Vietnam Journal of Computer Science
Vol. 6, No. 3 (2019) 1–25
© The Author(s)
DOI: 10.1142/S2196888819500155

Information Diffusion on Complex Networks: A Novel Approach Based on Topic Modeling and Pretopology Theory

Thi Kim Thoa Ho*
EA 4004 Human and Artificial Cognition
(CHArt) Laboratory
École Pratique des Hautes Études
PSL Research University
4-14 Rue Ferrus, 75014 Paris, France
**thi-kim-thoa.ho@etu.ephe.psl.eu*

Quang Vu Bui
University of Sciences, Hue University
77 Nguyen Hue Street, Hue City, Vietnam

Marc Bui
EA 4004 Human and Artificial Cognition
(CHArt) Laboratory
École Pratique des Hautes Études
PSL Research University, Les Patios Saint-Jacques
4-14 Rue Ferrus, 75014 Paris, France

Received 31 December 2018
Accepted 30 April 2019
Published 3 June 2019

In this research, we exploit a novel approach for propagation processes on a network related to textual information by using topic modeling and pretopology theory. We first introduce the *textual agent's network* in which each agent represents a node which contains specific properties, particularly the agent's interest. Agent's interest is illustrated through the topic's probability distribution which is estimated based on textual information using topic modeling. Based on *textual agent's network*, we proposed two information diffusion models. The first model, namely *Textual-Homo-IC*, is an expanded model of independent cascade model in which the probability of infection is formed on homophily that is measured based on agent's interest similarity. In addition to expressing the *Textual-Homo-IC* model on the static network, we also reveal it on dynamic agent's network where there is transformation of not only the structure but also the node's properties during the spreading process. We conducted experiments on two collected datasets from NIPS and a social network platform, Twitter, and have attained satisfactory

*Corresponding author.

results. On the other hand, we continue to exploit the dissemination process on a multi-relational agent's network by integrating the pseudo-closure function from pretopology theory to the cascade model. By using pseudo-closure or stochastic pseudo-closure functions to define the set of neighbors, we can capture more complex kind of neighbors of a set. In this study, we propose the second model, namely *Textual-Homo-PCM*, an expanded model of pretopological cascade model, a general model for information diffusion process that can take place in more complex networks such as multi-relational networks or stochastic graphs. In *Textual-Homo-PCM*, pretopology theory will be applied to determine the neighborhood set on multi-relational agent's network through pseudo-closure functions. Besides, threshold rule based on homophily will be used for activation. Experiments are implemented for simulating *Textual-Homo-PCM* and we obtained expected results. The work in this paper is an extended version of our paper [T. K. T. Ho, Q. V. Bui and M. Bui, Homophily independent cascade diffusion model based on textual information, in *Computational Collective Intelligence*, eds. N. T. Nguyen, E. Pimenidis, Z. Khan and B. Trawiski, Lecture Notes in Computer Science, Vol. 11055 (Springer International Publishing, 2018), pp. 134–145] presented in *ICCCI 2018* conference.

Keywords: Complex network; information diffusion; independent cascade model; agent-based model; latent Dirichlet allocation; author–topic model; pretopology; stochastic pretopology; pretopological cascade model.

1. Introduction

In recent years, research on the process of information diffusion through social networks has attracted the attention of researchers with applications in various fields including computer science, economy, and biology. Information propagation has been extensively researched in networks, with the objective of observing the information spreading among objects when they are connected with each other. Recently, there are numerous diffusion models which have been proposed including the linear threshold (LT) model,[1] independent cascade (IC) model[2] and so on. The IC model has been used extensively since it is the simplest cascade model and is successful at explaining diffusion phenomena in social networks.[2] The propagation process in IC occurs in discrete time steps t which include two major substeps: determine inactive neighborhood set $\eta^{\text{out}}(u)$ of newly active nodes u at step $t-1$ and each inactive node $v \in \eta^{\text{out}}(u)$ will be infected by u with a probability $P(u,v)$. In IC, each edge is associated with a probability of infection independently which is usually assigned by a uniform distribution.[3–5] Nevertheless, perhaps from the fact that the infected probability from one object to another depends on the similarity or homophily among them, for instance, the probability that two scientists in the common field incorporate to write a paper is higher in comparison with the different fields. In another instance, a user A on Twitter finds it easy to *follow* user B when A has common interests with B. Therefore, we estimate the probability of infection based on similarity or homophily.

Homophily is the tendency of individuals to associate with similar others.[6,7] There are two principal approaches to measure homophily including the first one based on a single characteristic and the combination of multiple features for the second. For the first approach, homophily is classified into two types including status homophily and value homophily in which the former refers to the similarity in socio-demographic traits, such as race, age, gender, etc., while the similarity in internal states for the latter, such as opinions, attitudes, and beliefs.[6,7] Besides, Laniado *et al.* analyzed the presence of gender homophily in relationships on the Tuenti Spanish social network.[8] On the other hand, with the second approach, Aiello *et al.*[9] discovered homophily from the context of tags of social networks including Flickr, Last.fm, and aNobii. Additionally, Cardoso *et al.*[10] explored homophily from hashtags on Twitter. However, in general, these methods have not exploited the textual information related to users yet while it contains significant information for similarity analysis, for instance, based on the content of papers, we can define whether the authors research in the same narrow subject or not, or we can determine which are common interests between two users on Twitter based on their tweets. For that reason, we propose a method of homophily measurement based on textual content. A fundamental technology for text mining is *vector space model* (VSM)[11] where each document is represented by word-frequency vector. Nevertheless, two principal drawbacks of VSM are the high dimensionality as a result of the high number of unique terms in text corpora and insufficiency to capture all semantics. Therefore, topic modeling was proposed to solve these issues. Recently, there are dissimilar methods of topic modeling which include *latent Dirichlet allocation* (LDA),[12]*author–topic model* (ATM),[13] etc. In this study, we chose LDA and ATM to estimate the topic's probability distribution of users.

In this study, we first propose an expanded model of independent cascade model, namely *Textual-Homo-IC*. The first step of the propagation process is network construction. We construct a heterogeneous network related to textual information, namely *textual agent's network*, where each node is represented by an agent. In *textual agent's network*, each agent contains specific characteristics including *ID*, *neighbors*, and *topic's probability distribution*. Textual information of users will be used to estimate the topic's distribution using *topic modeling*. Besides, there may be one or more relations between agents. In *Textual-Homo-IC*, we just consider the spreading process on single-relational textual agent's network, and mainly focus on infected probability estimation from an active object to inactive another based on their similarity or homophily. Particularly, homophily is measured based on the topic's distribution of agents. *Textual-Homo-IC* is demonstrated on static and dynamic textual agent's networks in which the network structure and characteristics of agents have remained during the propagation process for the former while there is a variation for the latter. Some experiments were implemented on co-author network and Twitter with the combinations of two methods LDA and ATM for estimating topic's distribution of users and two distance measurements Hellinger distance and Jensen–Shannon divergence for measuring homophily. On the static networks, the results demonstrated that the effectiveness of *Textual-Homo-IC* outperforms in comparison with random diffusion. Additionally, our results also illustrated the fluctuation of the active number for the diffusion process on a dynamic network instead of attaining and remaining in a stable state on a static network.

In addition, we propose an expanded model of *pretopological cascade model* (PCM) from our previous research,[14] namely *Textual-Homo-PCM*. *Textual-Homo-IC* is the standard IC model with infected probability based on textual homophily; however, we just exploit spreading phenomena on single-relational textual agent's network. Therefore, we continue to exploit the propagation process on multi-relational textual agent's network. In the previous study,[14] we proposed a general PCM model which is a cascade diffusion model on complex networks including stochastic graph or multi-relational network using stochastic pretopology (SP) theory. Stochastic pseudo-closure function defined is utilized to determine neighborhood set for the spreading process. The predecessor of stochastic pretopology is pretopology[15] which is a mathematical tool for modeling the concept of proximity. It is usually used for analyzing and modeling the structure of a complex network. The core of pretopology is the propagation operator called pseudo-closure function. It is usually defined for neighborhood collection. However, the phenomena in a complex system do not always conform to a certain mechanism that contains stochastic or uncontrolled factors. Therefore, in addition to modeling the dynamics of structural phenomena by pretopology, stochastic pretopology was also proposed with a combination of pretopology and random set theory.[14,15] In this study, we will illustrate in detail the PCM on multi-relational textual agent's network namely *Textual-Homo-PCM* in which we take into account a multi-relational textual agent's network as a complex network, and apply pretopology theory and stochastic pretopology in neighborhood determination through various pseudo-closure functions including *strong*, *weak*, and *stochastic pseudo-closure*. Besides, we apply the threshold rule to diffusion based on homophily. Several experiments were conducted on co-author network and Twitter. Experimental results illustrated the impact of neighborhood set determination on multi-relation networks for spreading process performance.

