


SCIENTIFIC REPORTS



OPEN

Information Spread and Topic Diffusion in Heterogeneous Information Networks

Soheila Molaei¹, Sama Babaei¹, Mostafa Salehi^{1,2}  & Mahdi Jalili³

Diffusion of information in complex networks largely depends on the network structure. Recent studies have mainly addressed information diffusion in homogeneous networks where there is only a single type of nodes and edges. However, some real-world networks consist of heterogeneous types of nodes and edges. In this manuscript, we model information diffusion in heterogeneous information networks, and use interactions of different meta-paths to predict the diffusion process. A meta-path is a path between nodes across different layers of a heterogeneous network. As its most important feature the proposed method is capable of determining the influence of all meta-paths on the diffusion process. A conditional probability is used assuming interdependent relations between the nodes to calculate the activation probability of each node. As independent cascade models, we consider linear threshold and independent cascade models. Applying the proposed method on two real heterogeneous networks reveals its effectiveness and superior performance over state-of-the-art methods.

Many real systems can be modeled by networks where a number of individuals interact through a connection graph. Examples of networked systems include the Internet, World Wide Web, the human brain, power grids, online social networks, transportation and water distribution networks. Various dynamical phenomena have been studied on complex networks including synchronisation¹, consensus², opinion formation^{3,4} and information spread⁵. Network topology has the major role in how such dynamical processes evolve on networks. Certain topologies might facilitate synchronisation or information spread, while some other network structures might disrupt such activities^{6,7}.

Information diffusion is one of the widely studied dynamical processes on networks, which has potential applications in fields. Information such as a news, innovation, virus or malware, starts from a set of seed nodes and propagates throughout the network. There is a rich literature on information diffusion on complex networks, where different models and their interplay with network topology have been studied¹. Previous research works have mainly considered homogeneous networks. An information network $G = (V, E)$ with V as the set of nodes and E as the set of edges, is a homogeneous network if the edges and nodes are of the same type. Networks with nodes and/or edges from more than one type are called heterogeneous networks^{8–10}. For example, in DBLP network, which is a major bibliography provider in computer science, the nodes are authors, papers, venues (journals/conferences). In this network, edges can be author-author relationship when they co-author a paper, or author-venue relationship when an author participates in a conference.

Here we model information diffusion or more specifically topic diffusion in heterogeneous information networks. To this end, we use the concept of meta-path, which is defined in heterogeneous networks. A meta-path P is a path defined over the general schema of the network $T_G = (A, R)$, where A and R denote the nodes and their relations, respectively. The meta-path is denoted by $A_1 \xrightarrow{R_1} A_2 \xrightarrow{R_2} \dots \xrightarrow{R_l} A_{l+1}$, where l is an index indicating the corresponding meta-path. The aggregated relationship is obtained as $R = R_1 \circ R_2 \circ \dots \circ R_l$ between different types of nodes A_1 to A_{l+1} , where \circ is the composition operator. For instance, in DBLP network, each of the author-paper-author and author-conference-author relations is considered to be an individual meta-path. Figure 1 is an example of “Data mining” topic propagation that authors can be connected to one another through different meta-paths in DBLP network.

Recently, much attention has been given to employing non-homogeneous networks in classification and ranking tasks. For instance, sentiment classification of product reviews using heterogeneous networks was addressed

¹Faculty of New Sciences and Technologies, University of Tehran, Tehran, Iran. ²School of Computer Science, Institute for Research in Fundamental Science (IPM), Tehran, Iran. ³School of Engineering RMIT University, Melbourne, Australia. Correspondence and requests for materials should be addressed to M.S. (email: mostafa_salehi@ut.ac.ir)

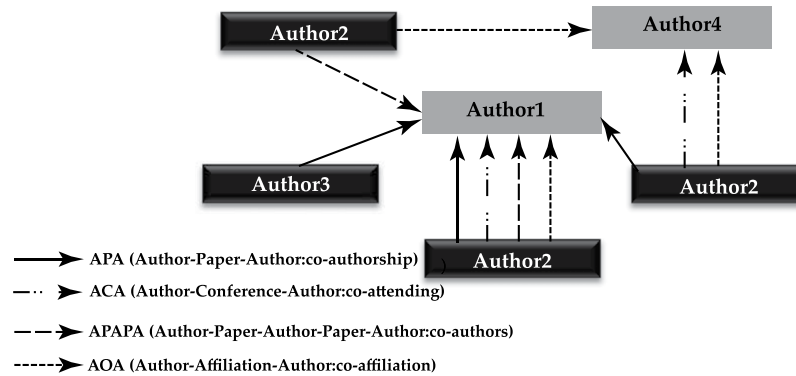


Figure 1. An Example of a heterogeneous network, where “Data mining” topic propagates along different types of relationships among authors. Black nodes are authors who have already pursued the topic, while gray nodes represent authors that may pursue the topic at the next timestamp.

by Zhou *et al.*¹¹. In this process, a heterogeneous network connects the users, products, and words, based on which the learning process is conducted using sentiment classification. In this regard, Zhou *et al.*¹¹ proposed a co-ranking method which classifies the authors and documents separately based on random walks. Angelova *et al.*¹² presented a new classification method for the DBLP heterogeneous network. Mining of heterogeneous networks was addressed in a number of studies^{13–15}. For example, Boccaletti and others¹⁶ studied mining of homogeneous information networks through their decomposition into multiple homogeneous networks. The idea of citation recommendation using mining in heterogeneous networks was proposed by Liu *et al.*¹⁷. Heterogeneous networks have also been employed in healthcare. Some papers^{18–20} focused on epidemic spreading on heterogeneous networks. Considering an epidemic threshold, Wang and Dai²¹ addressed virus spreading in heterogeneous networks based on the well-known susceptible-infected-susceptible model. Moreover, it was shown by Yang *et al.*²² that by considering heterogeneity between people, a heterogeneous network is created which is resistant against epidemic spread of virus. Epidemic spreading is important issue that was considered in other networks like time-varying networks²³ and adaptive network²⁴. Nadini *et al.*²³ used SIR and SIS models and investigated effects of modular and temporal connectivity patterns on epidemic spreading.

Link prediction in heterogeneous networks has also been addressed. Shakibian and Moghadam Charkari²⁵ used meta-paths for prediction and Jalili and Orouskhani²⁶ formulated drug response prediction as a link prediction problem using kernelised Bayesian multitask learning algorithm. Some works have considered information diffusion on these networks. Sermpetis and his colleagues²⁷ used degree distribution for the process of information diffusion assuming that diffusion takes place between two nodes at random times. Zhou and Liu²⁸ presented a social influence based clustering framework has been presented for analyzing heterogeneous information networks. Moreover, a heterogeneous network model was proposed for new product diffusion in two stages by Li and Jin²⁹; the first stage is transition of information concerning new products to customers through advertisement, and the second stage is changing customer priorities through persuasive advertisements.

As another definition, heterogeneous networks are referred to as multilayer networks, where the nodes and/or edges can be of different types. In many studies in this field, the concept of heterogeneous networks has been used to present a different definition for the infrastructure networks, based on which the concepts of diffusion are explained. Multilayer networks with all nodes from the same type are often called multiplex networks; a number of works have considered link prediction problem in multiplex networks^{16,26}.

Some works have studied topic diffusion in heterogeneous networks. The concept of similarity based on meta-paths (known as Pathsim), between each two nodes was utilised and predictions were made by generalising the Linear threshold (LT) model by Gui and *et al.*³⁰. Pathsim was considered as a weight between each two nodes in this method through which predictions were conducted^{31,32}. In our proposed method, each meta-path instance is considered as a path by considering different meta-paths, and the conditional probability model is used to calculate the activation probability of each node. Also, two different diffusion models are used including Independent Cascade (IC) and LT. In these models, first all nodes are considered to be inactive. Then, an initial set of seed nodes are activated and LT/IC is used to activate the subsequent nodes. In IC model, an inactive node is activated under the influence of the active node with the highest probability of influence³³. In this model, a probability is assigned to each active node for activating its neighbors; the probability of activation of node w triggered by node v is denoted as $P(v|w)$. Every newly-activated node v attempts to trigger its inactive neighbors. If successfully triggered, node w is activated in the next step and triggers its inactive neighbors. Once a node is activated, it has a single chance to independently influence each of its neighbors. In LT mode, each inactive node is activated if the portion of its activated neighbors is more than a threshold $\theta \in [0, 1]$ ³⁴. Indeed, an inactive node is activated if and only if the total weight of all its activated neighbors exceeds a given threshold θ_u , as equation (1).

$$\sum_{v \in \varepsilon_u} W_{u,v} \geq \theta_u \quad (1)$$

where ε_u is the active neighbors of node u and $W_{u,v}$ represents the weight of the link between nodes u and v . Watts³⁵ studied the role of threshold values and network structure in the information diffusion. Gui *et al.*³⁰

Dataset	Authors	Papers
DBLP	215222	105372
PubMed	1219686	459726

Table 1. Information of DBLP and PubMed datasets used in this work.

proposed a model called Multi-Relational Linear Threshold Model - Relation Level Aggregation (MLTM-R), which studied how LT model behaves in heterogeneous networks. Our proposed model is compared to this model. The major contributions of this study are:

1. We propose two novel topic diffusion models in heterogeneous networks considering different meta-paths, meaning that the influence of each relation is individually learned.
2. The dependency of active nodes to inactive ones is considered and conditional probability is employed to obtain the possibility of activation of each inactive node.
3. Two frequently used models (LT and IC) are studied in heterogeneous networks and their behavior is compared in two real datasets. We show that IC model has more accurate answer than LT model in properly modeling topic diffusion in heterogeneous networks.

