Hue University of Sciences, Hue University, Hue City 530000, Vietnam.
Email: bqvu288@gmail.com

Summary

We propose in this work a novel approach for dynamic social network analysis by combining an agent-based model, an author-topic model, and pretopology. We first introduce an analytical model for a dynamic social network associated with textual content using agent-based and author-topic models, namely, *Textual-ABM*. The purpose of Textual-ABM is to support for the concept exploitation of the *"dynamics"* of a social network, which contains not only network's structure transformation but also agent's interest variation over time. Agent's interest is revealed through topic probability distribution, which is estimated based on textual data using an author-topic model. In addition to demonstrating the fluctuation of the social network related to textual content, we also exploit information propagation phenomena by proposing two expanded spreading models. The first model is an expanded model of an independent cascade model in which probability of infection is formed on homophily, namely *H-IC*. We have implemented experiments on a collected dataset from the Neural Information Processing Systems Conference and have acquired satisfying results. Furthermore, we propose an extended model of pretopological cascade model from our previous work, namely, *Textual-PCM*. The advantage of *PCM* comparison with classical cascade model is to utilize pseudoclosure function built from pretopology to define the more complex set of neighborhoods. In this work, we expand *PCM* to apply detail for a social network related to textual information. A toy example with some experiments and discussion is illustrated for *Textual-PCM*. The work in this paper is an extended version of our paper dynamic social network analysis using author-topic model presented in I4CS 2018 Conference.

KEYWORDS

agent-based model, author-topic model, dynamic network, independent cascade model, information diffusion, pretopology, social network

1 | INTRODUCTION

Research on social networks is increasingly attracting consideration from scientists with the occurrence of scientific fields such as social network analysis (SNA),[1,2] community detection,[3,4] information diffusion,[5-7] and so on. Nevertheless, there are always variation states for social networks that are challenging to describe and model. Consequently, the analytical propensity is transferring from researching on static social networks to dynamic social networks. Recently, there are two principal approaches in exploiting the dynamic concept of social networks: structure transformation and the attribute fluctuation of nodes over time. The first approach is that a dynamic network is taken into account as a cumulation of snapshot networks[8] or concept-temporal networks.[9] On the other hand, the dynamics of a social network is discovered in the aspect that the characteristic of a node may change over time since each node is considered a living entity with capabilities such as communicating, learning, and adapting to the environment, unlike a static node in SNA. Agent-based modeling (ABM) is often utilized to reveal the evolution of the network. In this study, we will combine these two approaches to analyze a dynamic social network.

Most studies related on the two approaches mentioned have not exploited textual information in the interaction between users since they are mostly concentrated on topology[8,9] or only propose an ABM for illustrating dynamic social networks without content.[10] Nevertheless, textual

https://doi.org/10.1002/cpe.5321

information contains significant content; for instance, we can define whether authors research in the same narrow subject or not based on the content of their papers, or we can determine the common interests between two users on Twitter based on their tweets. Therefore, in this work, in addition to discovering fluctuation on network topology, we also exploit the transformation of user's interest based on textual information from interactions. We illustrate a user's interest under a topic probability distribution by utilizing one of the text mining technologies, including latent Dirichlet allocation (LDA)[11] and author-topic modeling (ATM).[12] LDA is a generative statistical model of a corpus in which each document is taken into account as a combination of multiple topics, and each topic is demonstrated by a probability distribution of words. In addition, the ATM is a generative model for documents and expands LDA to incorporate author's information in which each author is associated with a mixture of topics where the topics are multinomial distributions over words. The issue is to how to update the variation of the topic distribution of users after a time period to demonstrate user's interest transformation. In LDA, to measure user topic distribution, we consider that each user correlates with each document. Therefore, we cannot utilize LDA's update mechanism to update the topic distribution of users when users have more documents in interaction. Therefore, we choose the ATM to estimate user topic distribution and simultaneously update the mechanism to update user topic distribution since each author can own various documents.

In this work, we propose an analytical model for a dynamic social network associated with textual information, namely, *Textual-ABM*. *ABM* and *ATM* are major tools which are utilized to construct *Textual-ABM*. The dynamics of a social network is illustrated under the transformation of *Textual-ABM* including agent's network structure and agent's interest. Moreover, we construct online visualization application that corresponds to *Textual-ABM* to support for the user observe dynamic social networks related to textual information.

In addition to expressing social network dynamically, we exploit the information diffusion phenomena by proposing two expanded propagation models. The first model is an expanded model of an independent cascade (IC) model in which the probability of infection is formed on homophily, namely *H-IC*. The IC model is a propagation model in which infected probability is associated with each edge. The probability $P_{(u,v)}$ is the probability that u infects v. This probability is usually allocated by a uniform distribution.[13-15] However, in reality, propagation is perhaps largely dependent on similarity or homophily about interest; for instance, if two professors are both interested in a certain topic, then the probability to incorporate to research, discuss, and write a common paper is higher compared with the case of having different interests. Therefore, we propose the *H-IC* model in which infected probability is based on homophily. To measure homophily about the interest between two nodes, we based on the topic probability distribution of each node. Several experiments are conducted on static and dynamic coauthor networks. The results illustrated that the effectiveness of *H-IC* on the static network outperforms the IC with random infected probability. In addition, experimental results also revealed transformation about active proportion when the propagation process takes place on a dynamic network instead of obtaining and conserving steady state on a static network.

On the other hand, we propose an expanded model of the pretopological cascade model (PCM) from our previous research,[16] namely, *Textual-PCM*. Majority information diffusion models are performed through node's neighborhoods. Therefore, the first step of the spreading process is to define the neighborhood set of an active node set. The usual method is to aggregate neighbors of all members from the active node set. The problem is how to determine more complicated neighborhood sets. In our previous research,[16] we proposed PCM in which we can find out a more complex set of neighborhoods based on pretopology theory. We can define more complex neighborhoods set using the pseudoclosure function in pretopology since the classical approach is only a particular case in ways of determining neighborhoods from the pseudoclosure function. Therefore, in this work, we propose an expanded model of PCM applied for a social network related to textual information, namely *Textual-PCM*. *Textual-PCM* takes place on a multirelational network, where the set of neighborhoods is determined by a strong pseudoclosure function with different strong levels that depend on the number of relations. In addition, we can apply dissimilar diffusion rules, including a probability rule and a threshold rule based on homophily. Additionally, a toy example is demonstrated for *Textual-PCM*.

The structure of this paper is organized as follows. Section 2 reviews the background of the study. The model *Textual-ABM* and the toy example are proposed in Section 3. The online visualization application of *Textual-ABM* is presented in Section 4. Section 5 reveals the *H-IC* model, experiments, and results. The *Textual-PCM* and toy example are demonstrated in Section 6, and we conclude the study in Section 7.

2 | PRELIMINARIES

2.1 | Agent-based model

ABM is a class of computational models for simulating the actions and interactions of autonomous agents. There are numerous fields using ABM, such as biology, ecology, and social science.[17] ABM contains three principal components, including agents, an interactive environment, and an interactive mechanism between agents. An agent is a diverse entity that consists of various dissimilar properties and behaviors. Agents exchange information through interaction, which leads to the transformation of perception, characteristic, and behavior.

2.2 | Topic modeling

2.2.1 | Latent Dirichlet allocation (LDA)

LDA[11] is a generative statistical model of a corpus. In LDA, each document is illustrated as a combination of dissimilar topics, and each topic is a probability distribution over a vocabulary collection. The LDA's generative model is presented with the probabilistic graphical model in Figure 1A.

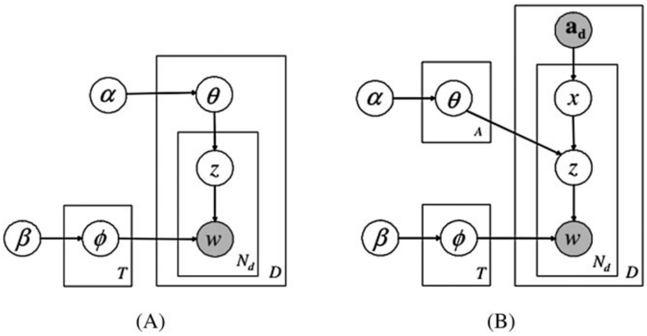

FIGURE 1 Text mining methods: LDA and ATM

Nevertheless, LDA has not exploited the author's information since it only takes into account each document as a composition of multiple topics. Therefore, ATM has been proposed as an extended model of LDA.

2.2.2 | Author-topic model (ATM)

ATM[12] is a generative model for the corpus that expands LDA to comprise the author's information. Each author is associated with the cooperation of multiple topics where each topic is a multinomial distribution over words. The words in a cooperative document are considered a result of the integration of the authors' topics. The generative model of ATM is illustrated with a graphical model in Figure 1B and proceeds as follows.