The structure of our paper is organized as follows: Section 2 reviews the preliminaries. Section 3 illustrates the *textual agent's network* construction. *Textual-Homo-IC* models are proposed in Sec. 4 with experiments, results, and evaluation. Section 5 demonstrates *Textual-Homo-PCM* with experiments, results, and discussion. Finally we conclude our work in Sec. 6.

2. Preliminaries

2.1. *The independent cascade*

2.1.1. *Model definition*

We assume a network $G = (V, \Gamma, W)$, where:

- V is a set of vertices.
- $\Gamma : V \to \mathcal{P}(V)$ is a neighborhood function. $\mathcal{P}(V)$ is the power set of set V:
  - $\Gamma(x)$ is the set of outcoming neighborhoods of node x.
  - $\Gamma^{-1}(x)$ is the set of incoming neighborhoods of node x.

- $W : V \times V \rightarrow \mathbb{R}$ is the weight function:
  - In LT model, $W(x, y)$ is the weight of edge between two nodes x and y.
  - In IC model, $W(x, y)$ is the probability of node y infected from node x.

The diffusion process occurs in discrete time steps t. If a node adopts a new behavior or idea, it becomes active, otherwise it is inactive. An inactive node has the ability to become active. The sets of active nodes, and newly active nodes at time t are considered as A_t and A_t^{new}, respectively. The tendency of an inactive node x to become active is positively correlated with the number of its active incoming neighbors $\Gamma^{-1}(x)$. Also, we assume that each node can only switch from inactive state to active state, and an active node will remain active for the rest of the diffusion process. In general, we start with an initial seed set A_0 and through the diffusion process, for a given inactive node x, its active neighbors attempt to activate it. The process runs until no more activations occur.

2.1.2. *Independent cascade model*

In IC model, there is a probability of infection associated with each edge. $W(x, y)$ is the probability of node x infecting node y. This probability can be assigned based on the frequency of interactions, geographic proximity, or historical infection traces. Each node, once infected, has the ability to infect its neighbor in the next time step based on the probability associated with that edge. At each time step t, each node $x \in A_{t-1}^{\text{new}}$ infects the inactive neighbors $y \in \Gamma(x)$ with a probability $W(x, y)$. The propagation continues until no more infection can occur (see Algorithm 2.1).

2.2. *Agent-based model*

An agent-based model (ABM) is a class of computational models for simulating the actions and interactions of autonomous agents. ABM has been utilized in numerous fields, for instance, biology, ecology, and social science.[16] ABM contains

Algorithm 2.1. Independent cascade model

```
Require: Network $G = (V, \Gamma, W)$, seed set $A_0$.
procedure IC-MODEL($G, A_0$)
    $t \leftarrow 0, A^{\text{total}} \leftarrow A_0, A^{\text{new}} \leftarrow A_0$
    while infection occurs do
        $t \leftarrow t + 1$; $A_t \leftarrow \emptyset$
        for $u \in A^{\text{new}}$ do
            $A_t(u) \leftarrow \{v^{\text{inactive}} \in \Gamma(u),\ q \leq W(u, v^{\text{inactive}})\}; q \sim U(0, 1)$    ▷ Newly active node $u$ infects inactive node $v$ with probability $W(u, v)$
            $A_t \leftarrow A_t \cup A_t(u)$
        $A^{\text{total}} \leftarrow A^{\text{total}} \cup A_t$; $A^{\text{new}} \leftarrow A_t$
    return $A^{\text{total}}$    ▷ Output
```

three principal elements including agents, their environment, and interactive mechanisms among agents. First, agents are heterogeneous entities which comprise diverse characteristics and behaviors. Second, agent's environment is a space that is responsible for reflecting the structure of the overall system and supplying agents their perceptions and enabling their actions. Third, interaction is a form of information exchange among agents which resulted in perception and behavior. Particularly, the essence of an ABM is the dynamics of the global system emerges from the local interactions among its composing parts.

2.3. *Topic modeling*

2.3.1. *LDA*

LDA[12] is a generative statistical model of a corpus. In LDA, each document may be taken into account as a combination of multiple topics and each topic is demonstrated by a probability distribution of words. The generative model of LDA, which is described with a graphical model in Fig. 1(a), proceeds as follows:

(1) Choose distribution over topics $\theta_i \sim \text{Dirichlet}(\alpha)$ for each document.
(2) Choose distribution over words $\phi_j \sim \text{Dirichlet}(\beta)$ for each topic.
(3) For each of the word position i, j:

  (3.1) Choose a topic $z_{ij} \sim \text{Multinomial}(\theta_i)$.
  (3.2) Choose a word $w_{i,j} \sim \text{Multinomial}(\phi_{z_{i;j}})$.

2.3.2. *ATM*

ATM[13] is a generative model for documents that expands LDA to incorporate author's information. Each author is associated with a mixture of topics where topics are multinomial distributions over words. The words in a collaborative paper are assumed to be the result of a mixture of the authors' topics. The generative model of ATM, which is described with a graphical model in Fig. 1(b), proceeds as follows:

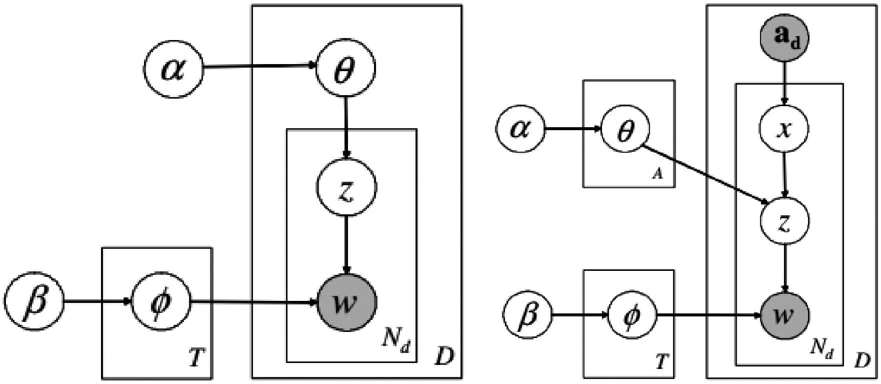
(1) For each author $a = 1, \ldots, A$ choose $\theta_a \sim \text{Dirichlet}(\alpha)$. For each topic $t = 1, \ldots, T$ choose $\phi_t \sim \text{Dirichlet}(\beta)$.
(2) For each document $d = 1, \ldots, D$:

  (2.1) Given the vector of authors a_d.
  (2.2) For each word $i = 1, \ldots, N_d$:

    (2.2.1) Choose an author $x_{di} \sim \text{Uniform}(a_d)$.
    (2.2.2) Choose a topic $z_{di} \sim \text{Discrete}(\theta_{x_{di}})$.
    (2.2.3) Choose a word $w_{di} \sim \text{Discrete}(\phi_{z_{di}})$.

2.3.3. *Update process of LDA and ATM*

LDA and ATM can be updated with additional documents after training has been finished. This update procedure is executed by expectation maximization



(a) LDA.

(b) ATM.

Fig. 1. Text mining methods: LDA and ATM.

(EM) — iterating over new corpus until the topics converge. This process is equal to the online training of Hoffman *et al.*[17] There are already several available packages for topic modeling including *topicmodels* or *lda* in R or *Gensim*[a] in Python. In this study, we chose Gensim for training and updating the topic modeling.

2.4. *Pretopology theory*

2.4.1. *Pretopological notions*

Definition 2.1. A pretopological space is an ordered pair (X, a), where X is a set and $a : \mathcal{P}(X) \rightarrow \mathcal{P}(X)$ is a *pseudo-closure* 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).