Methods

This study incorporates conditional probability for calculating the activation probability of inactive nodes by neighboring active nodes. This is in fact known as information propagation probability which defines the probability that an active node activates an inactive neighbor. This propagation probability is calculated considering meta-paths and using Bayesian framework. It is assumed that inactive nodes are dependent on the active ones. IC and LT models are employed for the process of information distribution. The stages involved in the proposed method are briefly presented in algorithm 1 with every stage being explained separately in the following subsections.

Datasets. A time-stamp of a year is defined for both datasets, based on which the training set and the test set are created as explained in the followings:

- **DBLP (computer science bibliography)**³⁶: Objects indicate authors in this network. Different meta-path such as APA (Author-Paper-Author), ACA (Author-Conference-Author), APAPA (Author-Paper-Author-Paper-Author), and ACACA (Author-Conference-Author-Conference-Author) are considered. Different topics are extracted from this dataset, and information diffusion about a specific topic is investigated. This dataset include information from 1954 to 2016.
- **PubMed Dataset**^{37,38}: In this network, the authors are represented by objects and meta-paths APA and APAPA are used. The dataset consists of information from 1950 to 2013. Information of both datasets is given in Table 1.

Evaluation criteria. All nodes with published papers on our particular topic of interest are tagged as active and the rest as inactive. Assuming the nodes to be predicted at time t , the training and test sets are considered as follows:

Training set: Those within the time period from $t - 4$ to $t - 2$ are considered as the training set.

Test set: Those within the time period from $t - 1$ to t are considered as the test set. Additionally, the nodes tagged as active up to the time $t - 2$ are considered as the seed nodes that are activated initially in the start of the diffusion process.

We use Precision and Recall, F-score, and Recall criteria to assess the performance. These metrics are defined as follows.

$$\text{Precision} = \frac{TP}{TP + FP}, \text{Recall} = \frac{TP}{TP + FN}, \text{F - Score} = \frac{2 * (\text{Precision} * \text{Recall})}{\text{Precision} + \text{Recall}} \quad (2)$$

where True Positive (TP) is the active nodes that are correctly tagged as active by the algorithm, True Negative (TN) is the inactive nodes that are correctly tagged as inactive by the algorithm, False Positive (FP) is the active nodes that are falsely tagged as inactive by the algorithm, and False Negative (FN) is the inactive nodes that are falsely tagged as active by the algorithm.

In IC model, let $S_t \subseteq V$ be the set of nodes that are activated at step $t \geq 0$, with $S_0 = S$. At step $t + 1$, every node $u \in S_t$ may activate its out-neighbors $v \in V$ with a propagation probability of $P(v|u)$. One should also consider the activation threshold for LT model. We study how the diffusion process depends on the threshold value. Initially, the optimal threshold limit is required to be calculated from the training set in order to obtain the evaluation criteria according to the third step of the algorithm 1. Figure 2 shows the F-scores as a function of the threshold value when considering diffusion of the selected topics in DBLP dataset. As it is seen, one can often obtain an optimal value for the threshold for which the F-score is the highest. Note that F-score scales in the range $[0, 1]$, where 1 indicates the best performance. This optimal threshold varies across different topics, which indicates that different topics have different propagation mechanisms in this dataset. The obtained optimal threshold value is

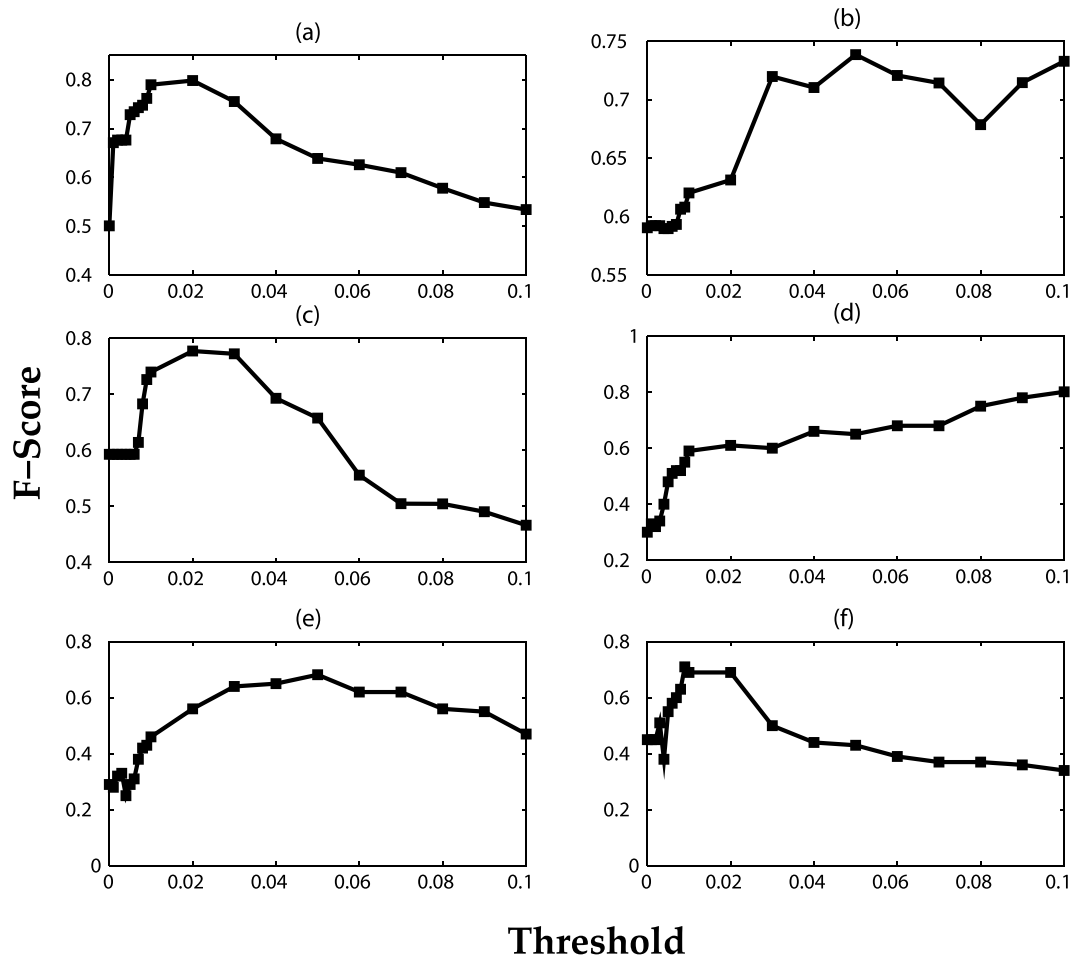


Figure 2. F-Score as a function of the threshold of LT model in DBLP dataset for selected topics. The figure shows F-score for topics (a) Data Mining, (b) Machine Learning, (c) Social Networks, (d) Healthcare, (e) DNA and (f) Infectious Disease. The optimal threshold for each topic is the one with the highest F-score.

then applied to the test set to assess the performance. We also use recall measure to obtain the optimal threshold and the results are similar to those obtained based on F-score (results not shown here).

Calculating propagation probability of Nodes. Propagation probabilities for all edges and nodes are calculated in this stage. In order to calculate the activation probability of each node according to its neighboring nodes, the influence probabilities of each node and edge are calculated considering meta-paths.

Edge Propagation Probability: In heterogeneous networks, different routes are available for meta-paths. Hence, for every pair of nodes v_1 and v_2 in meta-path k , the edge probability is equal to the number of path instances between the two nodes divided by all the existing path instances between them, as shown in equation 3.

Algorithm 1. Heterogeneous Probability Model (HPM).