1. For each author $a = 1, \ldots, A$, choose $\theta_a \sim \mathrm{Dirichlet}(\alpha)$.
   For each topic $t = 1, \ldots, T$, choose $\phi_t \sim \mathrm{Dirichlet}(\beta)$.
2. For each document $d = 1, \ldots, D$

   2.1. Given the vector of authors a_d

   2.2. For each word $\mathrm{i} = 1, \ldots, N_d$

      2.2.1. Choose an author $x_{\mathrm{DI}} \sim \mathrm{Uniform}(a_d)$

      2.2.2. Choose a topic $z_{\mathrm{DI}} \sim \mathrm{Discrete}(\theta_{x_{di}})$

      2.2.3. Choose a word $w_{\mathrm{DI}} \sim \mathrm{Discrete}(\phi_{z_{di}})$

2.2.3 | Update mechanism of author-topic model

We can estimate word's distribution on each topic and topic distribution on each author from a training set using ATM. In addition, ATM can be updated with supplemental documents after training has been finished. This update procedure is executed by expectation maximization, iterating over new corpus until the topics converge. The two models are then combined in proportion to the number of old and new documents. As an alternative, for stationary input, this process is equivalent to the online training of Hoffman et al.[18] Recently, there are several application programming interfaces (APIs) that are available for topic modeling, including *topicmodels* or *lda* in R, or *Gensim*[*] in Python. In this work, we chose Gensim to implement ATM.

2.3 | Pretopology theory

For dynamic SNA, scientists try to build the models that can capture the transformation process step by step. For instance, it is the case when one studies the spreading of information on complex networks. To deal with such process, usual topology does not seem very adequate since a closure function is an idempotent one in topology. Therefore, to follow intermediate steps between a set A and $F(A)$, ie, closure of A, we have to relax this assumption of idempotence. This is exactly what *pretopology* proposes.

Pretopology[19] is built based on a *pseudoclosure* function. Different to a *closure operator* with having only step from a set A to its closure, ie, $F(A)$, the *pseudoclosure* function, with its flexibility, can follow step by step the growth of a set; then it can be applied to solve some problems in complex systems such as data analysis,[19] clustering with multicriteria,[20] social network modeling for group,[21] topological analysis of complex systems,[16] etc.

[*] https://pypi.python.org/pypi/gensim

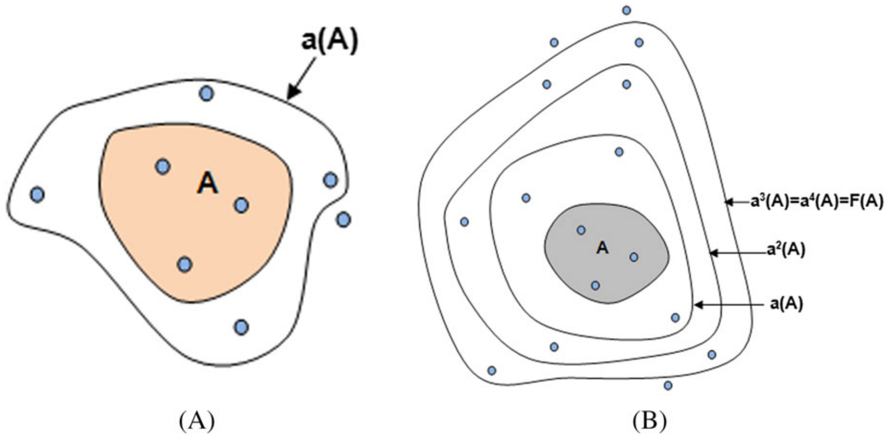

FIGURE 2 Pseudoclosure and closure function

2.3.1 | Pretopological notions

Definition 1. A pretopological space is an ordered pair (X, a), where X is a set and $\mathrm{a} : \mathcal{P}(X) \rightarrow \mathcal{P}(X)$ is a **pseudoclosure** operator, satisfying the two following axioms:

(P1): $a(\emptyset) = \emptyset$; (preservation of nullary union)
(P2): $A \subset a(A) \forall A, A \subset X$ (extensivity)

It is important to note that, by defining *pseudoclosure* $a(.)$ (see Figure 2A), we do not suppose that it is an idempotent transform. Then, conversely as it happens in topology, we can compute: $a(A), a(a(A)), a(a(a(A))), \ldots, a^k(A)$ (see Figure 2B). As a consequence, the *pseudoclosure* function can follow step by step the growth of a set by adding elements to a departure set according to the defined characteristics.

Definition 2. Given a pretopological space (X, a), $A \subset X$ is a closed subset if and only if $a(A) = A$.

Definition 3. Given a pretopological space (X, a), call the closure of A, when it exists, the smallest closed subset of X that contains A. The closure of A is denoted by $F(A)$.

Closure gives the information about the reachability of a set. Therefore, from the point of view of application, the existence of closure is very important since it ensures that the algorithm can stop after finite steps. In the most general case, closure does not necessarily exist. As a consequence, we need to build some pretopological spaces that are less general than the basic one but for which the neighbors satisfy some good properties and the closure always exists.

2.3.2 | Pretopological spaces

Definition 4. A pretopology space (X, a) is called $\mathcal{V}$-type space if and only if

$$(P3) \quad (A \subseteq B) \Rightarrow (a(A) \subseteq a(B)) \quad \forall A, B \in \mathcal{P}(X) \quad (\text{Isotonic}). \tag{1}$$

Definition 5. A pretopology space (X, a) is called $\mathcal{V}_D$-type space if and only if

$$(P4) \quad a(A \cup B) = a(A) \cup a(B) \quad \forall A \subset X, \quad \forall B \subset X \quad (\text{Additive}). \tag{2}$$

Definition 6. A pretopology space (X, a) is called $\mathcal{V}_S$-type space if and only if

$$(P5) \quad a(A) = \bigcup_{x \in A} a(\{x\}) \quad \forall A \subset E. \tag{3}$$

Proposition 1. *Any $\mathcal{V}$-type space is a $\mathcal{V}_S$-type space.*
Any $\mathcal{V}_S$-type space is a $\mathcal{V}_D$-type space.

2.3.3 | Some ways to build pseudoclosure function

In our previous work,[16] we proposed some ways to build pseudoclosure functions in many situations for applications.

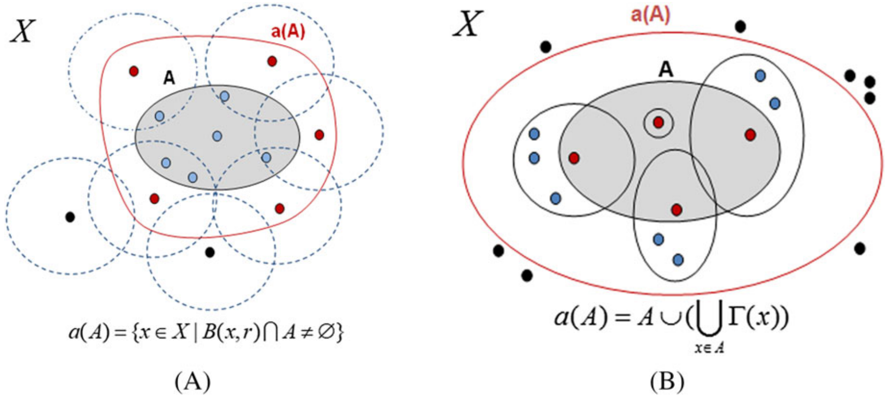
X
a(A)
A
$a(A)=\{x\in X \mid B(x,r)\bigcap A\neq\varnothing\}$
(A)
X
a(A)
A
$a(A)=A\cup(\bigcup_{x\in A}\Gamma(x))$
(B)

FIGURE 3 Pseudoclosure in metric space and a space equipped with a neighbor function

Pretopology in metric space

Let us consider a metric space (X, d), where d is a distance. $\forall x \in X, B(x, r) = \{y \in X | d(x, y) \leq r\}$ is a ball with the center x and a radius $r > 0$. Clearly, $\mathcal{B}(x) = (B(x, r)_{x\in X})$ is the basis of neighborhoods of x. We then can build a *pseudoclosure* function **a(.)** (see Figure 3A for an example) on X with $B(x, r)$ such as

$$\forall A \in \mathcal{P}(X), \quad a(A) = \{x \in X | B(x, r) \cap A \neq \emptyset\}. \tag{4}$$

Pretopological space built previously is of $\mathcal{V}_D$-type pretopological space ($\mathcal{V}_S$-type if X is finite).

Pretopology in a space equipped with a neighborhood function

We consider neighborhood function $\Gamma : X \to \mathcal{P}(X)$. Clearly, $\mathcal{B}(x) = ((x \cup \Gamma(x))_{x\in X})$ is the basis of neighborhoods of x. We then can build a *pseudoclosure* function **a(.)** (see Figure 3B for an example) such as

$$\forall A \in \mathcal{P}(X), \quad a(A) = A \cup \left(\bigcup_{x\in A} \Gamma(x)\right). \tag{5}$$

We can see that the pretopology space built from the pseudoclosure $a(.)$ function defined previously is of $\mathcal{V}_S$ type. The graph defined by *Claude Berge sense*[22] is a special case of this kind of pretopological space.

Pretopology and binary relationships

Suppose we have a family $(R_i)_{i=1,\dots,n}$ of binary relationships on a finite set X. Let us consider $\forall i = 1, 2, \dots, n, \forall x \in X, V_i(x)$ defined by

$$V_i(x) = \{y \in X | x R_i y\} \tag{6}$$

Then, the pseudoclosure $a_s(.)$ is defined by

$$a_s(A) = \{x \in X | \forall i = 1, 2, \dots, n, V_i(x) \cap A \neq \emptyset\} \quad \forall A \subset X \tag{7}$$

Pretopology defined on X by $a_s(.)$ using the intersection operator is called the strong pretopology. Figure 4A gives an example for strong pretopology built from two relationships.