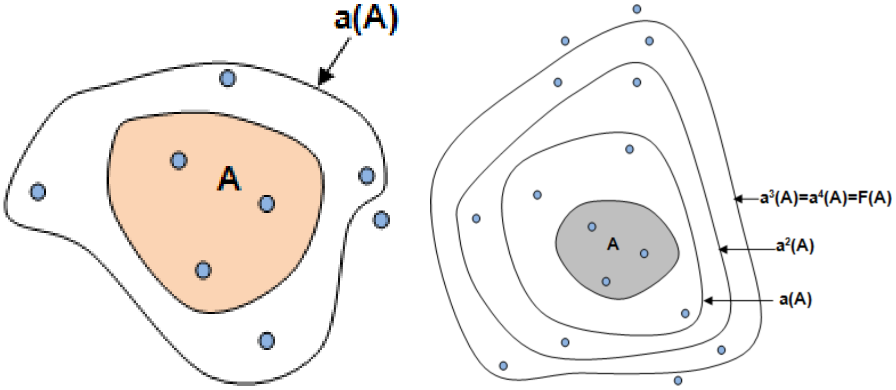
It is important to note that, by defining *pseudo-closure* $a(\cdot)$ (see Fig. 2(a)), we do not suppose that it is an idempotent transform. Then, conversely as it happens in topology, we can compute: $a(A), a(a(A)), a(a(a(A))), \ldots, a^k(A)$ (see Fig. 2(b)). So, *pseudo-closure* allows, for each of its applications, to add elements to a set departure according to defined characteristics. The starting set gets bigger but never reduces.

Definition 2.2. Let (X, a) be a pretopological space, $\forall\, A, A \subset X$. A is a closed subset if and only if $a(A) = A$.

Definition 2.3. Given a pretopological space (X, a), call the closure of A, when it exists, the smallest closed subset of X which contains A. The closure of A is denoted by $F(A)$.

Closure is very important because of the information it gives about the influence or reachability of a set, meaning, for example, that a set A can influence or reach the elements in $F(A)$, but not further (see Fig. 2(b)).

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



(a) Pseudo-closure of A.

(b) Closure of A.

Fig. 2. Pseudo-closure and closure functions.

Hence, it is necessary to build pretopological spaces in which the closure always exists. We present in the following different types of pretopological spaces which are less general than the basic ones but for which, good properties are fulfilled by neighborhoods as well as allowing the existence of closure.

2.4.2. *Pretopological spaces*

Definition 2.4. A pretopology space (X, a) is called $\mathcal{V}$-type space if and only if

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

Definition 2.5. A Pretopology space (X, a) is called $\mathcal{V}_D$-type space if and only if

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

Definition 2.6. A pretopology space (X, a) is called $\mathcal{V}_S$-type space if and only if

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

2.4.3. *Pretopology and binary relationships*

Suppose we have a family $(R_i)_{i=1,\dots,n}$ of binary reflexive relations on a finite set X. We call $L = \{R_1, R_2, \dots, R_n\}$ as a set of relations. For each relation R_i, we can define pretopological structure by considering the following subset: $\forall\, i = 1, 2, \dots, n$, $\forall\, x \in X$, $V_i(x)$ is defined by

$$V_i(x) = \{y \in X \mid x\, R_i\, y\}.$$

Then, the pseudo-closure $a(\cdot)$ is defined by

$$a(A) = \{x \in X \mid \forall\, i = 1, 2, \dots, n, V_i(x) \cap A \neq \emptyset\}, \quad \forall\, A \subset X. \tag{4}$$

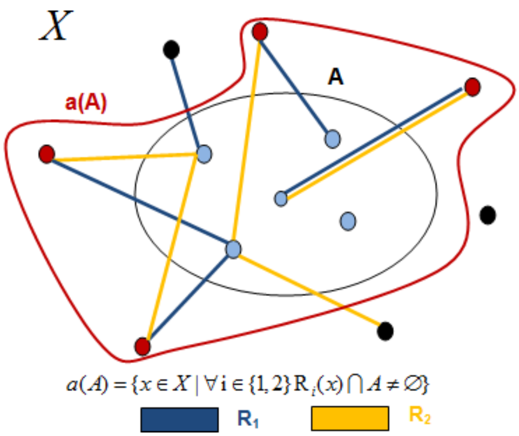

X
a(A)
A
$a(A) = \{x \in X \mid \forall i \in \{1,2\} R_i(x) \cap A \neq \varnothing\}$
R_1
R_2

Fig. 3. Pseudo-closure of A in a binary space.

Pretopology defined on X by $a(\cdot)$ using the intersection operator is often called the strong pretopology. Figure 3 gives an example for strong pretopology built from two relationships.

Proposition 2.1. *$a(\cdot)$ determines on X a pretopological structure and the space (X, a) is of V-type pretopological space.*

Please refer to the work of Belmandt[15] for more details about pretopology theory and to the work of Bui *et al.*[14] for some other ways to build pseudo-closure function in different spaces such as metric space, valued space, and space equipped with a neighbors function.

2.5. *Stochastic pretopology*

Stochastic pretopology was first basically introduced in Chap. 4 of the book *Basics of Pretopology*[15] by utilizing simple random set to propose three ways for defining stochastic pretopology. There are some applications of stochastic pretopology including modeling pollution phenomena[18] or studying complex networks via a stochastic pseudo-closure function defined from a family of random relations.[19] Additionally, in our previous research,[14] we proposed another approach for building stochastic pretopology by using finite random set (FRS) theory.[20]

2.5.1. *Finite random set*

From now on, V denotes a finite set. $(\Omega, \mathcal{A}, \mathbb{P})$ will be a *probability space*, where: Ω is a set, representing the *sample space* of the experiment; $\mathcal{A}$ is a σ-algebra on Ω, representing *events*; and $\mathbb{P} : \Omega \to [0, 1]$ is a *probability measure.*

Definition 2.7. An FRS with values in $\mathcal{P}(V)$ is a map $Y : \Omega \to \mathcal{P}(V)$ such that

$$Y^{-1}(\{A\}) = \{\omega \in \Omega : Y(\omega) = A\} \in \mathcal{A} \quad \text{for any } A \in \mathcal{P}(V). \tag{5}$$

The condition (5) is often called *measurability condition*. So, in other words, an FRS is a measurable map from the given probability space $(\Omega, \mathcal{A}, P)$ to $\mathcal{P}(V)$, equipped with a σ-algebra on $\mathcal{P}(V)$. We often choose that σ-algebra on $\mathcal{P}(V)$ is the discrete σ-algebra $\mathcal{E} = \mathcal{P}(\mathcal{P}(V))$. Clearly, a *finite random set* Y is a *random element* when we refer to the *measurable space* $(\mathcal{P}(V), \mathcal{E})$. This is because $Y^{-1}(\mathcal{E}) \subseteq \mathcal{A}$ since $\forall \mathbb{A} \in \mathcal{E}$, $Y^{-1}(\mathbb{A}) = \bigcup_{A\in\mathbb{A}} Y^{-1}(A)$.

2.5.2. *Definition of stochastic pretopology*

Definition 2.8. We define *stochastic pseudo-closure* defined on $\Omega \times V$, any function $a(\cdot,\cdot)$ from $\Omega \times \mathcal{P}(V)$ into $\mathcal{P}(V)$, such that:

(P1) $a(\omega, \emptyset) = \emptyset$, $\forall\, \omega \in \Omega$;
(P2) $A \subset a(\omega, A)$, $\forall\, \omega \in \Omega$, $\forall\, A, A \subset V$;
(P3) $a(\omega, A)$ is a finite random set, $\forall\, A, A \subset V$.

$(\Omega \times V, a(\cdot,\cdot))$ is then called the stochastic pretopological space.

Please refer to the work of Bui *et al.*[14] for some ways to build stochastic pseudo-closure function in different situations such as relational spaces, metric spaces, valued spaces, and spaces equipped with a neighbors function.

3. Agent's Network Related to Textual Information (Textual Agent's Network)

3.1. *Textual agent's network construction*

In this study, the network that we take into account for spreading process is the agent's network related to textual information, namely textual agent's network. A textual agent's network is given by $G(V, (R_i)_{i=1,2,\ldots,n}, (E_i)_{i=1,2,\ldots,n})$ in which V is the set of agents representing the nodes, $(R_i)_{i=1,2,\ldots,n}$ is a family of relations, and $(E_i)_{i=1,2,\ldots,n}$ is a set of edges among nodes corresponding to relations. Particularly, each agent in textual agent's network contains textual information, for instance, users on Twitter or authors on collaboration network. Construction of a textual agent's network will be illustrated in detail through the following steps.