Input : Datasets

Output : Propagation Probability of each inactive node; F-Score;

$\%P^k$ = Propagation Probability in meta-path k

$\%P$ = Final Propagation Probability

$\%a_k^{new}$ = Weight for meta-path k

$\%E_{v_i}$ = Active neighbors of node v_i

$\%n$ = The Number of active neighbors of node v_i

$\%nei_v$ = Neighbors of node v

$\%T_0$ = Initial time-stamp

$\%T_f$ = Final time-stamp

$\%Tr$ = Threshold

$\%n_{v_2 \rightarrow v_1}^k$ = Number of path instances in meta-path between node v_1 and v_2

$\%\phi$ = Learning step

1. Calculate Probability for nodes:
 - Find P for each pair of nodes v_1 and v_2

$$P^k(v_1|v_2) = \frac{P^k(v_1, v_2)}{P^k(v_2)} = \frac{n_{v_2 \rightarrow v_1}^k}{\sum_{r \in nei_v} n_{v_2 \rightarrow r}^k}$$

2. Create Propagation Flow Graph (PFG)
 - For t from T_0 to T_f
 - Insert edge from active nodes to inactive ones in the main graph
 - Delete any edges between active nodes (there is no dependency between active nodes)
 - Delete any edges between inactive nodes (there is no dependency between inactive nodes)

3. Calculate Propagation probabilities

- For t from T_0 to T_f
- For IC Model:
 - For each inactive nodes do:
 - Based on flow graph and α_k - find $P(v_i|\{\varepsilon_{v_i}\})$
 - For $M = 1$ to n : calculate $P(v_i|\{\varepsilon_{v_{iM}}\})$

$$P(v_i|\{\varepsilon_{v_{iM}}\}) = \frac{\sum_{k=1}^m \alpha_k n_{v_i \rightarrow \varepsilon_{iM}}^k}{\sum_{k=1}^m \alpha_k \sum_{r \in nei_{\varepsilon_{iM}}} n_{\varepsilon_{iM} \rightarrow r}^k}$$

- For $M = 1$ to n :
 - If node(v_i) was activated by one of active neighbors:
Select max ($P(v_i|\varepsilon_{v_{iM}})$) which activated v_i and consider v_i as active.
 - else:
Select max ($P(v_i|\varepsilon_{v_{iM}})$) of neighbors and consider v_i as inactive.

$$P(v_i|\varepsilon_{v_i}) = \max_{M=1:n} \left[P(v_i|\varepsilon_{v_{iM}}) \right]$$

- For LT Model:
 - For each inactive node do:
 - Based on flow graph and α_k - find $P(v_i|\{\varepsilon_{v_i}\})$

$$P^k(v_i|\{\varepsilon_{v_i}\}) = \frac{P^k(v_i, \varepsilon_{v_{i1}}, \varepsilon_{v_{i2}}, \dots, \varepsilon_{v_{im}})}{P^k(\{\varepsilon_{v_i}\})}$$

- If $P(v_i) \geq Tr$, consider v_i as active.

4. Learn α_k for each Meta-path k

- For t from T_0 to T_f :

$$\alpha_k^{new} = \alpha_k^{new} + \varphi \frac{\partial \log(P(U_t))}{\partial \alpha_k}$$

5. Calculate F-Score and Recall measures as equation (2)

$$P^k(v_1, v_2) = \frac{n_{v_1 \rightarrow v_2}^k}{\sum_{i=1}^{n_u} \sum_{j=i}^{n_u} n_{v_i \rightarrow v_j}^k} \tag{3}$$

The above fraction can be considered as the information propagation probability between nodes v_1 and v_2 . In equation (3), $P^k(v_1, v_2)$ denotes the probability of the edge between nodes v_1 and v_2 connecting in meta-path k . n_u is the total number of existing nodes and $n_{v_1 \rightarrow v_2}^k$ represents the path instances between these nodes in meta-path k .

Node Propagation Probability: The strength of each node, i.e. the amount of information propagation the node is capable of, according to each meta-path is expressed by:

$$P^k(v) = \frac{\sum_{r \in nei_v} n_{v \rightarrow r}^k}{\sum_{i=1}^{n_u} \sum_{j=i}^{n_u} n_{v_i \rightarrow v_j}^k} \tag{4}$$

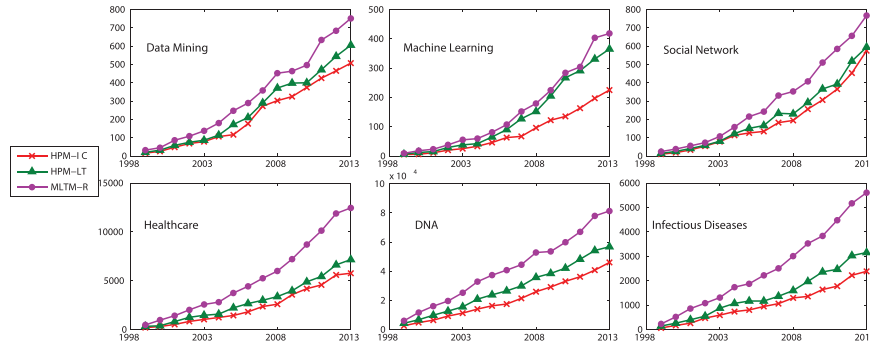


Figure 3. An illustrative example of Propagation Flow Graph where authors (active and inactive nodes) are connected to one another through papers (P). Edge of PFG illustrates that inactive nodes can be affected by active nodes.

For instance, an author with a higher number of published papers should be assigned higher influence strength for information spread. Probability of propagation from node v_1 to node v_2 in meta-path k is expressed using equation (5).

$$P^k(v_1|v_2) = \frac{P^k(v_1, v_2)}{P^k(v_2)} = \frac{n_{v_2 \rightarrow v_1}^k}{\sum_{r \in nei_v} n_{v_2 \rightarrow r}^k} \tag{5}$$

Propagation flow graph. The activation probability is assumed to be conditional as only an active node is capable of activating an inactive one, meaning that the direction of flow is always from the active node to the inactive one. Hence, the network is considered to be of Bayesian type. Additionally, we assume that active nodes are independent as an inactive node can only be activated by an active neighboring nodes and no flow may occur between two active nodes; hence no edge is considered between them. An implicit graph, with an example shown in Fig. 3, known as Propagation Flow Graph (PFG) is considered in this work. It should be noted that in order to calculate the node and edge propagation probabilities, the relationships between all nodes, both active or inactive ones are taken into account. In each state, if a node is activated, it is added to the PFG.

In our example shown in Fig. 3, nodes V2 and V4 are activate V1 as there are links from V2 and V4 to V1 on PFG. However, V3 can only be activated by V4 as there is no link from V1 to V3 on PFG. As V1 is activated in the first step, it can also affect V3 in the next step.

Propagation Probability. In this section, the activation probability for each node is calculated according to IC and LT diffusion model.

IC Model. In IC model, each inactive node has a single change to be activated by one of its active neighbors. In other words, if an inactive node is not activated by a recently activated neighbor node, it will not be considered in the next steps for being activated. Here, among the neighboring nodes of an inactive node that activated this node, the one with the maximum probability is selected as the activating node. Otherwise, if the state of an inactive node does not change we select the maximum probability of neighbors as the probability of this inactive node. The propagation probability from active neighboring nodes (ϵ_v) to an inactive node v_i through a given meta-path k is obtained according to:

$$P^k(v_i|\{\epsilon_{v_i}\}) = \max_{M=1:n} \left[\frac{P^k(v_i, \epsilon_{v_{iM}})}{P^k(\epsilon_{v_{iM}})} \right] \tag{6}$$

As mentioned before, we assume that active nodes are independent since no flow may occur between two active nodes. Since the overall probability is obtained as the sum of meta-paths, the overall activation probability of node v can be obtained as:

$$P(v) = \sum_{k=1}^{m=\text{number of metapaths}} \alpha_k P^k(v) \tag{7}$$

which means that a coefficient α_k is assigned to each meta-path to obtain the overall probability. Among the active neighboring nodes of inactive node v_i , the one with the maximum probability is selected as the activating node for node v_i .