Proposition 2. *$a_s(.)$ determines on X a pretopological structure, and the space (X, a_s) is of V-type pretopological space.*

Similarly, we can define weak pretopology from $a_w(.)$ by using the union operator, ie,

$$a_w(A) = \{x \in X | \exists i = 1, 2, \dots, n, V_i(x) \cap A \neq \emptyset\} \quad \forall A \subset X. \tag{8}$$

Proposition 3. *$a_w(.)$ determines on X a pretopological structure, and the space (X, a_s) is of $\mathcal{V}_D$-type.*

Pretopology and valued relationships

Let us consider the space X in which elements are linked by a valued relation. For this space, we first define a value function ν from $X \times X \to \mathbb{R}$ as: $(x, y) \to \nu(x, y)$. Then, we can build the *pseudoclosure* **a(.)** such as

$$\forall A \in \mathcal{P}(X), \quad a(A) = \left\{ y \in X - A | \sum_{x\in A} \nu(x, y) \geq s \right\} \cup A, s \in \mathbb{R} \tag{9}$$

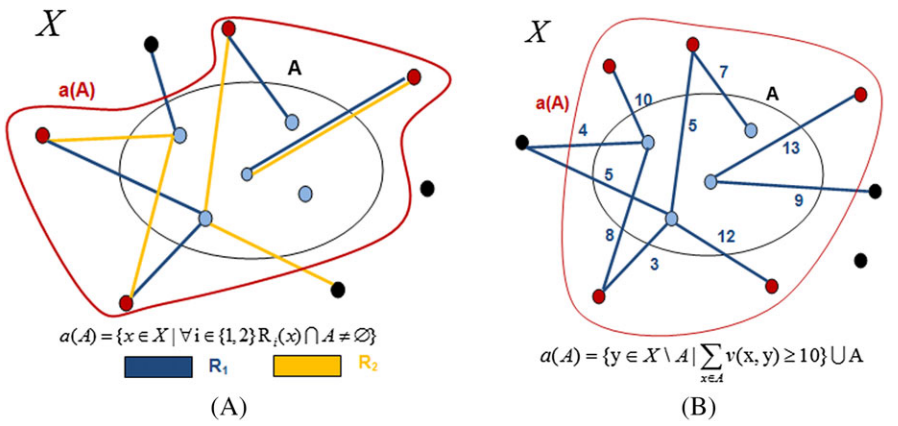
X
a(A)
A
$a(A) = \{x \in X \mid \forall i \in \{1,2\} R_i(x) \cap A \neq \emptyset\}$
R1
R2
(A)
X
a(A)
A
10
7
4
5
13
5
9
8
12
3
$a(A) = \{y \in X \setminus A \mid \sum_{x \in A} v(x,y) \geq 10\} \cup A$
(B)

FIGURE 4 Pseudoclosure in binary and valued space

The pseudoclosure $a(A)$ is a set that contains set A, and all elements $y \in X$ where the sum of valued edges between some elements of A and y is greater than the threshold s. Pretopological space built from this kind of pseudoclosure is $\mathcal{V}$-type space. Figure 4B shows an example of this pretopological space with $s = 10$.

3 | ANALYTICAL MODEL FOR DYNAMIC SOCIAL NETWORK ASSOCIATED WITH THE TEXTUAL INFORMATION USING AGENT-BASED MODEL AND AUTHOR-TOPIC MODEL (TEXTUAL-ABM)

In this section, we propose an analytical model for a dynamic social network associated with the textual information using ABM and ATM, namely, Textual-ABM (see Figure 5). Each agent will represent for each node in the social network. We illustrate the fluctuation of the social network under two aspects, in which the former is agent's network structure and the latter is agent's topic distribution. We reveal in the following the steps to construct and update the model in more details.

3.1 | Textual information extraction and topic modeling with ATM

In the first step, we will collect textual data from social networking sites or blogs using APIs, for instance, Twitter API, Facebook Graph API, and so on. Subsequently, some preprocessing techniques are conducted to clean data including stemming the words and removing stop-words and numeric symbols. Eventually, preprocessed data will be saved into training corpus. After textual data extraction and preprocessing step, we utilize ATM method to estimate user topic probability distribution from the training corpus. The output of ATM consists of two matrices: The author-topic distribution matrix θ and the topic-terms distribution matrix ϕ. The topic-term distribution matrix $\phi \in R^{KxV}$ consists of K rows, where the ith row $\phi_i \in R^V$ is the term's distribution on topic i. The author-topic distribution matrix $\theta \in R^{NxK}$ consists of N rows, where the ith row $\theta_i \in R^K$ is the topic distribution for author i. A high probability value of θ_{ij} means that author i is interested in topic j.

3.2 | Textual-ABM formation

From the results of ATM, Textual-ABM is constructed with the principal components are *agents*, *agent's network*, and *global environment surrounding the network*. First, in our model, a user on the social network will be represented by an agent who are heterogeneous with several

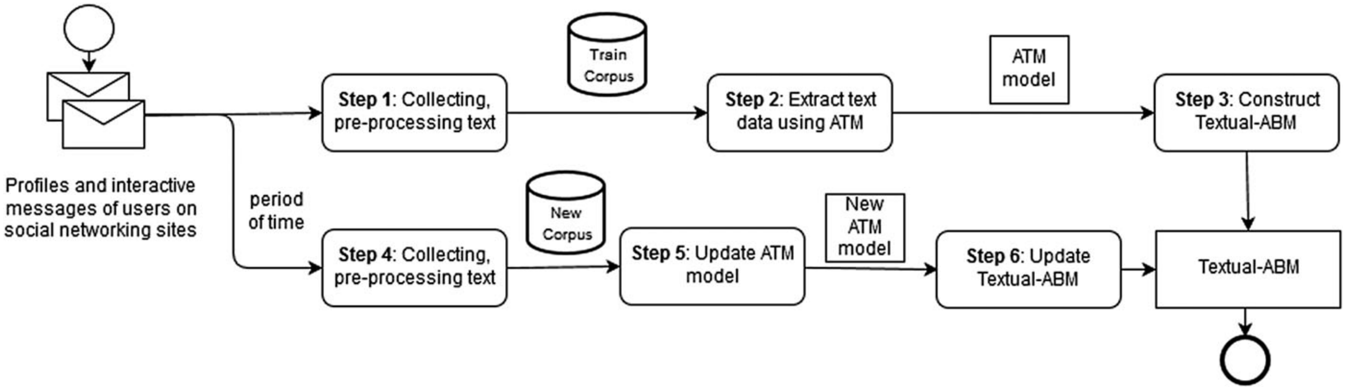
Train Corpus
ATM model
Step 1: Collecting, pre-processing text
Step 2: Extract text data using ATM
Step 3: Construct Textual-ABM
Profiles and interactive messages of users on social networking sites
period of time
New Corpus
New ATM model
Step 4: Collecting, pre-processing text
Step 5: Update ATM model
Step 6: Update Textual-ABM
Textual-ABM

FIGURE 5 Textual-ABM for analyzing a dynamic social network using ABM & ATM

particular attributes including *ID*, *Name*, *Neighborhoods*, *Corpus* (textual collection that an agent utilize in interactive process), and *TP-Dis* (agent's topic probability distribution). *Corpus* will be gathered over time through the process of interaction. Moreover, *TP-Dis* reveals the agent's interest on different topics. After agent's formation, we will construct the agent's network. There can be one or multiple relationships between two agents, for instance, directed interaction relation (R_{DI}) such as *follow*, *retweet*, or *reply* on Twitter, *like* and *share status* on Facebook, and *collaborate in a scientific article*, and major topic relation (R_{MTP}). $uR_{\mathrm{MTP}}v$ means that agent u and v are interested in a certain topic with probability larger than threshold p_0. It can be said that the network's structure transformation is the result of the interactive process of agents since there is the occurrence of new agents, the formation of new interactions, or increase in the number of interactions. Moreover, agents act as network nodes which can change their characteristics over time. Finally, we consider a space cover agent's network, namely, *global environment*. Global environment is a space that comprise *agents*, *interaction between agents*, and *system's resource*. In particular, we focus attention on the system's resource associated with textual data. System's resource consist a *system's corpus* and a *its generative model* in which the former is combined from all agent's corpus while *ATM* is utilized for the later.

3.3 | Update mechanism for textual-ABM

After constructing model *textual-ABM* to illustrate a social network associated with textual content, we propose an update mechanism with the purpose of revealing the transformation of a social network over time. After a period of time, the crawl process of new textual data and preprocessing are automatically conducted. Next, the *system's corpus* and the existed generative model *ATM* will be updated with the new corpus. Finally, the *Textual-ABM* will be updated with the occurrence of new agents and agent's attributes variation comprising *Neighborhoods*, *Corpus*, and *TP-Dis*. We illustrate steps from 1 to 3 for textual-ABM's formation and update mechanism equivalent to steps from 4 to 6 in Figure 5.

Since each social network has a specific characteristic, for instance, for Twitter network, the frequency that a user posts or shares tweets is usually daily, while a scientist can take one or more years to publish a paper. Therefore, textual information collection from different social networks is dissimilar, including APIs for scraping and the timescale for crawling and updating textual content. Depending on the network's feature, we set up an appropriate timescale for crawling textual information and updating network.