3.1.1. *Text analysis with topic modeling*

Since we consider a network related to textual information where each user contains textual content, therefore technologies for text mining need to be used. A fundamental technology is VSM,[11] but there are two principal drawbacks including the high dimensionality as a result of the high number of unique terms in text corpora and insufficiency to capture all semantics. Therefore, topic modeling has been proposed to solve these issues. In this study, we chose two methods of topic modeling LDA and ATM to estimate the topic's probability distribution of users.

3.1.2. *Defining agents*

Each agent represents each node in the network. Agents are heterogeneous entities with three principal properties including *ID*, *Neighbors*, and *TP-Dis* (topic's probability distribution). LDA and ATM can be utilized to estimate the *TP-Dis* of users from their textual information. Depending on textual information that each agent owns, they have different probability distribution for the topic which is considered as agent's interest on the topic.

3.1.3. *Defining relations*

We take into account the agent's network with a family of relations $(R_i)_{i=1,2,\ldots,n}$ which contains the subsets of *real* relations and hidden relations. Besides *real* links such as "follow", co-author, friend, etc., we consider the *hidden* link based on the property *TP-Dis* of agents. By using topic modeling, each agent may be characterized by its topic distribution and also be labeled by the topic with the highest probability. In this subsection, we use this information to define the relations between two agents based on the way we consider the "similarity" between them. First, based on the label information, we can consider connecting agents if they have the same label. However, in some cases the probability of label topic is very small and it is not really good if we use this label to represent agent's interest. Hence, we just use the label information if its probability is higher than the threshold p_0. We define the *major topic* of each agent as follows:

Definition 3.1. $\mathrm{MTP}(a_i)$ is the major topic of agent a_i if $\mathrm{MTP}(a_i)$ is the topic with highest probability in the topic distribution of agent a_i and this probability is greater than the threshold p_0, $p_0 \geq 1/K$, where K is the number of topics:

$$\mathrm{MTP}(a_i) = \{k \mid \theta_{ik} = \max_j \theta_{ij} \text{ and } \theta_{ik} \geq p_0\}. \tag{6}$$

Considering two agents a_m, a_n with their major topics $\mathrm{MTP}(a_m)$, $\mathrm{MTP}(a_n)$, we see that a_m is close to a_n if they have the same major topic. So, we proposed a definition of binary relationship R_{MTP} of two agents based on their major topics as follows:

Definition 3.2. Agent a_m has the binary relationship R_{MTP} with a_n if a_m and a_n have the same major topic.

To demonstrate the dynamics of textual agent's network, we exploit not only the structure of the network but also agent's properties. First, the structure of the network can be transformed with the appearance of new agents or new connections. Moreover, topic's distribution of agents can fluctuate since agents own more text information through the interactive process. The problem is how to update the transformation of topic's distribution of users after a time period based on the existing topic modeling. In LDA, to make an estimate of topic's distribution of users, we consider each user corresponds to each document. Therefore, we cannot utilize the update mechanism of LDA to update topic's distribution of users when users have more documents in interaction. Instead of unusable update mechanism of LDA, we can make use of ATM to estimate topic's distribution of users and simultaneously update the mechanism to update topic's distribution of users since each author can own various documents.

3.2. *Homophily measure*

In this study, we estimate homophily between two agents based on their topic's probability distribution. If we consider a probability distribution as a vector, we can choose some distance measures related to the vector distance such as Euclidean distance, cosine similarity, Jaccard coefficient, etc. However, experimental results in our previous work[21] demonstrated that it is better if we choose distance measures related to the probability distribution such as Kullback–Leibler divergence, Jensen–Shannon divergence, Hellinger distance, etc. In this study, we chose Hellinger distance and Jensen–Shannon divergence to measure distance.

Let the probability distributions on k topics $P = (p_1, p_2, \ldots, p_k)$ and $Q = (q_1, q_2, \ldots, q_k)$ correspond to users u and v, respectively.

3.2.1. *Hellinger distance*

The Hellinger distance is given by

$$d_H(P,Q) = \frac{1}{\sqrt{2}}\sqrt{\sum_{i=1}^{k}(\sqrt{p_i} - \sqrt{q_i})^2}. \tag{7}$$

The Hellinger distance satisfies the inequality of $0 \leq d_H(P,Q) \leq 1$. This distance is a metric for measuring the deviation between two probability distributions. The distance is 0 when $P = Q$. When P and Q are disjoint, it shows the maximum distance of 1. The lower the value of the Hellinger distance, the smaller is the deviation between two probability distributions.

3.2.2. *Jensen–Shannon divergence*

The Jensen–Shannon divergence is given by

$$d_{\mathrm{JS}}(P,Q) = \frac{1}{2}\sum_{i=1}^{k} p_i \log\frac{2p_i}{p_i + q_i} + \frac{1}{2}\sum_{i=1}^{k} q_i \log\frac{2q_i}{p_i + q_i}. \tag{8}$$

The Jensen–Shannon divergence is a symmetrized and smoothed version of the Kullback–Leibler divergence, relative to Shannon's concept of uncertainty or "entropy" $H(P) = \sum_{i=1}^{n} p_i \ln p_i$.

For log base 2, the Jensen–Shannon divergence is bounded by 1: $0 \leq d_{\mathrm{JS}}(P,Q) \leq 1$. For log base e, or ln, which is commonly used in statistical thermodynamics, the upper bound is $\ln(2) : 0 \leq d_{\mathrm{JS}}(P,Q) \leq \ln(2)$.

3.2.3. *Homophily*

Since both Hellinger distance and Jensen–Shannon divergence are bounded by 1 and the lower the value of the Hellinger distance or Jensen–Shannon divergence, the smaller is the deviation between two probability distributions, therefore, we can define the homophily between two users u, v as

$$\mathrm{Homo}(u, v) = 1 - d(P, Q), \tag{9}$$

in which $d(P, Q)$ is $d_H(P, Q)$ if we use Hellinger distance or $d_{\mathrm{JS}}(P, Q)$ if we use Jensen–Shannon divergence.

4. Homophily Independent Cascade Model Based on Textual Information (Textual-Homo-IC)

In this section, we proposed an expanded model of independent cascade diffusion model, namely *Textual-Homo-IC*. In this model, we concentrate on exploiting the infected probability estimation based on homophily. *Textual-Homo-IC* is demonstrated in detail on both static and dynamic single-relational textual agent's networks. We present their steps in more detail in Algorithms 4.1 and 4.2, respectively.

4.1. *Random-IC on static single-relational textual agent's network*

In this subsection, we illustrate the IC model on a static network with infected probability based on uniform distribution, namely *Random-IC*. This model plays as a benchmark model for comparing the performance with *Textual-Homo-IC* model that we will propose at Sec. 4.2. At each step t where I^{newest} is the set of newly active nodes at time $t-1$, each $u \in I^{\mathrm{newest}}$ infects the inactive neighbors $v \in \eta^{\mathrm{out}}(u)$ with a probability $P(u, v)$ randomly. The propagation continues until no more infection can occur (see Algorithm 4.1).

Algorithm 4.1. Random-IC on static single-relational textual agent's network

```
Require: Single-relational textual agent's network $G = (V, E)$, $I_0$: seed set.
procedure RANDOM-IC-STATIC-SINGLERELATION-NETWORK($G, I_0$)
    $t \leftarrow 0, I^{\mathrm{total}} \leftarrow I_0, I^{\mathrm{newest}} \leftarrow I_0$
    while infection occurs do
        $t \leftarrow t + 1$; $I_t \leftarrow \emptyset$
        for $u \in I^{\mathrm{newest}}$ do
            $I_t(u) \leftarrow \{v^{\mathrm{inactive}} \in \eta^{\mathrm{out}}(u), p \leq q\}; p, q \sim U(0, 1)$
            $I_t \leftarrow I_t \cup I_t(u)$
        $I^{\mathrm{total}} \leftarrow I^{\mathrm{total}} \cup I_t$; $I^{\mathrm{newest}} \leftarrow I_t$
    return $I^{\mathrm{total}}$                                        ▷ Output
```

Algorithm 4.2 . Textual-Homo-IC on static single-relational textual agent's network

```
Require: Single-relational textual agent's network $G = (V, E)$, $I_0$: seed set.
procedure TEXTUAL-HOMO-IC-STATIC-SINGLERELATION-NETWORK($G, I_0$)
    $t \leftarrow 0, I^{\text{total}} \leftarrow I_0, I^{\text{newest}} \leftarrow I_0$
    while infection occurs do
        $t \leftarrow t + 1$; $I_t \leftarrow \emptyset$
        for $u \in I^{\text{newest}}$ do
            $I_t(u) \leftarrow \{v^{\text{inactive}} \in \eta^{\text{out}}(u), p \leq \text{Homo}(u, v)\}; p \sim U(0, 1)$
            $I_t \leftarrow I_t \cup I_t(u)$
        $I^{\text{total}} \leftarrow I^{\text{total}} \cup I_t$; $I^{\text{newest}} \leftarrow I_t$
    return $I^{\text{total}}$                                              ▷ Output
```

4.2. *Textual-Homo-IC on static single-relational textual agent's network*

Propagation mechanism of *Textual-Homo-IC* on static textual agent's network is similar to *Random-IC*, but the difference is that each active agent $u \in I^{\text{newest}}$ infects the inactive neighbors $v \in \eta^{\text{out}}(u)$ with a probability $P(u, v)$ equal to Homophily(u, v) instead of a random probability (see Algorithm 4.2).