$$\begin{aligned}
 \{\varepsilon_{v_i}\} &= \{\varepsilon_{v_{i_1}}, \varepsilon_{v_{i_2}}, \dots, \varepsilon_{v_{i_n}}\} \& \varepsilon_{v_{i_1}} \perp \varepsilon_{v_{i_2}} \perp \dots \perp \varepsilon_{v_{i_n}} \\
 P(v_i|\{\varepsilon_{v_i}\}) &= \max_{M=1:n} \left[\frac{\sum_{k=1}^m \alpha_k P^k(v_i, \varepsilon_{v_{iM}})}{\sum_{k=1}^m \alpha_k P^k(\varepsilon_{v_{iM}})} \right] \\
 &\approx \frac{1}{\sum_{i=1}^{n_u} \sum_{j=1}^{n_u} n_{v_i \rightarrow v_j}} \sum_{k=1}^m \alpha_k n_{v_i \rightarrow \varepsilon_{v_{iM}}}^k \\
 &\approx \frac{1}{\sum_{i=1}^{n_u} \sum_{j=1}^{n_u} n_{v_i \rightarrow v_j}} \sum_{k=1}^m \alpha_k \sum_{r \in nei_{\varepsilon_{v_{iM}}}} n_{\varepsilon_{v_{iM}}}^k \rightarrow r \\
 &\approx \max_{M=1:n} \left[\frac{\sum_{k=1}^m \alpha_k n_{v_i \rightarrow \varepsilon_{v_{iM}}}^k}{\sum_{k=1}^m \alpha_k \sum_{r \in nei_{\varepsilon_{v_{iM}}}} n_{\varepsilon_{v_{iM}}}^k \rightarrow r} \right] \tag{8}
 \end{aligned}$$

LT Model. As a more intuitive and closer assumption to the real world, LT model assumes that a node is activated if at least certain percentage of its neighbors have already been activated. In DBLP network for example, this means that the total number of studied papers from different authors can influence the author to publish a paper on a particular topic. The general type of LT model is as equation (1). On the other hand, due to assuming the conditional probability, we can obtain the probability of each inactive node. In this section, we keep the properties of LT model and conditional probability together. In this case, calculations of propagation probability through active neighboring nodes of node v_i are as follows:

$$\begin{aligned}
 P^k(v_i|\{\varepsilon_{v_i}\}) &= \frac{P^k(v_i, \varepsilon_{v_{i_1}}, \varepsilon_{v_{i_2}}, \dots, \varepsilon_{v_{i_n}})}{P^k(\{\varepsilon_{v_i}\})} \\
 &= \frac{P^k(v_i|\varepsilon_{v_{i_1}}) \times P^k(v_i|\varepsilon_{v_{i_2}}) \times \dots \times P^k(v_i|\varepsilon_{v_{i_n}}) \times P^k(\varepsilon_{v_{i_1}}) \times P^k(\varepsilon_{v_{i_2}}) \times \dots \times P^k(\varepsilon_{v_{i_n}})}{P^k(\varepsilon_{v_{i_1}}) \times P^k(\varepsilon_{v_{i_2}}) \times \dots \times P^k(\varepsilon_{v_{i_n}})} \\
 &= P^k(v_i|\varepsilon_{v_{i_1}}) \times P^k(v_i|\varepsilon_{v_{i_2}}) \times \dots \times P^k(v_i|\varepsilon_{v_{i_n}}) = \prod_{q=1}^n P^k(v_i|\varepsilon_{v_{i_q}}) \tag{9}
 \end{aligned}$$

It is obvious that each node v_i should have more $\prod_{q=1}^n P(v_i|\varepsilon_{v_{i_q}})$ for obtaining more influence. This means that with higher probability, the neighbors of node v_i will have more influence on it, which leads to:

$$\prod_{q=1}^n P(v_i|\varepsilon_{v_{i_q}}) \geq \lambda_{v_i} \tag{10}$$

We can infer that if multiplication of the neighbors' probability of an inactive node v_i becomes more than the threshold λ_{v_i} , the inactive node is more likely to be activated. Let us multiply a constant value v in both sides of equation 10 which does not change the final result:

$$\prod_{q=1}^n v P(v_i|\varepsilon_{v_{i_q}}) \geq v^n \lambda_{v_i} \tag{11}$$

By making logarithm from both sides of the above equation, we have equation (12) as:

$$\log_n \left(\prod_{q=1}^n v P(v_i|\varepsilon_{v_{i_q}}) \right) \geq \log_n (v^n \lambda_{v_i}) \rightarrow \sum_{q=1}^n \log_n (v P(v_i|\varepsilon_{v_{i_q}})) \geq \theta_{v_i} \quad \& \quad \log_n (v P(v_i|\varepsilon_{v_{i_q}})) \leq 1 \tag{12}$$

Equation (12) shows that by considering W_i as $\log_n (v P(v_i|\varepsilon_{v_{i_q}}))$ we kept the LT conditions and also we used conditional probability.

Learning model. Since information diffuses from active to inactive nodes, the flow of propagation is considered as a directed graph from active to inactive nodes. Moreover, due to their active state, no edge is considered between active nodes. Hence, according to PFG, the probability of all nodes is obtained through individual multiplication of active and inactive nodes. In the following, we explain the learning process used for IC and LT models.

IC model. If U_t is the set of all graph nodes, V_t the set of active nodes and R_t the set of inactive nodes at time t , the propagation probability for nodes is obtained by:

$$P(U_t) = \prod_{t \in T} \prod_{v \in V_t} P(v) \prod_{r \in R_t} (1 - P(r|\{\varepsilon_r\})) \tag{13}$$

The objective is to maximise $P(U_t)$; the probability of active nodes ($P(V_t)$) as well as that of unity minus the probability of inactive nodes ($1 - P(r|\{\varepsilon_r\})$) should be maximised to obtain the best results. For convenience, the function can be converted to log-likelihood function as:

$$\begin{aligned} \log(P(U_t)) &= \sum_{t \in T} \left[\sum_{v \in V_t} \sum_{k=1}^m \log(\alpha_k P^k(v)) + \sum_{r \in R_t} (1 - P(r|\{\varepsilon_r\})) \right] \\ 1 - P(r|\{\varepsilon_r\}) &= 1 - \max_{j=1:n} \left[\frac{\sum_{k=1}^m \alpha_k P^k(r, \varepsilon_{r_j})}{\sum_{k=1}^m \alpha_k P^k(\varepsilon_{r_j})} \right] \\ &= \min_{j=1:n} \left[\frac{\sum_{k=1}^m \alpha_k P^k(\varepsilon_{r_j}) - \sum_{k=1}^m \alpha_k P^k(r, \varepsilon_{r_j})}{\sum_{k=1}^m \alpha_k P^k(\varepsilon_{r_j})} \right] \\ \varphi \frac{\partial \log(P(U_t))}{\partial \alpha_k} &= \sum_{t \in T} \left[\sum_{v \in V_t} \frac{P(v)}{\sum_{k=1}^m \alpha_k P^k(v)} + \sum_{r \in R_t} \min_{j=1:n} \left[\frac{P^k(\varepsilon_{r_j}) - P^k(r, \varepsilon_{r_j})}{\sum_{k=1}^m \alpha_k P^k(\varepsilon_{r_j}) - \sum_{k=1}^m \alpha_k P^k(r, \varepsilon_{r_j})} \right] \right] \end{aligned} \tag{14}$$

LT model. Similar to the log-likelihood function in IC model, this can also be obtained when using LT as influence model. The resulting equation is as follows:

$$\begin{aligned} P(U_t) &= \prod_{t \in T} \prod_{v \in V_t} P(v) \prod_{r \in R_t} (1 - P(r|\{\varepsilon_r\})) \\ \log(P(U_t)) &= \sum_{t \in T} \left[\sum_{v \in V_t} \sum_{k=1}^m \log(\alpha_k P^k(v)) + \sum_{r \in R_t} (1 - P(r|\{\varepsilon_r\})) \right] \\ \varphi \frac{\partial \log(P(U_t))}{\partial \alpha_k} &= \sum_{v \in V_t} \frac{P(v)}{\sum_{k=1}^m \alpha_k P^k(v)} + \sum_{r \in R_t} \phi \frac{\partial}{\partial \alpha_k} \log \left[1 - \frac{\sum_{k=1}^m \alpha_k \{P^k(r, \varepsilon_{r_1}) \times P^k(r, \varepsilon_{r_2}) \times \dots \times P^k(r, \varepsilon_{r_n})\}}{\sum_{k=1}^m \alpha_k \{P^k(\varepsilon_{r_1}) \times P^k(\varepsilon_{r_2}) \times \dots \times P^k(\varepsilon_{r_n})\}} \right] \\ &\quad \times \phi \frac{\partial}{\partial \alpha_k} \log \left[1 - \frac{\sum_{k=1}^m \alpha_k \{P^k(r, \varepsilon_{r_1}) \times P^k(r, \varepsilon_{r_2}) \times \dots \times P^k(r, \varepsilon_{r_n})\}}{\sum_{k=1}^m \alpha_k \{P^k(\varepsilon_{r_1}) \times P^k(\varepsilon_{r_2}) \times \dots \times P^k(\varepsilon_{r_n})\}} \right] \\ &= \frac{\{P^k(\varepsilon_{r_1}) \times P^k(\varepsilon_{r_2}) \times \dots \times P^k(\varepsilon_{r_n})\} - \{P^k(r, \varepsilon_{r_1}) \times P^k(r, \varepsilon_{r_2}) \times \dots \times P^k(r, \varepsilon_{r_n})\}}{\sum_{k=1}^m \alpha_k \{P^k(\varepsilon_{r_1}) \times P^k(\varepsilon_{r_2}) \times \dots \times P^k(\varepsilon_{r_n})\} - \sum_{k=1}^m \alpha_k \{P^k(r, \varepsilon_{r_1}) \times P^k(r, \varepsilon_{r_2}) \times \dots \times P^k(r, \varepsilon_{r_n})\}} \\ &\quad - \frac{\{P^k(\varepsilon_{r_1}) \times P^k(\varepsilon_{r_2}) \times \dots \times P^k(\varepsilon_{r_n})\}}{\sum_{k=1}^m \alpha_k \{P^k(\varepsilon_{r_1}) \times P^k(\varepsilon_{r_2}) \times \dots \times P^k(\varepsilon_{r_n})\}} \end{aligned} \tag{15}$$