3.4 | Toy example

In this subsection, we simulate a dynamic network with ten users of *theguardian.com*[†] under analytical model *Textual-ABM*. Users are identified with IDs from 0 to 9. To illustrate the transformation of user's interest on the small number of topics, we crawl textual data from a political blog *"Resentful Americans turn a blind eye to Trump's faults"*[‡] in *theguardian.com*. Since this political blog attracted discussion from users in a short time period from 4:00 PM 25 April 2016 to 12:00 AM 26 April 2016, we first crawl textual content from 4:00 PM to 6:00 PM for training corpus, and the next two periods for crawling and updating are set up every hour. *Textual-ABM* is formed as soon as user topic distribution are estimated from training corpus using ATM. In *Textual-ABM*, agent's network is considered with two kinds of relations, including R_{DI} and R_{MTP} with $p_0 = 0.1$. Additionally, to demonstrate the fluctuation of the network, we performed update processes with new corpus in the next two periods in which the second period lasted from 6:00 PM to 7:00 PM and the third period lasted from 7:00 PM to 8:00 PM.

We can see the fluctuation of the network through the significant transformation about agent's network topology (see Figure 6) over three periods. There is notable transformation from the first stage to the second stage: occurrence of new agent *9*; new interactive formation for instances (0, 9) and (4, 9); increase in the number of interactions, such as (4, 7), (5, 8), or (0, 1); and especially, appearance $0R_{\mathrm{MTP}}1$ with topic [2]. Moreover, there is the remarkable appearance about relation R_{MTP} between the second stage and the third stage, for instance, (4, 7) are interested in topic [2]. On the other hand, the fluctuation of the network is also revealed under agent's interest variation. We demonstrated topic distributions θ of four typical agents over three stages in Table 1. We can see that, at the beginning, agent 0 concerned in three topics 2, 3, and 1. Nevertheless, there is a notable fluctuation about probability distribution on these three topics from the first stage to the second stage, keeping this state to the third stage. In opposition, agent 4 maintains interests in the first two stages and transforms drastically in the third stage. Perhaps, these transformations are results of the interactive process since agent 0 mostly interacts with each other in the second stage while agent 4 in the third stage. In short, we utilized *Textual-ABM* to analyze and simulate a dynamic social network where users communicate with each other under textual content. The fluctuation is not only demonstrated by the network's topology aspect but also by agent's interest.

4 | ONLINE VISUALIZATION APPLICATION OF TEXTUAL-ABM

In this section, we construct an online visualization application to support the user in the observation of dynamic social networks related to textual information. Major features of the application are that it automatically generates and updates the network's transformation after the time period. Network's fluctuation not only is illustrated by topology transformation but also topic probability distribution variation.

[†] https://www.theguardian.com/international

[‡] https://www.theguardian.com/us-news/blog/2016/aug/25/resentful-americans-turn-blind-eye-donald-trump

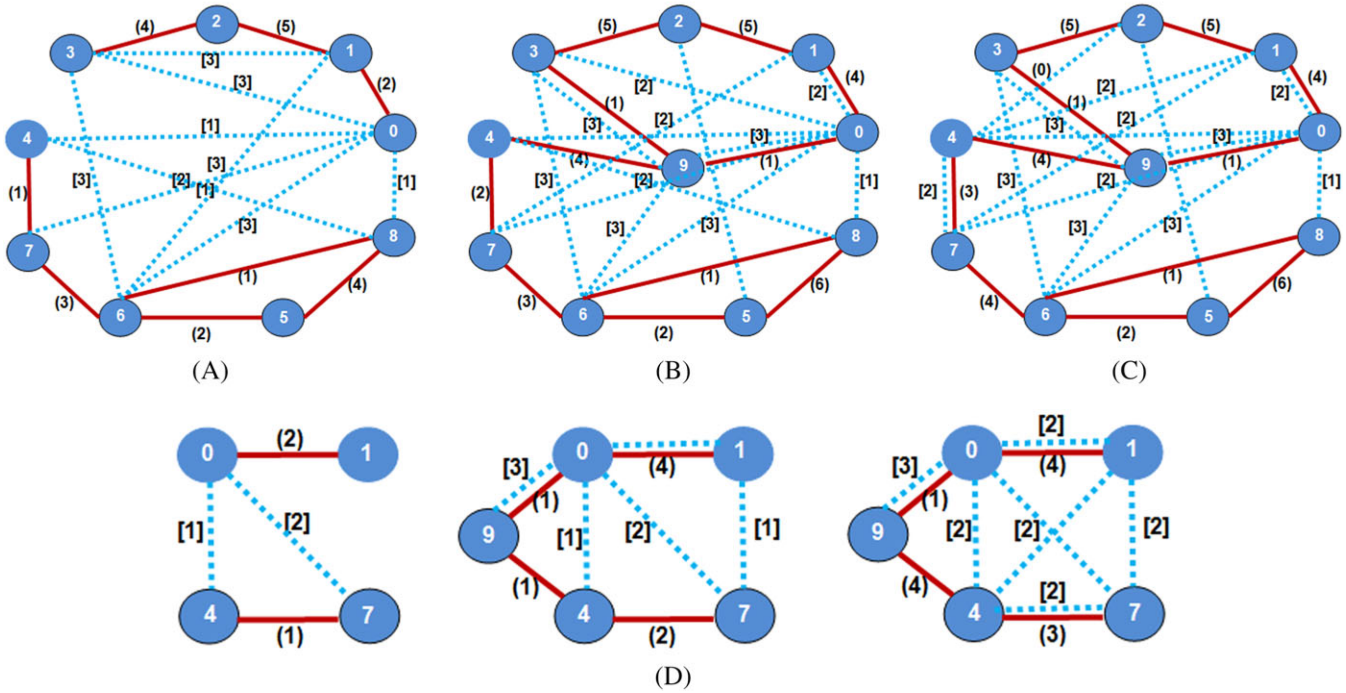

FIGURE 6 Structural fluctuation of network over three stages. The solid red line with label '*(number interaction)*' represents R_{DI}. The dash blue line with label '*[major topics]*' represents R_{MTP}. A, Period 1; B, Period 2; C, Period 3; D, Structural dynamic of several typical nodes *(0, 1, 4, 7, 9)* over three stages

TABLE 1 Topic-author distribution θ over three periods

User ID 0				User ID 1				User ID 4				User ID 7			
Topic	Prob.			Topic	Prob.			Topic	Prob.			Topic	Prob.		
	Period				Period				Period				Period		
	1	2	3		1	2	3		1	2	3		1	2	3
2	**0.542**	**0.397**	**0.397**	3	0.991	0.231	0.231	1	**0.919**	**0.919**	**0.003**	2	0.968	0.977	0.983
3	**0.277**	**0.412**	**0.412**	0	0.002	0.015	0.015	0	**0.02**	**0.02**	**0.491**	4	0.008	0.006	0.004
1	**0.175**	**0.187**	**0.187**	4	0.002	0.001	0.001	2	**0.02**	**0.02**	**0.501**	3	0.008	0.006	0.004
4	0.003	0.002	0.002	1	0.002	0.001	0.001	4	0.02	0.02	0.003	0	0.008	0.006	0.004
0	0.003	0.002	0.002	2	0.002	0.752	0.752	3	0.02	0.02	0.003	1	0.008	0.006	0.004

To illustrate facility of our application, we visualize a dynamic social network, ie, Twitter, which comprises of ten well-known users in social fields including politics, music, movie, education, and health (see Figure 7). Two main relations are considered in which the first is relation "*follow*" and the second is relation "*major topic*" (R_{MTP}). Relation R_{MTP} will appear when two agents are interested in the common topic with a probability greater than threshold p_0. We considered that each user contains a probability distribution of five topics and R_{MTP} with $p_0 = 0.3$. Initially, a network is generated as soon as text information is crawled, preprocessed, and text mined by ATM. There are numerous Python libraries for Twitter API, including Tweepy, Python Twitter Tools, Python-Twitter, Twython, and so on. Visualization demonstrates the structure of the network with two relations. Moreover, the user can click on each node to see the user topic distribution. As explained in Section 3.3, we set up after 24 hours, new tweets of users will automatically be scrapped, and the existing ATM model will be updated. Therefore, the network will also be updated including relation *"follow"*, relation "*major topic*" (R_{MTP}), and the topic distribution of users. We published all the source codes of application on GitHub.[§]

In this study, we just implement a visualization application of Textual-ABM on a small-scale network of Twitter. We can expand visualization on a more large-scale network with more nodes. However, it is difficult to visualize relations between nodes on a large-scale network because of the large number of nodes and two kinds of relation. The density of nodes and relations lead to the difficulty for users to observe the dynamic of the network. Moreover, we just suppose that the number of topics is fixed in estimating topic probability distribution, but in fact, it can change over time depending on the textual content of the users. In the future, we will try to develop and improve the features of the application, including visualization on numerous different social networks and on large-scale networks, analysis of the number of topics in each period, and more complete interface.