4.3. *Textual-Homo-IC on dynamic single-relational textual agent's network*

Although IC model on the dynamic network has been researched,[22,23] the dynamic concept of a network has only been considered under the structure transformation while the activated probability from an active node to inactive another is always fixed during the spreading process. Therefore, we propose *Textual-Homo-IC* on a dynamic textual agent's network in which we can discover not only the variation of network's structure but also the agent's topics distribution. It can be said that the infected probability among agents can change over time because of their homophily transformation.

There is a resemblance in the propagation mechanism of *Textual-Homo-IC* on the dynamic network in comparison with the static network; however, in the spreading process at step $t \in C$, textual agent's network G will be updated as shown in Sec. 3 (see Algorithm 4.3).

4.4. *Experiments*

In this subsection, we implement the experiments to test our models *Textual-Homo-IC* on real-data networks. First, the goal of the experiment is to test the performance of *Textual-Homo-IC* on static single-relational textual agent's network. Moreover, we simulate experiments to demonstrate *Textual-Homo-IC* on dynamic single-relational textual agent's network. Steps of the experiment are demonstrated sequentially

Algorithm 4.3. Textual-Homo-IC on dynamic single-relational textual agent's network

```
Require: Single-relational textual agent's network $G = (V, E)$; $I_0$: seed set.
Require: $C = \{k_1, k_2, \ldots, k_n\}$, at step $k_i$ $G$ is updated; $n$: number of steps of
    diffusion.
procedure Textual-Homo-IC-Dynamic-SingleRelation-Network($G$, $I_0, C$)
    $t \leftarrow 0, I^{\text{total}} = I_0, I^{\text{newest}} = I_0$
    while $t < n$ do                                              ▷ $(n > \max\{C\})$
        $t \leftarrow t + 1; I_t = \emptyset$
        if $t \in C$ then:
            Update $G$; $I^{\text{newest}} = I^{\text{total}}$
        for $u \in I^{\text{newest}}$ do
            $I_t(u) \leftarrow \{v^{\text{inactive}} \in \eta^{\text{out}}(u), p \leq \text{Homo}(u, v)\}; p \sim U(0, 1)$
            $I_t \leftarrow I_t \cup I_t(u)$
        $I^{\text{total}} = I^{\text{total}} \cup I_t$; $I^{\text{newest}} = I_t$
    return $I^{\text{total}}$                                     ▷ Output
```

from collecting data, seting up experiments to evaluating the results based on baseline model.

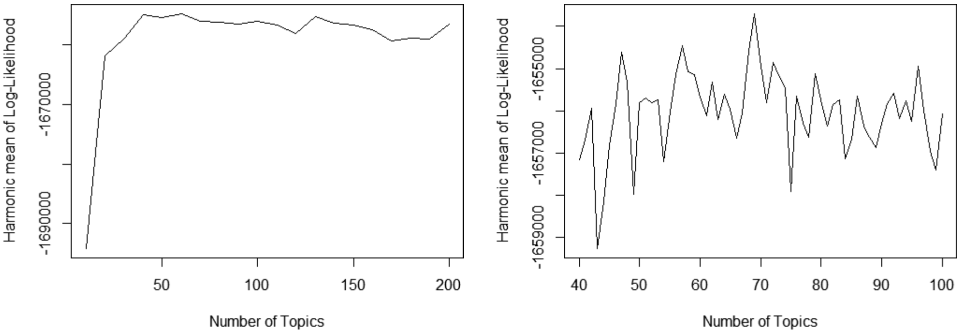
4.4.1. *Data collection*

The propose *Textual-Homo-IC* models have been tested on a well-known social network platform, Twitter, and co-author network. For Twitter network, we have aimed for 1,524 users in which links are the *follow* relations. We crawled over 100 tweets for each user and the textual data stretched from 2011 to April 2018. For co-author network, we have targeted the authors who have participated in Neural Information Processing Systems Conference (NIPS) from 2000 to 2012. The dataset contains 1,740 papers which are contributed by 2,479 scientists.

4.4.2. *Experimental setup*

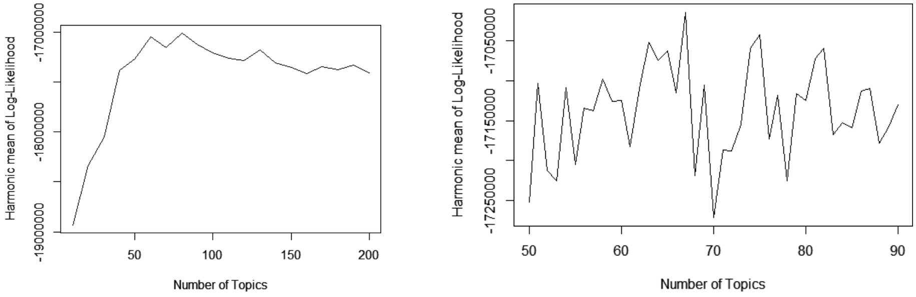
First, we defined the number of topics for the whole corpus based on the harmonic mean of log-likelihood (HLK).[24] We calculated HLK with the number of topics in the range [10, 200] with sequence 10. We realized that the best number of topics is in the range [40, 100] for Twitter network (Fig. 4(a)) and [50, 90] is in the range (Fig. 5(a)) for co-author network. Therefore, we ran HLK again with the sequence 1 and obtained the best as 69 for Twitter network (Fig. 4(b)) and as 67 for co-author network (Fig. 5(b)).

Textual-Homo-IC diffusion is implemented on the static Twitter network and co-author network. Textual agent's networks with single relation are constructed as



(a) No. of topics in the range [10, 200], seq. = 10. (b) No. of topics in the range [40, 100], seq. = 1.

Fig. 4. Log-likelihood for Twitter network.



(a) No. of topics in the range [10, 200], seq. = 10. (b) No. of topics in the range [50, 90], seq. = 1.

Fig. 5. Log-likelihood for co-author network.

shown in Sec. 3 in which there exist *follow* relation for Twitter network and *co-author* relation for co-author network. For each network, we implemented four experiments of *Textual-Homo-IC* with the combinations of two methods for estimating topic's distribution (LDA and ATM) and two kinds of distance measurements (Hellinger distance and Jensen–Shannon divergence). Besides, we also conducted *Random-IC* as a benchmark to compare the performance with *Textual-Homo-IC*.

To simulate *Textual-Homo-IC* on dynamic textual agent's network, we conducted experiments on the dynamic Twitter network and co-author network. For co-author network, we collected textual data between 2000 and 2009 for training the corpus and estimating the author's topic distribution using ATM. Textual agent's network is formed with the *co-author* relation. On the other hand, for Twitter network, textual data is gathered from 2011 to January 2018 for training the corpus. Unfortunately, it is impossible to get the exact date when a user starts to follow another on Twitter, including from the API or Twitter's Web interface. This leads to the inability to express the fluctuations in network structure with the *follow* relation. Therefore, we took into account textual agent's network with the relation R_{MTP} with $p_0 = 0.1$.