Ultimately, both models use equation (16) for calculating the coefficient of each meta-path (α_k).

$$\alpha_k^{new} = \alpha_k^{new} + \frac{\partial \log(P(U_t))}{\partial \alpha_k} \& \sum_{k \in \text{Metapaths}} \alpha_k = 1 \tag{16}$$

Example. In this section, we provide the above analysis on a sample network shown in Fig. 3. In this network, nodes V2 and V4 are active nodes, and thus can influence the inactive nodes V1, V3, V5 and V6, and activate them. Considering two meta-paths APA and APAPA, the probability of activation for each node can be calculated as follows:

APA:

$$\begin{aligned} P^{APA}(v_1|v_2) &= \frac{n_{v_2 \rightarrow v_1}^{APA}}{\sum_{r \in nei_{v_2}} n_{v_2 \rightarrow v_r}^{APA}} = \frac{1}{2}, P^{APA}(v_3|v_4) \\ &= \frac{n_{v_4 \rightarrow v_3}^{APA}}{\sum_{r \in nei_{v_4}} n_{v_4 \rightarrow v_r}^{APA}} = \frac{1}{8}, P^{APA}(v_1|v_4) = \frac{n_{v_4 \rightarrow v_1}^{APA}}{\sum_{r \in nei_{v_4}} n_{v_4 \rightarrow v_r}^{APA}} = \frac{1}{8}, \\ P^{APA}(v_6|v_4) &= \frac{n_{v_4 \rightarrow v_6}^{APA}}{\sum_{r \in nei_{v_4}} n_{v_4 \rightarrow v_r}^{APA}} = \frac{1}{8}, P^{APA}(v_5|v_4) = \frac{n_{v_4 \rightarrow v_5}^{APA}}{\sum_{r \in nei_{v_4}} n_{v_4 \rightarrow v_r}^{APA}} = \frac{1}{4} \end{aligned}$$

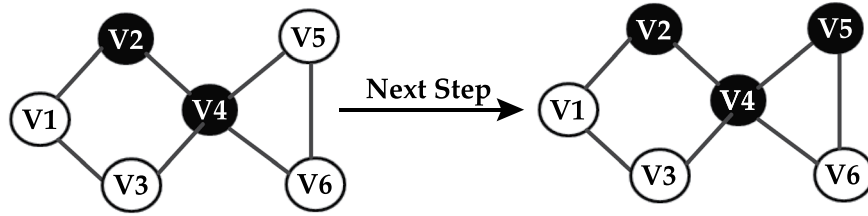


Figure 4. Process of activating inactive nodes affected by active neighbors for “Data Mining” topic.

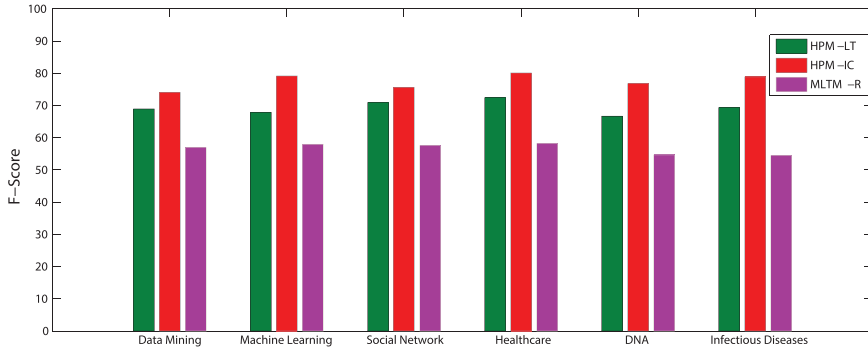


Figure 5. F-Score of the proposed method (HPM) with IC and LT as diffusion models (HPM-IC and HPM-LT) and the state-of-the-art method (MLTM-R) on DBLP dataset.

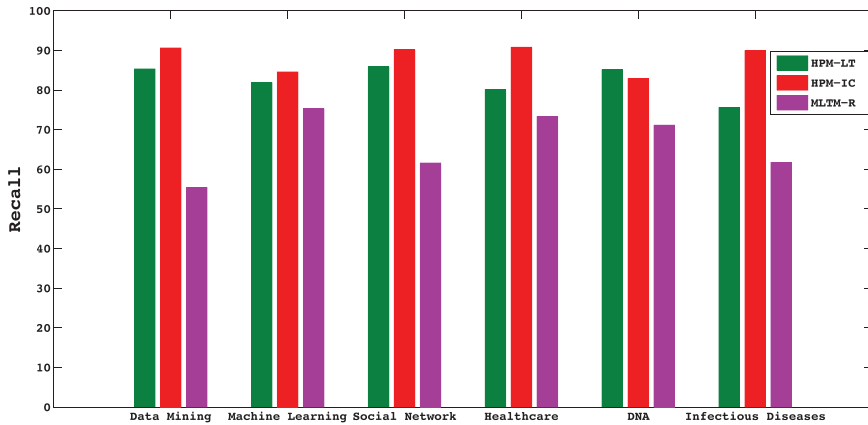


Figure 6. Recall of the methods on DBLP dataset.

$$\begin{aligned}
 & \text{APAPA:} \\
 & P^{APAPA}(v_1|v_2) = \frac{n_{v_2 \rightarrow v_1}^{APAPA}}{\sum_{r \in nei_{v_2}} n_{v_2 \rightarrow v_r}^{APAPA}} = \frac{1}{8}, P^{APAPA}(v_3|v_4) \\
 & = \frac{n_{v_4 \rightarrow v_3}^{APAPA}}{\sum_{r \in nei_{v_4}} n_{v_4 \rightarrow v_r}^{APAPA}} = \frac{2}{25}, P^{APAPA}(v_1|v_4) = \frac{n_{v_4 \rightarrow v_1}^{APAPA}}{\sum_{r \in nei_{v_4}} n_{v_4 \rightarrow v_r}^{APAPA}} = \frac{11}{25}, \\
 & P^{APAPA}(v_6|v_4) = \frac{n_{v_4 \rightarrow v_6}^{APAPA}}{\sum_{r \in nei_{v_4}} n_{v_4 \rightarrow v_r}^{APAPA}} = \frac{6}{25}, P^{APAPA}(v_5|v_4) = \frac{n_{v_4 \rightarrow v_5}^{APAPA}}{\sum_{r \in nei_{v_4}} n_{v_4 \rightarrow v_r}^{APAPA}} = \frac{3}{25}
 \end{aligned}$$

where the values of α_{APA} and α_{APAPA} are learned for the corresponding meta-paths. Assuming the learned values for α_{APA} and α_{APAPA} as 0.6 and 0.4, respectively, the final activation probability is obtained as:

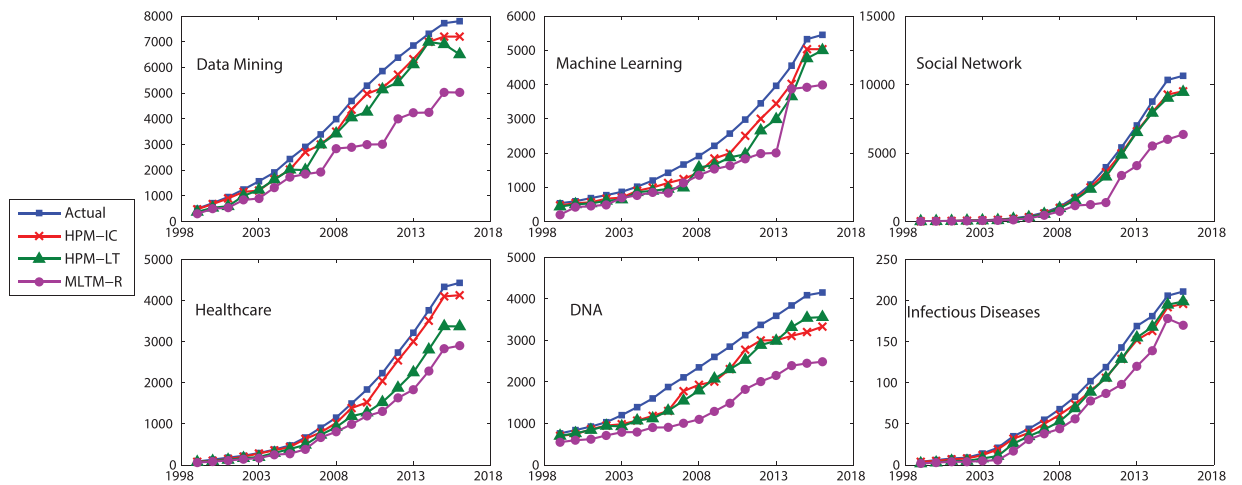


Figure 7. The number of correctly predicted active authors (TP) for the selected topics on DBLP dataset.

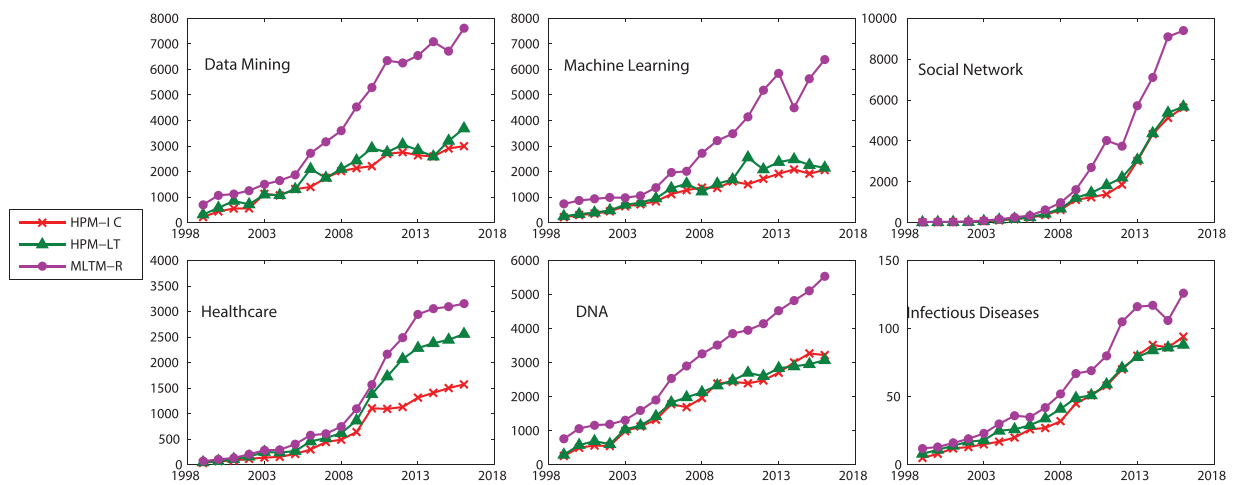


Figure 8. The number of authors who have been tagged incorrectly as active (FP) or inactive (FN) for the selected topics on DBLP dataset.

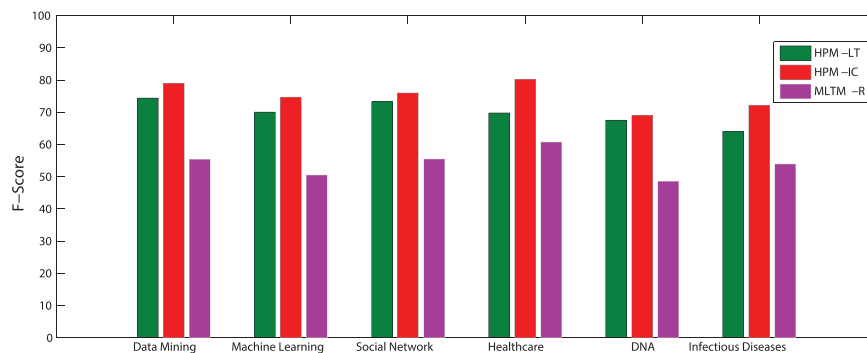


Figure 9. F-Score of the proposed method (HPM) with IC and LT as diffusion models (HPM-IC and HPM-LT) and the state-of-the-art method (MLTM-R) on PubMed dataset.

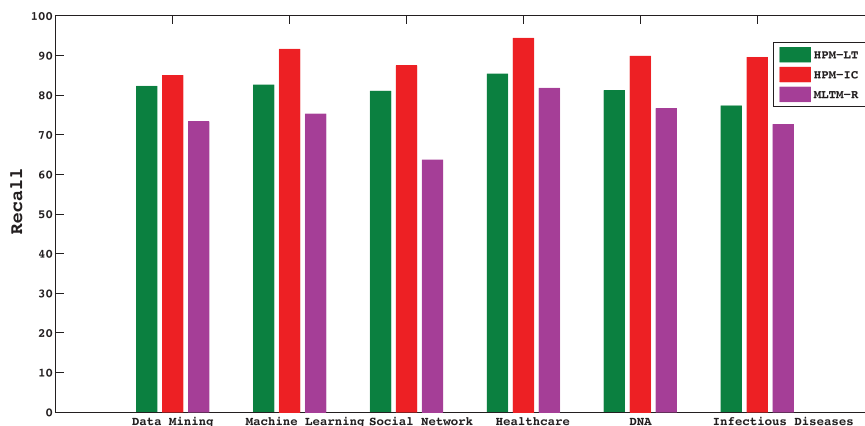


Figure 10. Recall of the methods on PubMed dataset.

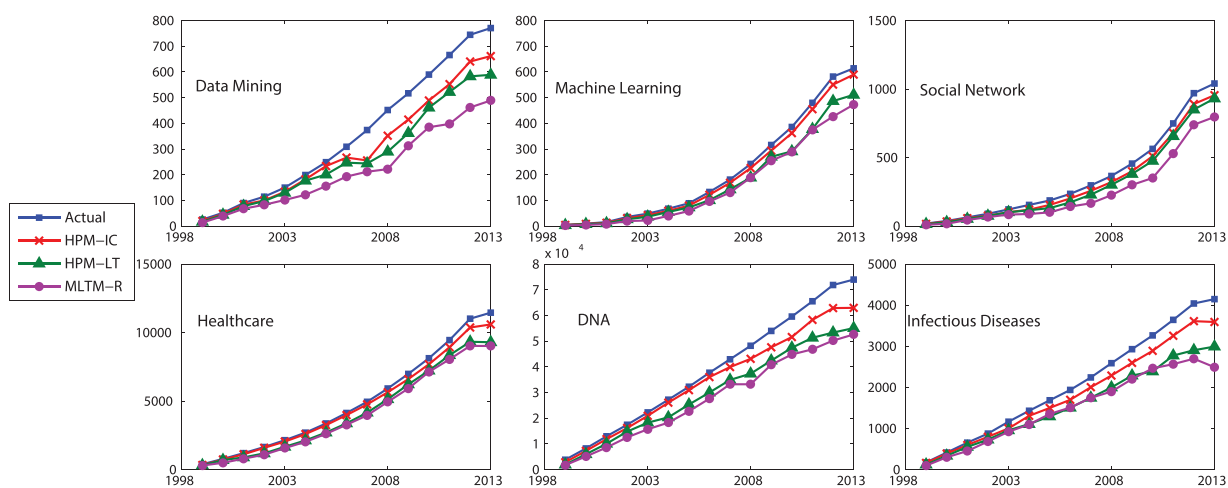


Figure 11. The number of correctly predicted active authors (TP) for the selected topics on PubMed dataset.