[§] https://github.com/CooperationProject/DynamicSocialNetworkSimulation

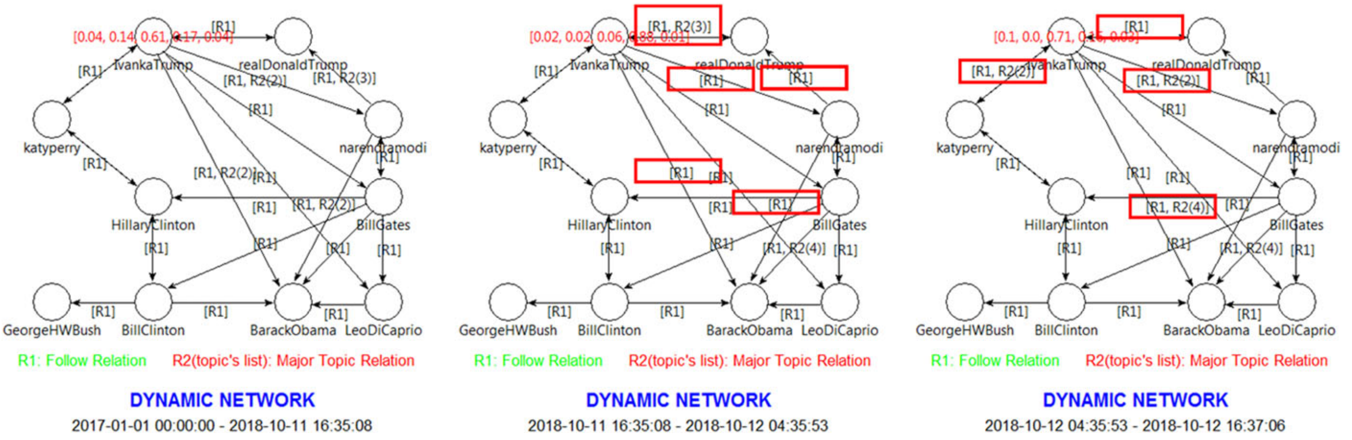

FIGURE 7 Visualization sample

5 | INDEPENDENT CASCADE MODEL BASED ON HOMOPHILY (H-IC)

Researching on information propagation phenomena is one of the principal branches of SNA. The purpose of this research is to observe the spreading of information among objects when they are connected. Recently, there are numerous models that have been proposed comprising linear threshold (LT) model, IC model, and so on.[23] In the IC model, the probability of infection is associated with each edge and usually assigned by a uniform distribution. However, perhaps, in reality, this probability mostly relies on similarity or homophily; for instance, if two scientists research in a narrow field, the probability for discussing, incorporating, and writing common articles will be higher compared with dissimilar fields. Therefore, we propose the IC model on the agent's network in which the probability of infection is based on homophily, namely *H-IC*. Homophily is estimated based on agent's topic distribution. Moreover, *H-IC* is revealed to be detailed on both the static and dynamic agent's networks.

5.1 | Independent cascade model (IC)

We assume a network $G = (V, E, P)$, where $P : V \times V \rightarrow [0, 1]$ is the probability function. $P(u, v)$ is the probability that node u infected node v. The propagation process takes place through discrete steps t. If a node adopts a new behavior or idea, it becomes active; otherwise, it is inactive. The set of active nodes at step t is considered as A_t.

Under the IC model, at each time step t, where A^{new}_{t-1} is the set of the newly activated nodes at time $t - 1$, each $u \in A^{new}_{t-1}$ infects the inactive neighbors $v \in \eta^{out}(u)$ with a probability $P_{(u,v)}$

5.2 | Homophily measure between two agents

Homophily is the propensity that people connect more with those who are similar to them on several characteristics compared with those who are not. In this work, homophily between two agents will be measured based on their topic probability distribution. If we take into account a topic probability distribution as a vector, we can select several distance measurements associated with the vector distance, for instance, Euclidean distance, cosine similarity, Jaccard coefficient, etc. Nevertheless, the results from our previous work[24] proved that it is better if we select distance measurement related to the probability distribution such as Kullback-Leibler Divergence, Jensen-Shannon divergence, Hellinger distance, etc. In this work, we choose Hellinger distance and Jensen-Shannon divergence to measure distance. Let two discrete probability distributions, $P = (p_1, p_2, \dots, p_k)$, $Q = (q_1, q_2, \dots, q_k)$

Hellinger distance:

$$d_H(P, Q) = \frac{1}{\sqrt{2}} \sqrt{\sum_{i=1}^{k} \left(\sqrt{p_i} - \sqrt{q_i} \right)^2} \tag{10}$$

Jensen-Shannon distance:

$$d_{JS}(P, Q) = \frac{1}{2} \sum_{i=1}^{k} p_i ln \frac{2p_i}{p_i + q_i} + \frac{1}{2} \sum_{i=1}^{k} q_i ln \frac{2q_i}{p_i + q_i} \tag{11}$$

Homophily:

$$\mathbf{Homo(P,Q) = 1 - d(P, Q)} \tag{12}$$

where P and Q are topic probability distribution of the two agents, respectively.

5.3 | Random independent cascade model on static agent's network

In this subsection, we illustrate the IC model in which infected probability is assigned randomly, namely *R-IC* (see Algorithm 1). This model is utilized as a benchmark model to compare effectiveness with model *H-IC* that we will propose in Section 5.4.

At each step t, where A^{new} is the set of the newly active nodes at step $t - 1$, each $u \in A^{\text{new}}$ will infect its inactive neighborhoods $v \in \eta^{\text{out}}(u)$ with a random probability $P(u, v)$. The spreading will stop when no more activation occurs.

Algorithm 1 R-IC diffusion on static agent's network

```
Require: A Textual-ABM (include agent's network G = (V, E)), $A_0$: seed set
procedure R-IC-STATIC-NETWORK($G, A_0$)
    $t = 0, A^{all} = A_0, A^{new} = A_0$
    while activation occur do
        $t = t + 1; A_t^{all} = \emptyset$
        for u $\in A^{new}$ do
            Calculate $A_t(u) = \{v \in \eta^{out}(u), \boldsymbol{p <= q}\}; \boldsymbol{p, q \sim U(0, 1)}$
            $A_t^{all} = A_t^{all} \bigcup A_t(u)$
        end for
        $A^{all} = A^{all} \bigcup A_t^{all}; A^{new} = A_t^{all}$
    end while
    return $A^{all}$                                        ▷ Output
end procedure
```

5.4 | H-IC diffusion model on static agent's network

There is a resemblance between the propagation mechanism of *H-IC* and *R-IC* on static agent's network, but the dissimilarity is that infected probability in *H-IC* based on homophily instead of a random probability (see Algorithm 2). This means at each step t where each active agent $u \in A^{\text{new}}$ infects v with a probability equal $homophily(u, v)$.

Algorithm 2 H-IC diffusion on static agent's network

```
Require: A Textual-ABM (include agent's network G = (V, E)), $A_0$: seed set
procedure H-IC-STATIC-NETWORK($G, A_0$)
    $t = 0, A^{all} = A_0, A^{new} = A_0$
    while activation occur do
        $t = t + 1; A_t^{all} = \emptyset$
        for u $\in A^{new}$ do
            Calculate $A_t(u) = \{v \in \eta^{out}(u), \boldsymbol{p <= homo(u, v)}\}; \boldsymbol{p \sim U(0, 1)}$
            $A_t^{all} = A_t^{all} \bigcup A_t(u)$
        end for
        $A^{all} = A^{all} \bigcup A_t^{all}; A^{new} = A_t^{all}$
    end while
    return $A^{all}$                                        ▷ Output
end procedure
```

5.5 | H-IC diffusion on dynamic agent's network

Despite that the IC model on the dynamic network has been discovered,[25,26] the network's dynamic concept is only exploited under the structure fluctuation while the probability of infection from this object to another is always fixed during the propagation process. Therefore, we propose the *H-IC* spreading model on a *dynamic agent's network*, which is not only comprised of the network's topology fluctuation but also the transformation of agent's topic distribution (see Algorithm 3). This means that the infected probability between two agents can change over time since their homophily can fluctuate.

The spreading mechanism of *H-IC* on the dynamic agent's network has similarity with that of the static agent's network; however, the principal dissimilarity is that, at some steps of diffusion process $k \in U$, the agent's network G will be updated.

Algorithm 3 H-IC diffusion on dynamic agent's network

```
Require: A Textual-ABM (include agent's network $G = (V, E)$); $A_0$: seed set
Require: $U = \{k_1, k_2, \dots, k_n\}$, at step $k_i$ $G$ is updated; $n$: number steps of diffusion
procedure H-IC-DYNAMIC-NETWORK(Textual-ABM, $A_0$, $U$)
    $t = 0, A^{all} = A_0, A^{new} = A_0$
    while $t < n$ do                                        ▷ $(n > max\{U\})$
        $t = t + 1; A_t^{all} = \emptyset$
        if $t \in U$ then:
            Update Textual-ABM; $A^{new} = A^{all}$
        end if
        for u $\in A^{new}$ do
            Calculate $A_t(u) = \{v \in \eta^{out}(u), \boldsymbol{p <= homo(u, v)}\}; \boldsymbol{p \sim U(0, 1)}$
            $A_t^{all} = A_t^{all} \bigcup A_t(u)$
        end for
        $A^{all} = A^{all} \bigcup A_t^{all}; A^{new} = A_t^{all}$
    end while
    return $A^{all}$                                        ▷ Output
end procedure
```

5.6 | Experiments

5.6.1 | Dataset

We implemented experiments on the coauthor network where each node is represented by a scientist and the link is relation *'coauthor'*. We constructed a coauthor network from authors who have participated in the Neural Information Processing Systems Conference (NIPS) from 2000 to 2012. The dataset consists of 1740 papers which are contribution from 2479 authors.