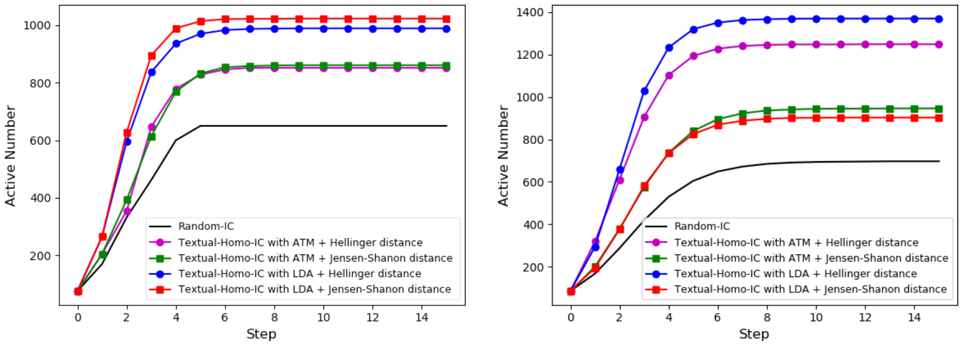
The diffusion process on dynamic network starts as soon as a textual agent's network is formed. For each kind of distance measurement, we implemented four experiments in which the first one is the propagation on textual agent's network without dynamics. The last investigations are those after every 5, 10, and 15 steps of diffusion when the textual agent's network will fluctuate once the *follow* mechanism exists presented in Sec. 3. Textual agent's networks will be updated three times corresponding to each month from February to April 2018 for the Twitter network and each year from 2010 to 2012 for the co-author network.

4.4.3. *Model evaluation*

To evaluate the performance of diffusion models, we can use the number of active nodes or the active percentage which are the standard metrics in information diffusion field.[22,23] In this research, we utilize the active number to evaluate the performance of spreading models. We compare the performance of proposed *Textual-Homo-IC* diffusion model with the baseline model (*Random-IC*).

4.4.4. *Results and discussion*

Comparison of Textual-Homo-IC and Random-IC diffusions. The results of *Textual-Homo-IC* on static textual agent's networks are shown in Fig. 6. For both networks, we can see that the active number of *Textual-Homo-IC* is always greater than *Random-IC* in the four cases which are the combinations of two methods of topic modeling and two distance measurements. First, in Twitter network (Fig. 6(a)), the number of active agents reaches approximately 650 for *Random-IC* diffusion while *Textual-Homo-IC* attains about 862 for both cases where ATM combines with the two distances. Particularly, the cases where *Textual-Homo-IC*



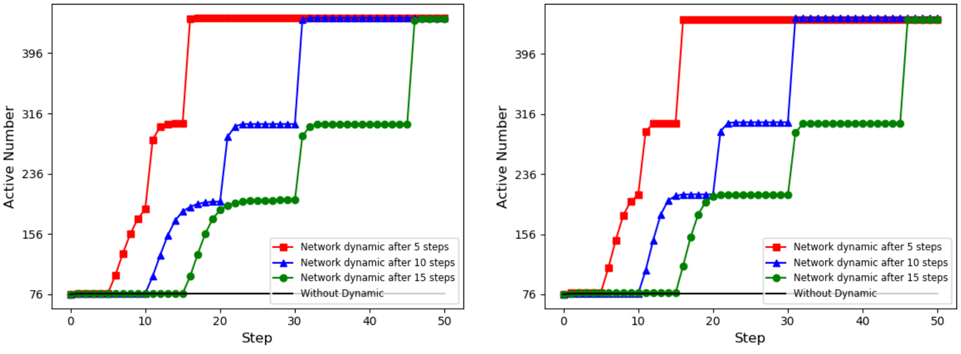
(a) Twitter network.

(b) Co-author network.

Fig. 6. Textual-Homo-IC diffusion on static networks.

incorporates LDA with Hellinger distance and Jensen–Shannon divergence obtain the higher numbers of active agents in comparison with the cases utilizing ATM, of about 989 and 1,023 active agents, respectively. On the other hand, in co-author network (Fig. 6(b)), the number of active agents reaches approximately 700 for *Random-IC* diffusion while *Textual-Homo-IC* attains about 903 for the case using LDA combined with Jensen–Shannon divergence. Besides, 946 active agents are reached by making ATM and Jensen–Shannon divergence to collaborate. In addition, *Textual-Homo-IC* with ATM and Hellinger distance obtains approximately 1,248 active agents while the highest number belongs to *Textual-Homo-IC* with LDA and Hellinger distance, of around 1,369 active agents. In summary, we can conclude that *Textual-Homo-IC* diffusion outperforms when compared with Random-IC.

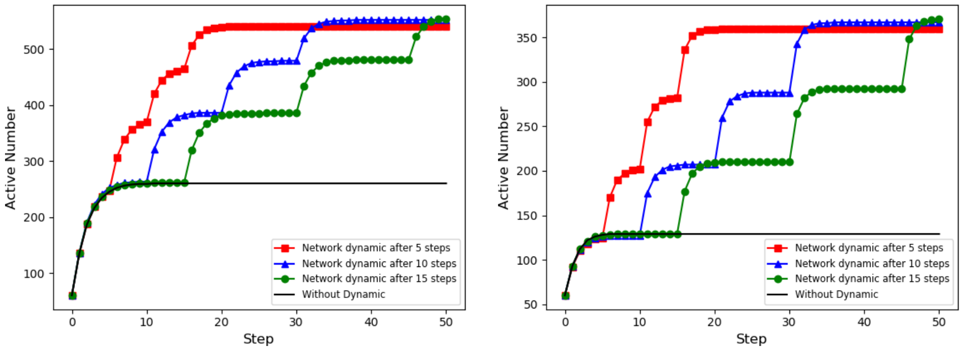
Textual-Homo-IC diffusion on dynamic textual agent's network. Results are shown in Figs. 7 and 8 which illustrate *Textual-Homo-IC* diffusions on the dynamic Twitter network and co-author network, respectively. For Twitter network, there is only one agent that can be activated from the seed set on the static network for both cases of distance measurements. The reason is the number of connections with R_{MTP} in the initial stage is too low for diffusion. However, if there exists network's transformation in the next three stages with the arrival of many new connections, there is a significant increase in the active number. For co-author network, *Textual-Homo-IC* on a static network reaches a steady state from the 12th step and seventh step onwards with using Hellinger distance and Jensen–Shannon divergence, respectively. However, if there is network's fluctuation in the next three stages, the active number increases significantly. In short, we can conclude that the propagation process without dynamics of network reached and maintained the steady state while there is a significant transformation in the active number if the textual agent's network has fluctuation in the diffusion process.



(a) ATM and Hellinger distance.

(b) ATM and Jensen–Shannon divergence.

Fig. 7. Textual-Homo-IC diffusion on dynamic Twitter network.



(a) ATM and Hellinger distance.

(b) ATM and Jensen–Shannon divergence.

Fig. 8. Textual-Homo-IC diffusion on dynamic co-author network.

5. Homophily Pretopological Cascade Model for Information Diffusion Based on Textual Information (Textual-Homo-PCM)

In Sec. 4, we introduced the propagation process on a social network related to textual content through *Textual-Homo-IC* model. Nevertheless, we just consider spreading process on a single-relational network where the classical way for defining the neighborhood set of active set A is a union of neighbors of each element in A. In fact, the diffusion process can take place on a network with numerous different relations and perhaps that the neighborhood determination may become more complicated. Therefore, in this section, we will exploit dissemination on multi-relational textual agent's network. The question is how to define neighborhood set of an active set on the multi-relational network.

In our previous research,[14] we have proposed PCM as a general cascade diffusion model that can take place in more complex networks such as multi-relational networks or stochastic graphs. In PCM, stochastic pretopology is used to capture random neighborhoods set based on stochastic pseudo-closure function.

Pretopology[15] is a theory that generalizes for both topology and graph theories and is commonly used to model complex propagation phenomena. A usual pseudo-closure function in pretopology is usually defined for neighborhood aggregation. In addition to modeling the dynamics of structural phenomena by pretopology theory, stochastic aspects that affect phenomena also were considered. Therefore, stochastic pretopology was introduced[14,15] where there are different approaches for building stochastic pseudo-closure function by combining pretopology and finite random set theory.

In this research, we would like to present the PCM on specific network related to textual information that we called textual agent's network. Therefore, we propose Textual-Homo-PCM which is PCM on textual agent's network with multi-relations. Particularly, we introduce more strong and weak pseudo-closure functions built from pretopology to be utilized to determine the neighborhood set with stochastic pseudo-closure from PCM.