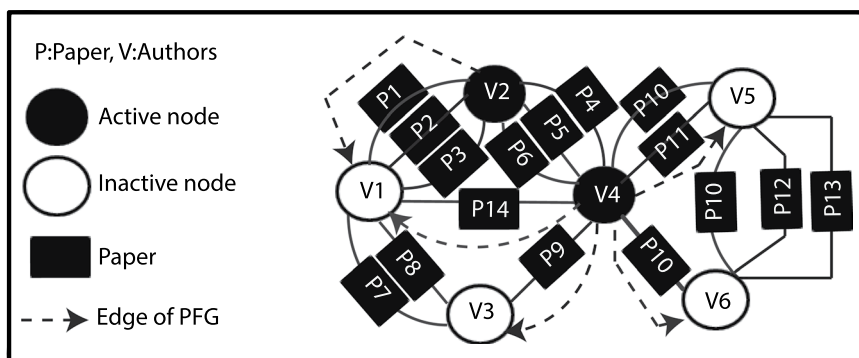


Figure 12. The number of authors who have been tagged incorrectly as active (FP) or inactive (FN) for the selected topics on PubMed dataset.

$$\begin{aligned}
 & \text{IC:} \\
 P(v_1|\{v_2, v_4\}) &= \max_{M=1:n} \left[\frac{\sum_{k=1}^m \alpha_k n_{v_1 \rightarrow v_2}^k}{\sum_{k=1}^m \alpha_k \sum_{r \in \text{nei}_{v_2}} n_{v_2 \rightarrow r}^k}, \frac{\sum_{k=1}^m \alpha_k n_{v_1 \rightarrow v_4}^k}{\sum_{k=1}^m \alpha_k \sum_{r \in \text{nei}_{v_4}} n_{v_4 \rightarrow r}^k} \right] \\
 &= \max_{M=1:n} \left[\frac{0.6 \times 3 + 0.4 \times 3}{0.6 \times 6 + 0.4 \times 24}, \frac{0.6 \times 1 + 0.4 \times 11}{0.6 \times 8 + 0.4 \times 25} \right] \\
 P(v_3|\{v_4\}) &= \frac{0.6 \times 1 + 0.4 \times 2}{0.6 \times 8 + 0.4 \times 25}, \\
 P(v_5|\{v_4\}) &= \frac{0.6 \times 2 + 0.4 \times 3}{0.6 \times 8 + 0.4 \times 25}, P(v_6|\{v_4\}) = \frac{0.6 \times 1 + 0.4 \times 6}{0.6 \times 8 + 0.4 \times 25}
 \end{aligned}$$

LT:

$$\begin{aligned}
 P^{APA}(v_1|\{v_2, v_4\}) &= \frac{1}{2} \times \frac{1}{8} = \frac{1}{16}, \\
 P^{APAPA}(v_1|\{v_2, v_4\}) &= \frac{1}{8} \times \frac{11}{25} = \frac{11}{200}
 \end{aligned}$$

Experimental Results

As mentioned for DBLP example, the authors are connected to one another according to the specified meta-path. For topic diffusion in such a graph, initially we need to select a special topic like “data mining”. The authors with papers related to the selected topic are considered as active nodes. These authors affect their neighbors in a way that of an inactive node (author) might be encouraged to write a paper in this field affected by active author(s). If a neighbor writes paper in this field, they will be active and will then affect their neighbors.

Figure 4 shows an example in which in the first step nodes V2 and V4 are activated by “data mining” topic. Node V2 can activate node V1 while node V4 can activate nodes V3, V5 and V6. In this example, node V4 activates node V5, hence node V5 is persuaded to write paper in “data mining” topic. In this section, we apply the proposed model on two real datasets and discuss the results. We consider two popular datasets, DBLP and PubMed, which include information on authors, papers and venues. We also consider some topics including data mining, machine learning, social networks, healthcare, DNA, and infectious disease, for which the diffusion process is modeled. The topic selection is mainly due to their convenient frequency in the datasets and the considerable amount of data available for comparison and conclusion.

DBLP. In this dataset, information diffusion is investigated on the selected topics. The results of the proposed method is compared to the state-of-the-art method introduced by Gui *et al.*³⁰, known as MLTM-R. Figures 5 and 6 compare the performance of the proposed model, Heterogeneous Probability Model (HPM), with MLTM-R in terms of F-score and Recall, respectively. Note that the original MLTM-R method is based on LT model for diffusion, while HPM works for both LT and IC models. As it can be seen, HPM significantly outperforms MLTM-R by providing much better F-score and Recall when IC model is used. An improvement of about 30–50% is obtained in HPM as compared to MLTM-R. Furthermore, these results show that one can obtain much better performance when IC model is used rather than LT. This indicates that IC model is better capable of modeling topic diffusion in this dataset.

Figure 7 compares TP, i.e., the number of correctly predicted active authors, of the methods. It also include the actual TPs for different years, where the closer is the predicted value to these actual values, the better is the performance of the method. As it is seen, the proposed method with IC model (HPM-IC) has the closest predicted values to the actual ones, followed by HPM-LT and then MLTM-R. This performance is observed across all the selected topics and all years. Figure 8 shows the number of authors who have been incorrectly identified as active or inactive, where HPM-IC has the lowest values (i.e., the best performance) while MLTM has the worst performance.

PubMed. We apply the methods on PubMed dataset with the same selected topics. Figures 9 and 10 show the F-score and Recall of the methods, respectively. Similar to the other dataset, HPM significantly outperforms MLTM-R in all topics. Also, HPM-IC performs better HPM-LT. Figures 11 and 12 show the correctly identified active authors (TP) and incorrectly identified active and inactive authors (FN and FP), respectively. As it is seen, similar to the other dataset, HPM-IC has the best performance.

Analysis. Compared to MLTM-R method³⁰, HPM-LT and HPM-IC methods significantly improve the F-score and Recall of the prediction, which is mainly due to the following reasons. MLTM-R uses pathsim to calculate the weight of each edge. Pathsim is not accurate in some cases^{31,32}, as it does not obtain similarity value (or obtain low similarity scores) between two similar nodes in certain circumstances. However, in our proposed method, each meta-path instance is considered as a path by considering different routes between nodes, which eliminates the problems of Pathsim as there is no need to calculate the similarity for weights. The proposed method instead uses the conditional probability model to calculate the activation probability of each node. The inactive nodes are considered to be dependent on the active neighboring nodes. This is a realistic scenario as if an author decides to write a paper about an issue, they should have already be aware of the existence papers written by others (active nodes). Unlike the other method, in the proposed algorithm we separately consider the node and edge influence.

The node influence is considered by having IC and LT models in which activation of inactive nodes is based on neighboring active nodes. The edge influence is considered as the extent to which the relation between two nodes is important for diffusion process i.e. a relation is more impressive if larger number of multipaths are found between two nodes. Topological properties of networks have significant influence on the way information propagates on them. DBLP has larger average degree than PubMed and having more connections facilitates spread. Our results also confirms this as the performance of the methods is better for DBLP than PubMed.

Better performance of the proposed strategy over of the previous model is due to considering information extracted from meta-paths. A meta-path is a path between any two nodes from different layers of an heterogeneous network. As meta-path traverses between different type of object, it can extract useful information on the structure of the network. method based on meta- paths have already been used for network analysis such as link prediction⁵. Our experiments shows that meta-paths are also important in the way information spread across layers and different object types. We also consider importance of the nodes by taking into account the paths passing through them (equation 4).

Our proposed method use meta-paths with different lengths. Two non-adjacent nodes of the same type, e.g. two authors in DBLP example, might be connected through meta-paths of length two or three. For example, in DBLP network when two authors who do not have any co-authored papers, both have papers with another authors, there is a meta-path of length two between these two authors. Considering such meta-paths allows one to account for such indirect connections between the nodes and taking into account the cross-layer information at the same time.

Conclusion

This paper studied information spread and diffusion of scientific topics in heterogeneous networks. To this end, a novel method called HPM, was developed based on meta-paths and conditional probability. Moreover, propagation flow graph was defined to illustrate the diffusion flow from active to inactive nodes. Propagation probability was then calculated based on this graph and the coefficients of meta-paths were learned using the log-likelihood function. We considered two well-known diffusion models: Linear Threshold (LT) and Independent Cascade (IC) models. In LT model, inactive nodes are activated if the portion of their active neighbors is higher than a certain threshold. In IC model, the recently activate nodes activates its inactive neighbors with a certain probability. We considered the problem of topic diffusion in two real-world networks: DBLP and PubMed. The performance of the proposed model was compared with a state-of-the-art method, where our experimental results showed that the proposed method significantly outperform the other one. Also, Using IC as the diffusion model led to better performance than LT model.