5.6.2 | Setup

In the first step, we collected textual data from 2000 to 2008 for training corpus and estimated topic probability distribution of authors using ATM. The number of topics is chosen based on the harmonic mean of log-likelihood (HLK).[27] First, we implemented HLK with the number of topics in the range [10, 200] with sequence 10. We realized that the best number of topics are in the range of [50, 90] of a maximum value of HLK (see Figure 8A). Therefore, we calculated one more time with the number of the topics in [50, 90] with sequence 1 and reached that the best is 67 (see Figure 8B). Based on the results of ATM, a *Textual-ABM* is formed. Notice that we only took into account the spreading on relation *'coauthor'*.

We experimented *H-IC* on four dissimilar static agent's networks equivalent to years from 2009 to 2012. The first experiment is *H-IC* on the agent's network that is established as soon as the *Textual-ABM* is updated with new corpus in 2009. Similarity, the last three experiments are *H-IC* on three agent's networks that are formed after *Textual-ABM* are updated two, three, and four times in correspondence to the new corpus from 2010 to 2012. We implemented the spreading process on the largest community *H* and measure homophily with two distances mentioned in Subsection 5.2. Seed set *A*0 is chosen from a million random samples, and the propagation process is performed a million times for each *A*0. Additionally, we also implemented *R-IC* as a baseline to compare the performance with *H-IC*.

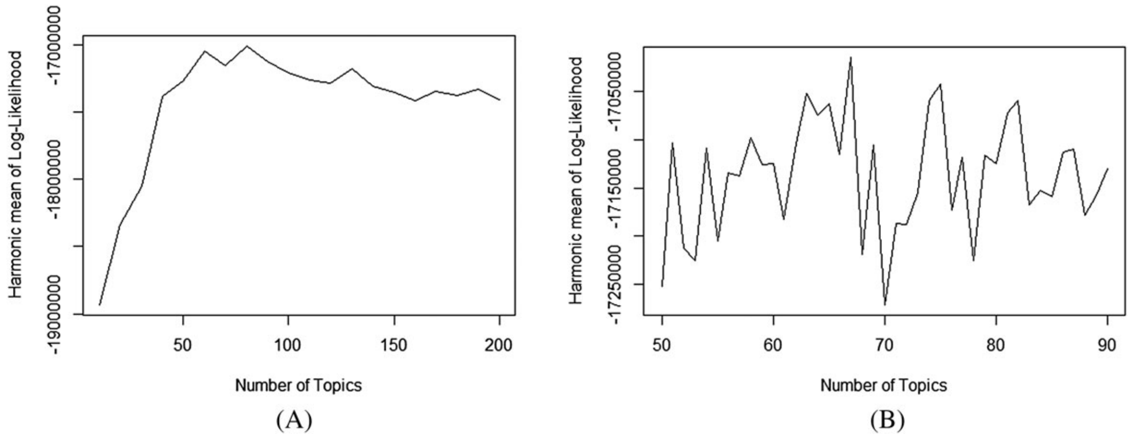

FIGURE 8 Log-likelihood. A, Number topics in range [10, 200]; B, Number topics in range [50, 90]

For investigations to simulate *H-IC* on the dynamic agent's network, the propagation process begins as soon as the *Textual-ABM* is generated. For each kind of distance measurement, we conducted four experiments in which the first is spreading on the agent's network without dynamic. The last experiments are that after every 5, 10, and 15 steps of spreading, *Textual-ABM* will be updated once.

5.6.3 | Evaluation

There are two standard metrics for evaluating the performance of spreading models: the number of active nodes and the active proportion.[25,26] In this work, we utilized the active percentage to evaluate the performance of propagation models on the static network and the number of active nodes for spreading models on a dynamic network. We compared the performance of *H-IC* with baseline model *R-IC*.

5.6.4 | Experimental results

H-IC diffusion on static agent's network

Spreading the results of *H-IC* on static agent's network are illustrated in Figure 9. Experimental results demonstrate that the active proportion of *H-IC* is always higher than *R-IC* in four dissimilar networks with two distance measurements. In 2009, the percentage of active nodes obtains around 0.08% for *R-IC* while *H-IC* reaches approximately 0.22% and 0.45% for Hellinger distance and Jensen-Shannon distance correspondingly. Similarly, the active proportion of *H-IC* diffusion with both distance measurements obtain superiority comparison with *R-IC* in the next three years. In particular, we can see that the combination between *H-IC* and Jensen-Shannon distance always bring higher performance *H-IC* and Hellinger distance. In short, we can conclude that *H-IC* outperforms in comparison with *R-IC*.

Additionally, the results illustrated that the spreading is equivalent to the network's transformation over time since active percentage for both *H-IC* and *R-IC* in the current year is always greater than the previous year. The active percentage of *R-IC* reaches about 0.08% in 2009 and increases up to 0.117%, 0.25%, and 0.4% in the next three years respectively. Moreover, the active percentage of *H-IC* in 2010 approximately is twice as large as in 2009 for both cases of distance measurements. Moreover, there is a significant increase in the active percentage of *H-IC* in the last two years compared with 2010. In 2011, *H-IC* with Hellinger distance reaches 0.622% and 0.81% for the Jensen-Shannon distance. Those percentages peak at about 0.7% and 0.86% in 2012. It can be said that the transformation about the active percentage in the spreading process results from the fluctuation of agent's network structure over time.

H-IC diffusion on dynamic agent's network

Results are shown in Figure 10, which reveal *H-IC* on the dynamic coauthor network with the combination of two distance measurements. We can see that *H-IC* without dynamic of the network quickly obtains the stable state, while there is a significant transformation in the active number if the network has fluctuation after some steps of spreading. For *H-IC* with Hellinger distance (Figure 10A), the dissemination process obtains around 227 active agents at step 8, and this state remains onwards in a static network. On the other hand, there are approximately 450 active agents for the spreading process on three dynamic networks. In addition, for *H-IC* with Jensen-Shannon distance (Figure 10B), the diffusion process obtains steady state at step 6 for the static network, while the active number reaches approximately 370 on the dynamic network. In

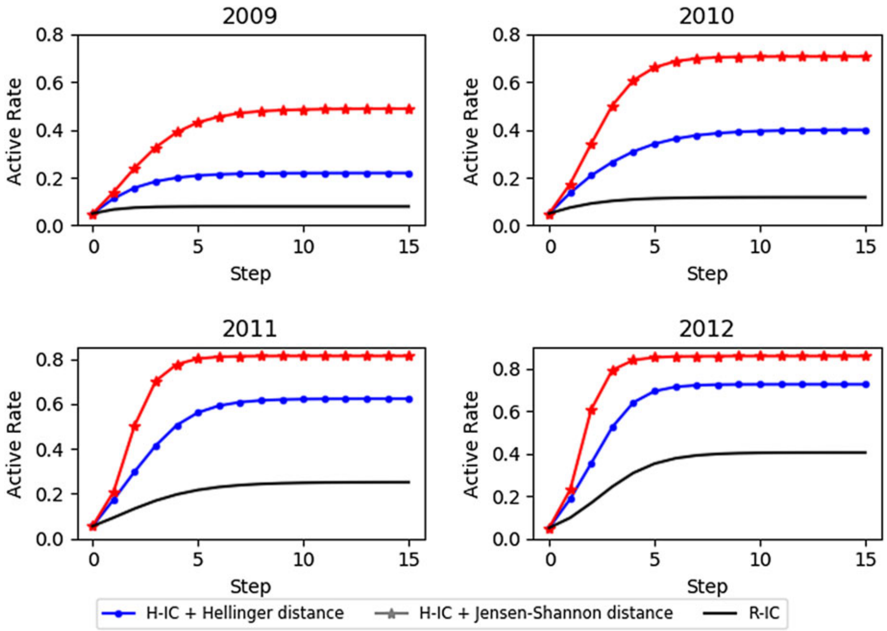

FIGURE 9 H-IC Diffusion on static co-author network

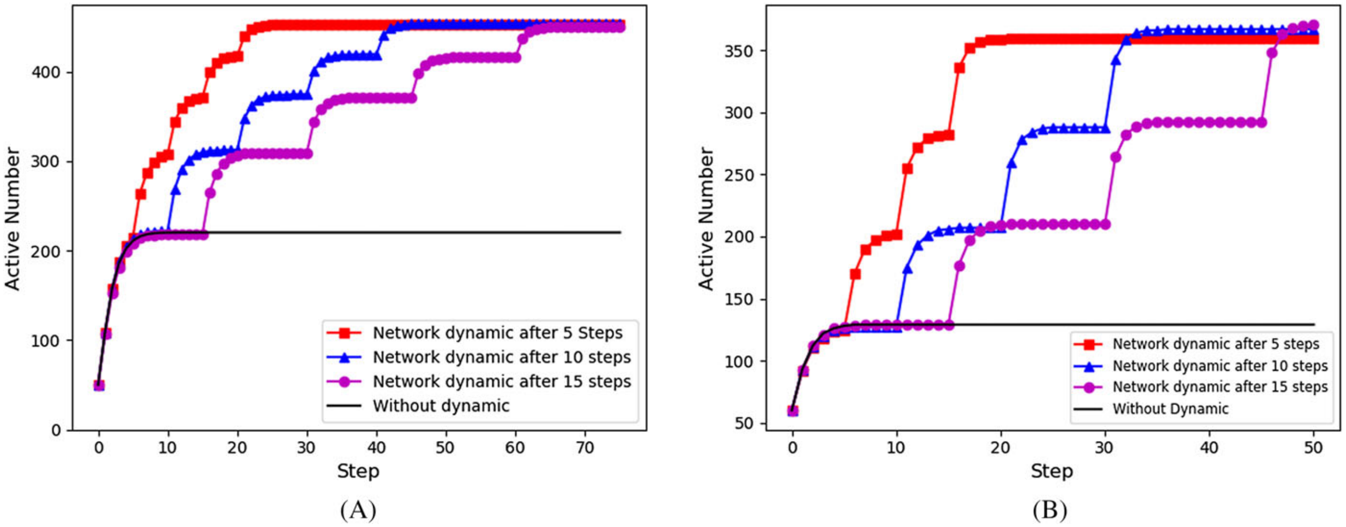

FIGURE 10 H-IC diffusion on dynamic coauthor network. A, H-IC + Hellinger distance; B, H-IC + Jensen-Shannon distance