Therefore, the contributions of this section are as follows:

- We present a method to build three kinds of pseudo-closure function based on the family of relationships.
- We propose an expanded model of PCM which will be presented on multi-relational textual agent's network, namely *Textual-Homo-PCM*, in which *strong*, *weak* pseudo-closure functions built from pretopology and *stochastic* pseudo-closure from stochastic pretopology are utilized to determine the neighborhood set.
- We conduct a small experiment with two small datasets to illustrate our approach.

5.1. *Building pseudo-closure functions*

In Sec. 3, we constructed a textual agent's network $G(V, (R_i)_{i=1,\dots,n}, (E_i)_{i=1,\dots,n})$. Let us consider a family of binary relations between agents $(R_i)_{i=1,\dots,n}$. For each relation R_i, we can define pretopological structure by considering the following subset: $\forall\, i = 1, 2, \dots, n$, $\forall\, x \in X$, $V_i(x)$ is defined by

$$V_i(x) = \{y \in X \mid x\, R_i\, y\}.$$

Then, the strong pseudo-closure $a_s(\cdot)$ is defined by

$$a_s(A) = \{x \in X \mid \forall\, i = 1, 2, \dots, n, V_i(x) \cap A \neq \emptyset\}, \quad \forall\, A \subset X. \tag{10}$$

Similarly, we can define weak pretopology from $a_w(\cdot)$ by using the union operator:

$$a_w(A) = \{x \in X \mid \exists\, i = 1, 2, \dots, n, V_i(x) \cap A \neq \emptyset\}, \quad \forall\, A \subset X. \tag{11}$$

To building stochastic pseudo-closure function, let us define a *random relation* R: $\Omega \to L$ as a random variable:

$$P(R(\omega) = R_i) = p_i; \quad p_i \geq 0; \quad \sum_{i=1}^{n} p_i = 1.$$

For each $x \in V$, we can build a random set of neighbors of x with the random relation R:

$$\Gamma_{R(\omega)}(x) = \{y \in V \mid x\, R(\omega)\, y\}.$$

We can define a *stochastic pseudo-closure* $a(\cdot, \cdot)$ as

$$\forall\, A \in \mathcal{P}(V), \quad a(\omega, A) = \{x \in V \mid \Gamma_{R(\omega)}(x) \cap A \neq \emptyset\}. \tag{12}$$

There are three pseudo-closure functions in Eqs. (10)–(12) which are used to determine the neighborhood set for the spreading process.

5.2. *Textual-Homo-PCM*

In our approach, from the textual agent's network $G(V, (R_i)_{i=1,\dots n}, (E_i)_{i=1,\dots n})$ constructed at Sec. 3, we can build a model namely *Textual-Homo-PCM* as an information diffusion model that integrated pretopology, stochastic pretopology, topic modeling, and cascade model via three steps:

> Step 1: Define pseudo-closure functions based on the family of relations $(R_i)_{i=1,\ldots,n}$. In this step, we will define three kinds of pseudo-closure functions presented in the previous subsection: *strong pseudo-closure* $a_s(\cdot)$ in Eq. (10), *weak pseudo-closure* $a_w(\cdot)$ in Eq. (11), and *stochastic pseudo-closure* $a(w, \cdot)$ in Eq. (12).
>
> Step 2: Define the set of neighbors $N(I_{t-1})$ of active set I_{t-1} at step $t-1$ based on the pseudo-closure functions built from step 1.
>
> Step 3: Each inactive node $v \in N(I_{t-1}) \backslash I_{t-1}$ will be influenced by I_{t-1} to become an active node based on the threshold rule: each inactive node $v \in N(I_{t-1}) \backslash A_{t-1}$ will be influenced by I_{t-1} to become an active node if the sum of all homophily from all active nodes in I^{t-1} to it is greater than the activation threshold θ_v.

We can see that with *strong pseudo-closure* function in Eq. (10), we can determine the neighborhood set of active set I^{total} by $a_s(A)$ where $A = I^{\text{total}}$. A node $v \in V$ will become the neighbor of active set I^{total} when it has all relations $(R_i)_{i=1,\ldots,n}$ with I^{total}. Therefore, a strong pseudo-closure function will narrow the scope of neighborhood set, leading to slower propagation process. In contrast, *weak pseudo-closure* function in Eq. (11) expands the area of neighborhood set since a node $v \in V$ will become the neighbor of I^{total} when there is just one relationship with I^{total}. This can promote the spreading process faster. Moreover, perhaps that in reality information diffusion process from one person to another takes place with a stochastic relationship at each time t. Therefore, we use $a(\omega, \cdot)$ from Eq. (12) to determine a random neighborhood set of active set with a random relation R. The *Textual-Homo-PCM* algorithm is presented in Algorithm 5.1.

Algorithm 5.1. Textual-Homo-PCM

```
Require: Multi-relational textual agent's network $G(V, (R_i)_{i=1,\ldots,n}, (E_i)_{i=1,\ldots,n})$.
Require: $I_0$: seed set.
procedure TEXTUAL-HOMO-PCM$(G, I_0)$
    $t \leftarrow 0, I^{\text{total}} \leftarrow I_0$
    while infection occurs do
        $t \leftarrow t + 1$; $I_t \leftarrow \emptyset$
        $N_t \leftarrow a_s(I^{\text{total}}), a_w(I^{\text{total}})$ or $a(\omega, I^{\text{total}})$
        for $v \in N_t - I^{\text{total}}$ do
            if $\sum_{u \in I_{\text{total}}} \text{Homo}(u, v) > \theta_v$ then
                $I_t \leftarrow I_t \cup \{v\}$
        $I^{\text{total}} \leftarrow I^{\text{total}} \cup I_t$;
    return $I^{\text{total}}$                                        ▷ Output
```

5.3. *Experiments*

5.3.1. *Data collection*

The proposed *Textual-Homo-PCM* model has been simulated on a well-known social network platform, Twitter, and the co-author network. For Twitter, we have aimed for 100 users and crawled over 100 tweets for each user. For the co-author network, we have targeted 100 authors who have participated in Neural Information Processing Systems Conference from 2000 to 2012.

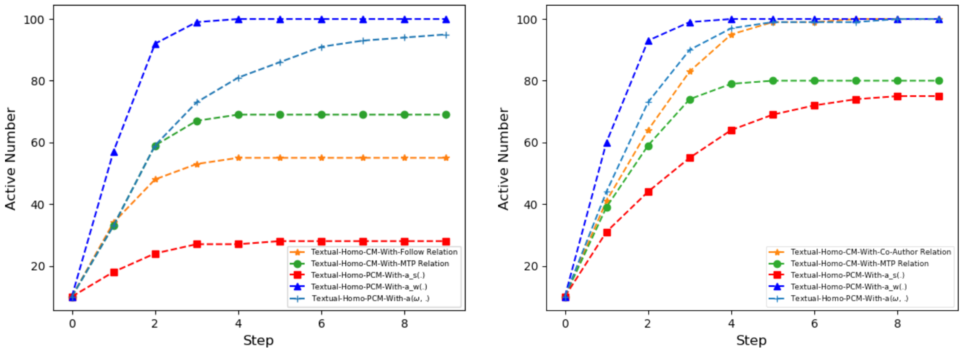
5.3.2. *Experimental setup*

Textual agent's networks are constructed as shown in Sec. 3 in which there are two relations R_1 and R_2 between agents. We use ATM to estimate topic's probability distribution of agents and Hellinger distance for homophily measurement. For the co-author network, we consider R_1 as *co-author* relation and the relation R_{MTP} for R_2. For Twitter network, *follow* and R_{MTP} relation correspond to R_1 and R_2, respectively. We consider R_{MTP} with probability threshold $p_0 = 0.35$.

The diffusion process will start with a seed set $|A_0| = 10$. One hundred random samples are used for seed set and the diffusion process is executed 100 times for each seed set. The spreading follows two major steps in which the first step is to capture a set of neighborhoods through pseudo-closure function and apply the threshold rule based on homophily under the second step. We estimate the activation threshold for each node $v[\theta(v)]$ from the product of average homophily on the whole network and the average degree of v. For random neighborhood set $a(\omega, \cdot)$, we defined random index distribution ω corresponding to probability distribution $[0.5, 0.5]$. In addition to experiments of *Textual-Homo-PCM*, we also implement cascade model on single-relational networks (*Textual-Homo-CM*) to compare the results.