References

- Arenas, A., Díaz-Guilera, A., Kurths, J., Moreno, Y. & Zhou, C. Synchronization in complex networks. *Physics Reports* **469**, 93–153, <https://doi.org/10.1016/j.physrep.2008.09.002> (2008).
- Olfati-Saber, R., Fax, J. A. & Murray, R. M. Consensus and Cooperation in Networked Multi-Agent Systems. *Proceedings of the IEEE* **95**, 215–233, <https://doi.org/10.1109/JPROC.2006.887293> (2007).
- Jalili, M. Social power and opinion formation in complex networks. *Physica A: Statistical Mechanics and its Applications* **392**, 959–966, <https://doi.org/10.1016/j.physa.2012.10.013> (2013).
- Jalili, M. Effects of leaders and social power on opinion formation in complex networks. *Simulation* **89**, 578–588, <https://doi.org/10.1177/0037549712462621> (2013).
- Jalili, M. & Perc, M. Information cascades in complex networks. *Journal of Complex Networks*, <https://doi.org/10.1093/comnet/cnx019> (2017).
- Kirst, C., Timme, M. & Battaglia, D. Dynamic information routing in complex networks. *Nature Communications* **7**, 11061, <https://doi.org/10.1038/ncomms11061> (2016).
- Zhang, Z.-K. *et al.* Dynamics of information diffusion and its applications on complex networks. *Physics Reports* **651**, 1–34, <https://doi.org/10.1016/j.physrep.2016.07.002> (2016).
- Deng, H., Han, J., Zhao, B., Yu, Y. & Lin, C. X. Probabilistic Topic Models with Biased Propagation on Heterogeneous Information Networks. *Kdd* 1271–1279, <https://doi.org/10.1145/2020408.2020600> (2011).
- Han, J. Mining heterogeneous information networks by exploring the power of links. In *Discovery Science*, 13–30, https://doi.org/10.1007/978-3-642-04747-3_2 (2009).
- Sun, Y. & Han, J. Mining heterogeneous information networks: a structural analysis approach. *ACM SIGKDD Explorations Newsletter* **14**, 20–28, <https://doi.org/10.1145/2481244.2481248> (2013).
- Zhou, D., Orshanskiy, S. A., Zha, H. & Giles, C. L. Co-ranking Authors and Documents in a Heterogeneous Network. In *Seventh IEEE International Conference on Data Mining (ICDM 2007)*, 739–744, <https://doi.org/10.1109/ICDM.2007.57> (2007).
- Angelova, R., Kasneci, G. & Weikum, G. Graffiti: graph-based classification in heterogeneous networks. *World Wide Web* **15**, 139–170, <https://doi.org/10.1007/s11280-011-0126-4> (2012).
- Sun, Y. & Han, J. Mining Heterogeneous Information Networks: Principles and Methodologies. *Synthesis Lectures on Data Mining and Knowledge Discovery* **3**, 1–159, <https://doi.org/10.2200/S00433ED1V01Y201207DMK005> (2012).
- Sun, Y. & Han, J. Mining heterogeneous information networks. *ACM SIGKDD Explorations Newsletter* **14**, 20, <https://doi.org/10.1145/2481244.2481248> (2013).
- Kralj, J., Robnik-Šikonja, M. & Lavrač N. HINMINE: heterogeneous information network mining with information retrieval heuristics. *Journal of Intelligent Information Systems*, <https://doi.org/10.1007/s10844-017-0444-9> (2017).
- Boccaletti, S. *et al.* The structure and dynamics of multilayer networks. *Physics Reports* **544**, 1–122, <https://doi.org/10.1016/j.physrep.2014.07.001> (2014).
- Liu, X., Yingying, Yu, Guo, C., Sun, Y. & Gao, L. Full-text based context-rich heterogeneous network mining approach for citation recommendation. In *IEEE/ACM Joint Conference on Digital Libraries*, 361–370, <https://doi.org/10.1109/JCDL.2014.6970191> (2014).
- Yang, R. *et al.* Epidemic spreading on heterogeneous networks with identical infectivity. *Physics Letters A* **364**, 189–193, <https://doi.org/10.1016/j.physleta.2006.12.021> (2007).
- Moreno, Y., Pastor-Satorras, R. & Vespignani, A. Epidemic outbreaks in complex heterogeneous networks. *The European Physical Journal B* **26**, 521–529, <https://doi.org/10.1140/epjb/e20020122> (2002).
- Salehi, M. *et al.* Spreading processes in multilayer networks. *IEEE Transactions on Network Science and Engineering* **2**, 65–83, <https://doi.org/10.1109/TNSE.2015.2425961> (2015).

21. Wang, L. & Dai, G. Z. Global stability of virus spreading in complex heterogeneous networks. *Siam Journal on Applied Mathematics* **68**, 1495–1502, <https://doi.org/10.1137/070694582> (2008).
22. Yang, H., Tang, M. & Gross, T. Large epidemic thresholds emerge in heterogeneous networks of heterogeneous nodes. *Scientific Reports* **5**, 13122, <https://doi.org/10.1038/srep13122> (2015).
23. Nadini, M. *et al.* Epidemic spreading in modular time-varying networks. *Scientific Reports* **8**, 2352, <https://doi.org/10.1038/s41598-018-20908-x> (2018).
24. Demirel, G., Barter, E. & Gross, T. Dynamics of epidemic diseases on a growing adaptive network. *Scientific reports* **7**, 42352, <https://doi.org/10.1038/srep42352> (2017).
25. Shakibian, H. & Moghadam Charkari, N. Mutual information model for link prediction in heterogeneous complex networks. *Scientific Reports* **7**, 44981, <https://doi.org/10.1038/srep44981> (2017).
26. Jalili, M., Orouskhani, Y., Asgari, M., Alipourfard, N. & Perc, M. Link prediction in multiplex online social networks. *Royal Society Open Science* **4**, 160863, <https://doi.org/10.1098/rsos.160863> (2017).
27. Sermpezis, P. & Spyropoulos, T. Information diffusion in heterogeneous networks: The configuration model approach. In *Proceedings - IEEE INFOCOM*, 3261–3266, <https://doi.org/10.1109/INFOCOM.2013.6567148> (2013).
28. Zhou, Y. & Liu, L. Social influence based clustering of heterogeneous information networks. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, 338–346, <https://doi.org/10.1145/2487575.2487640> (2013).
29. Li, S. & Jin, Z. Modeling and Analysis of New Products Diffusion on Heterogeneous Networks. *Journal of Applied Mathematics* **2014**, 1–12, <https://doi.org/10.1155/2014/940623> (2014).
30. Gui, H., Sun, Y., Han, J. & Brova, G. Modeling Topic Diffusion in Multi-Relational Bibliographic Information Networks. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management - CIKM '14*, 649–658, New York, New York, USA), <https://doi.org/10.1145/2661829.2662000> (2014).
31. Shang, J. *et al.* Meta-path guided embedding for similarity search in large-scale heterogeneous information networks. *preprint at* <https://arxiv.org/abs/1610.09769> (2016).
32. Kuck, J., Zhuang, H., Yan, X., Cam, H. & Han, J. Query-Based Outlier Detection in Heterogeneous Information Networks. *Advances in database technology: proceedings. International Conference on Extending Database Technology* **2015**, 325–336, <https://doi.org/10.5441/002/edbt.2015.29> (2015).
33. Goldenberg, J., Libai, B. & Muller, E. Talk of the network: A complex systems look at the underlying process of word-of-mouth. *Marketing letters* **12**, 211–223, <https://doi.org/10.1023/A:1011122126881> (2001).
34. Granovetter, M. S. Threshold Models of Collective Behavior. *American Journal of Sociology* **83**, 1420–1443, <https://doi.org/10.1086/226707> (1978).
35. Watts, D. J. A simple model of global cascades on random networks. *Proceedings of the National Academy of Sciences* **99**, 5766–5771, <https://doi.org/10.1073/pnas.082090499> (2002).
36. Tang, J. *et al.* Arnetminer: extraction and mining of academic social networks. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, 990–998, <https://doi.org/10.1145/1401890.1402008> (2008).
37. Light, R. P., Polley, D. E. & Börner, K. Open data and open code for big science of science studies. *Scientometrics* **101**, 1535–1551, <https://doi.org/10.1007/s11192-014-1238-2> (2014).
38. LaRowe, G., Ambre, S., Burgoon, J., Ke, W. & Börner, K. The scholarly database and its utility for scientometrics research. *Scientometrics* **79**, 219–234, <https://doi.org/10.1007/s11192-009-0414-2> (2008).

Acknowledgements

S.M., S.B. and M.S. contribute equally in this work. This research was in part supported by a grant from IPM (No. CS1396-4-49) and Mahdi Jalili is supported by Australian Research Council through project No. DP170102303.

Author Contributions

They conceived the study, performed the experiments, analysed the data, and wrote the manuscript. M.J. analysed the results and wrote the paper. All authors approved the final version of the manuscript.

Additional Information

Supplementary information accompanies this paper at <https://doi.org/10.1038/s41598-018-27385-2>.

Competing Interests: The authors declare no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2018