particular, the diffusion process sometimes reaches a steady state, for instance, start from step 8 or step 16 for the network where the fluctuation occurs after every 10 steps of spreading. However, we can see that this state is broken since the agent's network is updated. Consequently, we can conclude that the steady state in the diffusion process can be broken if there is network's transformation.

6 | PRETOPOLOGICAL CASCADE MODEL FOR SOCIAL NETWORK RELATED TO TEXTUAL INFORMATION (TEXTUAL-PCM)

In Section 5, we revealed the diffusion process on a social network related to textual information through the *H-IC* model. Nevertheless, we have only exploited aspects where the classical way to define the set of neighbors is a union of neighbors of elements from the active set and the spreading process takes place on a single relational network. The issue is to how to demonstrate the propagation process on a complex social network with multiple relations and more complicated way of neighborhood determinations. In our previous research,[16] we have proposed PCM, which can capture more complex neighborhoods set based on *pseudoclosure* function built from pretopology theory. PCM has been demonstrated in detail on the stochastic graph and multirelational network. Therefore, in this section, we propose an expanded model of PCM which will be applied for a social network related to textual content, namely *Textual-PCM*. The dissemination mechanism of *Textual-PCM* will be illustrated in detail in Section 6.1.

6.1 | Pretopological cascade model for social network related to textual information (Textual-PCM)

First, we construct multirelational agent's network in which each node have a probability distribution on T different topics. Suppose that there is a family of binary reflexive relations between agents $(R_i)_{i=1,\dots,n}$, which contain T relations ($T <= n$) related to T topics $(R_{topic_j})_{j=0,\dots,T-1}$. $uR_{topic_j}v$ illustrate that u and v are interested in common topic j with probability greater than threshold p_0. Moreover, $n - T$ relations will be considered '*real*' relations such as friend, follow, co-author, etc. The Textual-PCM algorithm is presented in Algorithm 4.

The diffusion process with *Textual-PCM* will follow two major steps:

Step 1: Target neighborhood set from active set at step $t - 1$ for activation.

Define set of neighbors $N(A_{t-1})$ of A_{t-1} based on strong pseudoclosure function (Equation 13) with different strong levels. A strong level of the pseudoclosure function depends on the threshold value k, which represents for the number relations that each element x needs to be satisfied. If the value of k increases, then the number of relationships is considered for defining neighborhoods set $|I|$ increases. This is synonymous with the $|a_s(A)|$ decrease since an inactive agent x will become a neighbor of A_{t-1} if it have all relationships $i \in I$ with A_{t-1}. Therefore, a stronger pseudoclosure function will narrow the scope of the neighborhood set and can lead to a slower propagation process.

$$a_s(A,k) = \{x \in X | \exists I \subset I_n, |I| >= k, \forall i \in I, V_i(x) \cap A \neq \emptyset\} \quad \forall A \subset X \tag{13}$$

Step 2: Each element $v \in N(A_{t-1}) \setminus A_{t-1}$ will be influenced by A_{t-1} to be an active node if it satisfies the following criteria:

- Probability rule: active element $u \in A_{t-1}$ infects the inactive elements $v \in N(A_{t-1})$ with a probability $P_{u,v}$ (see Equation 14).

$$P_{u,v} = \text{Homo}(u,v) \tag{14}$$

- ∗ Threshold rule: Inactive elements $v \in N(A_{t-1})$ will be active if the sum of all influence of all incoming elements of v is greater than a threshold θ_v (see Equation 15).

$$\sum_{u \in A_{t-1}} \text{Homo}(u, v) > \theta_v \tag{15}$$

$$\theta_v = \text{AverageHomo} * \text{AverageDegree}(v) \tag{16}$$

Homo(u, v) is defined at Section 5.2. Moreover, we estimate the diffusion threshold for each node v (θ_v) based on average homophily on the whole network and the average degree of v (see at Equation 16)

```
Algorithm 4 Textual-PCM
Require: Multi-relational textual agent's network $G(V, (R_i)_{i=1,\dots,n}, (E_i)_{i=1,\dots,n})$
Require: $A_0$: seed set; $k$
procedure TEXTUAL-PCM($G, I_0$)
    $t \leftarrow 0, A^{total} \leftarrow A_0$
    while infection occur do
        $t \leftarrow t + 1; A_t \leftarrow \emptyset$
        $N_t \leftarrow a_s(A^{total}, k)$
        for $v \in N_t - A^{total}$ do
            if satisfy activation condition then
                $A_t \leftarrow A_t \bigcup \{v\}$
            end if
        end for
        $A^{total} \leftarrow A^{total} \bigcup A_t$;
    end while
    return $A^{total}$                                        ▷ Output
end procedure
```

6.2 | Toy example

In this subsection, we illustrate *Textual-PCM* on a small multirelational network built from the Twitter. We construct agent's network with 100 nodes in which each node have a probability distribution on ten different topics. Eleven relations are considered: R_{topic_j} $j = \overline{0, 9}$ (correspond $R_{i=1,\dots,10}$), and R_{follow} (correspond R_{11}). $uR_{\text{follow}}v$ mean that *u follow v* while $uR_{\text{topic}_j}v$ illustrate that u and v are interested in common topic j with probability greater than threshold $p_0 = 0.2$. The diffusion process will follow two steps mention at 6.1 in which the first step is to capture the set of neighborhoods based on a pseudoclosure function and apply threshold rule for the second step.

Setup

Firstly, we randomly chose 100 users from Twitter and crawled 100 tweets for each user. ATM is utilized to estimate the topic probability distribution when the number of topics is 10. After that, we created a multirelational network that include 100 nodes (are identified with ID from 0 to 99) and 11 relations. We implemented the spreading process with three different seed sets $|A0| = 5, |A0| = 10$ & $|A0| = 15$. At each time step t, we examined the determination of neighborhood set $N(A_{t-1}$ *of* $N(A_{t-1}$ (Equation 13) with different values of $k, k = \overline{1, 4}$.

Result

Figure 11 demonstrates the results of *Textual-PCM* on the multirelational network with three different seed sets and four values of k. We can see that the value of k is higher than the ability to spread harder. In the first case with $A0 = [38, 25, 10, 24, 46]$ (see Figure 11A), the diffusion process with $k = 1$ reaches 90 and 100 active agents at step 2 and step 3, respectively. Moreover, the spreading process with $k = 2$ is slower compared with $k = 1$ since it sequentially obtains 40 *and* 78 active agents at steps 2 *and* 3 and a stable status with 82 active agents at step 4. However, there is a slight increase in the number of active agents in case $k = 3$, approximately a quarter of $k = 2$. In particular, the propagation process activate only one agent with $k = 4$. Similarly, the spreading process with the other two cases of seed sets (see Figures 11B and C) also illustrate that propagation speed is slower if value k increases. Therefore, we can conclude that the pseudoclosure function allows to define more complex neighborhood set compared with the classical method, and if the pseudoclosure function is more strong, then it is more difficult in the spreading process.