5.3.3. *Results and discussion*

Experimental results of *Textual-Homo-PCM* are shown in Fig. 9. We can see that cascade diffusions on single-relational textual agent's networks (*Textual-Homo-CM*) and *Textual-Homo-PCM* with $a(\omega, \cdot)$ always achieve lower results than *Textual-Homo-PCM* with $a_w(\cdot)$ while they reach higher performance in comparison with *Textual-Homo-PCM* with $a_s(\cdot)$. *Textual-Homo-PCM* with $a_s(\cdot)$ always reaches lowest performance while the highest result belongs to *Textual-Homo-PCM* with $a_w(\cdot)$. For Twitter network, *Textual-Homo-CM* obtained 55 and 69 active nodes for the network with *follow* and *major topic* relations, respectively. Besides, *Textual-Homo-PCM* with $a_s(\cdot)$ just reached half the value of active number of *Textual-Homo-CM* with *follow* relation while for *Textual-Homo-PCM* with $a_w(\cdot)$ and $a(\omega, \cdot)$ nearly all nodes are active in the network. Similarity, for co-author network, *Textual-Homo-PCM* with $a_s(\cdot)$ obtained the lowest result, approximately 75 active nodes. Next, *Textual-Homo-CM* with *major topic* relation reached slightly higher when compared with *Textual-Homo-PCM* with $a_s(\cdot)$, of around 80 active agents. Finally, the highest



(a) Twitter network.

(b) Co-author network.

Fig. 9. Textual-Homo-PCM diffusion.

results belonged to *Textual-Homo-CM* with *co-author* and *Textual-Homo-PCM* with $a_w(\cdot)$ and $a(\omega, \cdot)$. These results can be explained by the criteria for neighborhood set determination from pseudo-closure functions that we described in Sec. 5.2. In short, we can apply pretopology theory and stochastic pretopology to simulate diffusion process on multi-relational network. There are various ways to determine the neighborhood set through different pseudo-closure functions. Obviously, the performance of propagation process depends on the criteria combination for neighborhood determination from the pseudo-closure function.

6. Conclusion

In this research, we propose two expanded models of IC diffusion model and PCM, namely *Textual-Homo-IC* and *Textual-Homo-PCM*, respectively. These models are applied on network related to textual information, namely *textual agent's network*. In *textual agent's network*, each agent corresponds to one node which has specific properties including ID, neighbors, and topic's probability distribution. Topic modeling is utilized to estimate the topic's distribution of agents from the textual content. In *Textual-Homo-IC*, we just take into account the spreading process on single-relational textual agent's network and concentrate on estimating the probability of an active node infecting another inactive one based on their homophily which is measured from their topic's distribution. *Textual-Homo-IC* has revealed details on both static and dynamic single-relational textual agent's networks. Experimental results demonstrated that the effectiveness of *Textual-Homo-IC* on the static network outperforms *Random-IC*. In addition, experiments also illustrated the fluctuation of active number on the dynamic agent's network instead of reaching and remaining in a steady state in a static network. On the other hand, we exploit the spreading process on multi-relational textual agent's network through *Textual-Homo-PCM*. In *Textual-Homo-PCM*, pretopology and stochastic pretopology theory are applied to define the neighborhood set for spreading by the pseudo-closure function. Experimental results illustrated that we can combine multi-criteria based on relations between agents to determinate the neighborhood set, leading to expected diffusion results. In future works, we will conduct experiments on other large-scale networks and compare our models with more other baseline models.

References

1. M. Granovetter, Threshold models of collective behavior, *Am. J. Soc.* **83**(6) (1978) 1420–1443.
2. J. Goldenberg, B. Libai and E. Muller, Talk of the network: A complex systems look at the underlying process of word-of-mouth, *Mark. Lett.* **12**(3) (2001) 211–223.
3. D. Kempe, J. Kleinberg and É. Tardos, Maximizing the spread of influence through a social network, in *Proc. Ninth ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining* (2003), pp. 137–146.
4. D. Kempe, J. M. Kleinberg and É. Tardos, Influential nodes in a diffusion model for social networks, in *Automata, Languages and Programming: ICALP 2005*, Lecture Notes in Computer Science, Vol. 3580 (Springer, Berlin, 2005), pp. 1127–1138.
5. M. Kimura and K. Saito, Tractable models for information diffusion in social networks, in *PKDD 2006: Knowledge Discovery in Databases*, Lecture Notes in Computer Science, Vol. 4213 (Springer, Berlin, 2006), pp. 259–271.
6. P. F. Lazarsfeld and R. K. Merton, Friendship as a social process: A substantive and methodological analysis, *Freedom Control Mod. Soc.* **18**(1) (1954) 18–66.
7. M. McPherson, L. Smith-Lovin and J. M. Cook, Birds of a feather: Homophily in social networks, *Annu. Rev. Sociol.* **27**(1) (2001) 415–444.
8. D. Laniado, Y. Volkovich, K. Kappler and A. Kaltenbrunner, Gender homophily in online dyadic and triadic relationships, *EPJ Data Sci.* **5**(1) (2016) 19.
9. L. M. Aiello, A. Barrat, R. Schifanella, C. Cattuto, B. Markines and F. Menczer, Friendship prediction and homophily in social media, *ACM Trans. Web* **6**(2) (2012) 9:1–9:33.
10. F. M. Cardoso, S. Meloni, A. Santanch and Y. Moreno, Topical homophily in online social systems, arXiv:1707.06525 [physics.soc-ph].
11. G. Salton and C. Buckley, Term-weighting approaches in automatic text retrieval, *Inf. Process. Manage.* **24**(5) (1988) 513–523.
12. D. M. Blei, A. Y. Ng and M. Jordan, Latent Dirichlet allocation, in *J. Mach. Learn. Res.* **3** (2001) 601–608.
13. M. Rosen-Zvi, T. L. Griffiths, M. Steyvers and P. Smyth, The author-topic model for authors and documents. in *Proc. 20th Conf. Uncertainty in Artificial Intelligence* (2004), pp. 487–494.
14. Q. V. Bui, S. B. Amor and M. Bui, Stochastic pretopology as a tool for topological analysis of complex systems, in *ACIIDS 2018: Intelligent Information and Database Systems*, Lecture Notes in Computer Science, Vol. 10752 (Springer, Cham, 2018), pp. 102–111.
15. Z. Belmandt, *Basics of Pretopology* (Hermann, 2011).
16. M. Niazi and A. Hussain, Agent-based computing from multi-agent systems to agent-based models: A visual survey, *Scientometrics* **89**(2) (2011) 479.
17. M. D. Hoffman, D. M. Blei, C. Wang and J. W. Paisley, Stochastic variational inference, arXiv:1206.7051 [stat.ML].

18. M. Lamure, S. Bonnevay, M. Bui and S. B. Amor, A stochastic and pretopological modeling aerial pollution of an urban area, *Stud. Inf. Univ.* **7**(3) (2009) 410–426.
19. C. Basileu, S. B. Amor, M. Bui and M. Lamure, Prétopologie stochastique et réseaux complexes, *Stud. Inf. Univ.* **10**(2) (2012) 73–138.
20. H. T. Nguyen, *An Introduction to Random Sets* (CRC Press, 2006).
21. Q. V. Bui, K. Sayadi, S. B. Amor and M. Bui, Combining latent Dirichlet allocation and K-means for documents clustering: Effect of probabilistic based distance measures, in *Intelligent Information and Database Systems: ACIIDS 2017*, Lecture Notes in Computer Science, Vol. 10191 (Springer, Cham, 2017), pp. 248–257.
22. N. T. Gayraud, E. Pitoura and P. Tsaparas, Diffusion maximization in evolving social networks, in *Proc. 2015 ACM Conf. Online Social Networks* (2015), pp. 125–135.
23. H. Zhuang, Y. Sun, J. Tang, J. Zhang and X. Sun, Influence maximization in dynamic social networks, in *Proc. 2013 IEEE 13th Int. Conf. Data Mining* (2013), pp. 1313–1318.
24. W. Buntine, Estimating likelihoods for topic models, in *Proc. 1st Asian Conf. Machine Learning: Advances in Machine Learning* (2009), pp. 51–64.
25. T. K. T. Ho, Q. V. Bui and M. Bui, Homophily independent cascade diffusion model based on textual information, in *Computational Collective Intelligence*, eds. N. T. Nguyen, E. Pimenidis, Z. Khan and B. Trawinski, Lecture Notes in Computer Science, Vol. 11055 (Springer International Publishing, 2018), pp. 134–145.