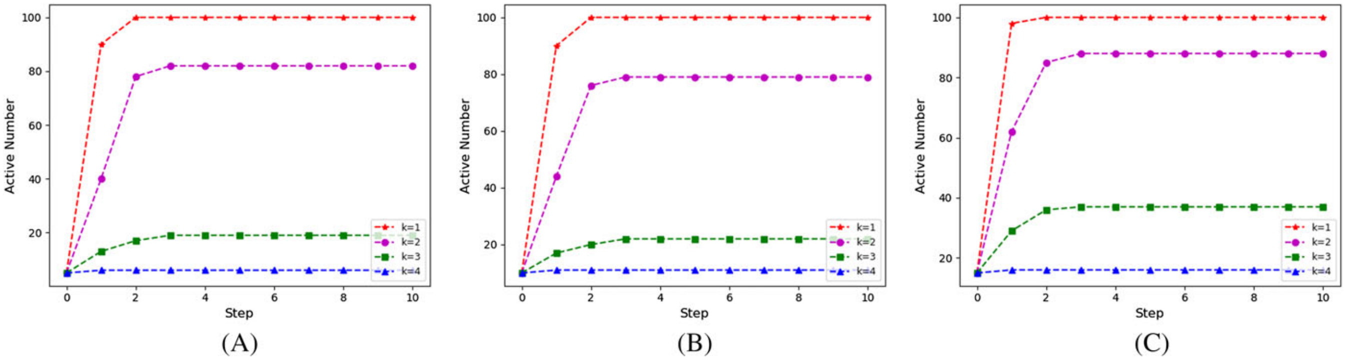

FIGURE 11 Textual-PCM on multirelational network. A, Different values of *k* & |*A*0| = 5; B, Different values of *k* & |*A*0| = 10; C, Different values of *k* & |*A*0| = 15

6.3 | Discussion

We propose the Textual-PCM model to demonstrate the propagation process on a complex network where contains textual information and multirelations. Since pretopology is a mathematical tool for modeling the concept of proximity, it is usually utilized for analyzing and modeling the structure of a complex network. The core of pretopology is the propagation operator called pseudoclosure function. The advantage of applying pretopology theory is that it allows to capture more complex neighborhood sets instead of the classical way, which is the combination of the neighborhood of each element. In this section, we propose a pseudoclosure function with different strength levels that depend on the constraint about the number of relations. This function allows to capture a complex neighborhood set with different tightness levels. The contributions of this section are as follows:

- We present a method to a pseudoclosure function based on the family of relationships with different strong levels.
- We propose an expanded model of PCM which will be presented on a multirelational network related to textual information, namely Textual-PCM, in which a strong pseudoclosure function built from pretopology is utilized to determine the neighborhood set. Besides, the probability rule and threshold rule based on homophily are proposed for activation.
- We conduct a toy example to illustrate our approach.

However, besides the advantages of Textual-PCM, there are two drawbacks, as follows:

- Strong pseudoclosure takes a lot of time to calculate. Therefore, the spreading process spends a long time if it takes place on a large-scale network.
- In several cases, the diffusion cannot take place by too tight constraints in the pseudoclosure function.

7 | CONCLUSION

In this work, we proposed a fresh approach for dynamic SNA in combination with ABM, ATM, and Pretopology. First, we proposed an analytical model for a dynamic social network associated with the textual information using ABM and ATM, namely, *Textual-ABM*. The advantage of *Textual-ABM* is to provide a process for constructing and simulating the fluctuation of a social network that includes not only network structure transformation but also agent's interest variation. Agent's interest is revealed through topic probability distribution, which is estimated based on textual information using ATM. Besides, an online visualization application of *Textual-ABM* is introduced to support for the user observing the dynamics of the social network. On the other hand, we exploited information dissemination phenomena on social networks related to textual information by introducing two expanded spreading models. The first is IC model based on homophily, namely *H-IC*. In *H-IC*, the infected probability between two agents is estimated by homophily based on agent's topic distribution. *H-IC* is illustrated on both static and dynamic agent's network. Experimental results proved that the effectiveness of *H-IC* on the static network outperforms comparison with *R-IC*. Moreover, experimental results also illustrated the transformation about the active number for the propagation on the dynamic agent's network instead of reaching and maintaining a steady state on a static network. The second is an extended model of PCM from our previous work,[16] namely *Textual-PCM*. PCM is a cascade propagation model in which neighborhood set is defined through pseudoclosure function in pretopology. The pseudoclosure function provides a general way to determine the set of the neighborhood instead of the classical method. In this work, we expand PCM detail for the social network associated with the textual information. *Textual-PCM* is performed on multirelational agent's network where neighborhoods set is defined through the strong pseudoclosure function with different strong levels. Besides, we can apply different diffusion rules, including probability rule and threshold rule based on homophily. A toy example is given, and experiments and discussion are demonstrated for *Textual-PCM*.

ORCID

Thi Kim Thoa Ho https://orcid.org/0000-0002-1564-6520

REFERENCES

1. Grandjean M. A social network analysis of Twitter: mapping the digital humanities community. *Cogent Arts Humanit*. 2016;3(1):1171458. https://doi.org/10.1080/23311983.2016.1171458
2. Otte E, Rousseau R. Social network analysis: a powerful strategy, also for the information sciences. *J Inf Sci*. 2002;28(6):441-453.
3. Girvan M, Newman MEJ. Community structure in social and biological networks. *Proc Natl Acad Sci USA*. 2002;99(12):7821-7826.
4. Newman MEJ. Fast algorithm for detecting community structure in networks. *Phys Rev E*. 2004;69:066133.
5. Li M, Wang X, Gao K, Zhang S. A survey on information diffusion in online social networks: models and methods. *Information*. 2017;8(4):118. https://doi.org/10.3390/info8040118
6. Goldenberg J, Libai B, Muller E. Talk of the network: a complex systems look at the underlying process of word-of-mouth. *Mark Lett*. 2001;12(3):211-223. https://doi.org/10.1023/A:1011122126881
7. Granovetter M. Threshold models of collective behavior. *Am J Sociol*. 1978;83(6):1420-1443. https://doi.org/10.1086/226707
8. Legendi RO, Gulyás L. Agent-based dynamic network models: validation on empirical data. In: Kamiński B, Koloch G, eds. *Advances in Social Simulation*. Berlin, Germany:Springer Berlin Heidelberg; 2014:49-60. Advances in Intelligent Systems and Computing; 229.
9. Holme P, Saramäki J. Temporal networks. *Phys Rep*. 2012;519(3):97-125.
10. Carley KM, Martin MK, Hirshman BR. The etiology of social change. *Top Cogn Sci*. 2009;1(4):621-650. https://doi.org/10.1111/j.1756-8765.2009.01037.x
11. Blei DM, Ng AY, Jordan MI. Latent Dirichlet allocation. *J Mach Learn Res*. 2003;3:993-1022.
12. Rosen-Zvi M, Griffiths T, Steyvers M, Smyth P. The author-topic model for authors and documents. Paper presented at: 20th Conference on Uncertainty in Artificial Intelligence; 2004; Arlington, VA.
13. Kempe D, Kleinberg J, Tardos É. Maximizing the spread of influence through a social network. Paper presented at: Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2003; New York, NY.
14. Kempe D, Kleinberg J, Tardos É. Influential nodes in a diffusion model for social networks. In: Caires L, Italiano GF, Monteiro L, Palamidessi C, Yung M, eds. *Automata, Languages and Programming*. Berlin, Germany:Springer Berlin Heidelberg; 2005:1127-1138. Lecture Notes in Computer Science; 85.
15. Kimura M, Saito K. Tractable models for information diffusion in social networks. In: Fürnkranz J, Scheffer T, Spiliopoulou M, eds. *Knowledge Discovery in Databases: PKDD 2006*. Berlin, Germany:Springer Berlin Heidelberg; 2006:259-271. Lecture Notes in Computer Science; 4213.
16. Bui QV, Ben Amor S, Bui M. Stochastic pretopology as a tool for topological analysis of complex systems. In: Nguyen NT, Hoang DH, Hong T-P, Pham Hoang, Trawiński B, eds. *Intelligent Information and Database Systems*. Cham, Switzerland: Springer International Publishing; 2018:102-111. Lecture Notes in Computer Science; 10752.
17. Niazi M, Hussain A. Agent-based computing from multi-agent systems to agent-based models: a visual survey. *Scientometrics*. 201189:479.
18. Hoffman M, Blei DM, Wang C, Paisley J. Stochastic variational inference. *J Mach Learn Res*. 2013;14(1):1303-1347.
19. Belmandt Z. *Basics of Pretopology*. Paris, France:Hermann; 2011.
20. Bui QV, Sayadi K, Bui M. A multi-criteria document clustering method based on topic modeling and pseudoclosure function. *Informatica*. 2016;40(2).
21. Levorato V. Modeling groups in social networks. Paper presented at: 25th European Conference on Modelling and Simulation; 2011; Krakow, Poland.
22. Berge C. *The Theory of Graphs*. Mineola, NY:Dover Publications; 1962.
23. Shakarian P, Bhatnagar A, Aleali A, Shaabani E, Guo R. *Diffusion in Social Networks*. Cham, Switzerland:Springer International Publishing; 2015.
24. Bui QV, Sayadi K, Amor SB, Bui M. Combining Latent Dirichlet Allocation and K-Means for Documents Clustering: Effect of Probabilistic Based Distance Measures. In: Nguyen NT, Tojo S, Nguyen LM, Trawiński B, eds. *Intelligent Information and Database Systems*. Cham, Switzerland:Springer International Publishing; 2017:248-257. Lecture Notes in Computer Science; 10191.
25. Gayraud NT, Pitoura E, Tsaparas P. Diffusion maximization in evolving social networks. Paper presented at: 2015 ACM Conference on Online Social Networks; 2015; Stanford, CA.
26. Zhuang H, Sun Y, Tang J, Zhang J, Sun X. Influence maximization in dynamic social networks. Paper presented at: IEEE 13th International Conference on Data Mining; 2013; Dallas, TX.
27. Buntine W. Estimating likelihoods for topic models. In: Zhou Z-H, Washio T, eds. *Advances in Machine Learning*. Berlin, Germany:Springer Berlin Heidelberg; 2009:51-64. Lecture Notes in Computer Science; 5828.

How to cite this article: Ho TKT, Bui QV, Bui M. Dynamic social network analysis: A novel approach using agent-based model, author-topic model, and pretopology. *Concurrency Computat Pract Exper*. 2019;e5321. https://doi.org/10.1002/cpe.5321