Expert Systems with Applications 160 (2020) 113580

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Identifying influential nodes in heterogeneous networks

Soheila Molaei [a], Reza Farahbakhsh [b], Mostafa Salehi [a,c,*], Noel Crespi [b]

[a] Faculty of New Sciences and Technologies, University of Tehran, Tehran, Iran
[b] Institut Polytechnique de Paris, Telecom SudParis, Evry, France
[c] Institute for Research in Fundamental Science (IPM), Tehran, Iran

ARTICLE INFO

Article history:
Received 21 August 2019
Revised 14 May 2020
Accepted 17 May 2020
Available online 20 June 2020

Keywords:
Social media
Influence
Influential nodes
Scholar
Heterogeneous networks

ABSTRACT

Identifying influential users and measure the influence of nodes in social networks have become an interesting and important topic of research. It is crucial to find out to what extent individuals influence each other because it can be used to control rumors, diseases, and diffusion. There are numerous relevant models most of which are based on a homogeneous network. However, in the real world, we face heterogeneous networks where the nodes and edges are different types. A network is homogeneous if and only if the edges and nodes are of the same type, and it is considered heterogeneous if the nodes and edges are different. In heterogeneous networks, there is a concept known as meta-path, which indicates the type of communication between two nodes. In this paper, we aim to locate influential nodes by calculating the entropy of different meta-paths. To evaluate information diffusion in a heterogeneous network, we used the known susceptible-infectious model. The results of our experiments on three real-world networks' dataset show that the proposed method outperforms state-of-the-art influence maximization algorithms.

1. Background & summary

Social networks are shaping and reshaping people's everyday life and frequently we see a new one borrowing the key features (friendships, relations, posts, reactions, etc.) but introducing new features (e.g. privacy, circles, etc.). In this era, based on popularity, they can influence others opinion and behavior. Social influence (Althoff, Jindal, & Leskovec, 2017; Wen & Deng, 2020) occurs when individuals change their behavior or opinion under the influence of others depending on a variety of factors such as the relationship between individuals and the type of network (Mohammadinejad, Farahbakhsh, & Crespi, 2018). Health care (Shakya et al., 2017), online recommendation (Flanagin, 2017), career choices (Eesley & Wang, 2017), rumor spreading (Fraigniaud & Natale, 2019), political campaigning to reach out to a maximum audience (Vitak et al., 2011), and virus spreading (Zhu, Li, & Gan, 2018) are affected by social influence. Social influence (Liu, Jing, Zhao, Wang, & Song, 2017) has been broadly analyzed in social networks (Pei, Teng, Shaman, Morone, & Makse, 2017). Social networks mainly intend to identify the most influential users. Influential nodes are located to control rumors (Borge-Holthoefer & Moreno, 2012; Wei & Deng, 2019), outbreak of diseases (Gu, Lee, Saramäki, & Holme, 2017; Molaei, Khansari, Veisi, & Salehi, 2019) and, marketing (Chen, Wang, & Wang, 2010).

Finding Influential nodes would be effective in the context of Expert Systems. Tselykh and et al. (Tselykh, Tselykh, Vasilev, & Barkovskii, 2018; Mohammadinejad et al., 2018) used the maximization of the spread of influence in an expert system for solving the knowledge-discovery problem. A bottleneck of expert systems (ESs) is the problem of knowledge acquisition (KA) and the automation of this process, which remains a topical concern. They (Tselykh, Vasilev, & Tselykh, 2020) also proposed a new influence productivity assessment methodology that is a cognitive intelligence system for the scenario planning of control impacts (generation and choice). Besides, some papers (Dinh & Thai, 2011; Shen, Nguyen, Xuan, & Thai, 2012; Tselykh, Vasilev, & Tselykh, 2019) tried to find influential nodes for assessing network vulnerability.

There are two different groups of related problems. The first group aim is to identify the single node of greatest importance. Usually, the process starts from a single node in applications such as an epidemic outbreak or spreading rumors. The second group tries to find a set of nodes to maximize the final influence of spreading. The first group of problems is generally referred to as the 'most influential node identification' in which most approaches rank all nodes heuristically first based on some metric of node importance and then select the top-ranking node as the most important. Finding nodes to maximize influence in a network is

* Corresponding author at: Faculty of New Sciences and Technologies, University of Tehran, Tehran, Iran.
E-mail addresses: soheila.molaei@ut.ac.ir (S. Molaei), reza.farahbakhsh@it-sudparis.eu (R. Farahbakhsh), mostafa_salehi@ut.ac.ir (M. Salehi), noel.crespi@it-sudparis.eu (N. Crespi).

https://doi.org/10.1016/j.eswa.2020.113580
0957-4174/

known as the Influence Maximization (IM) problem. In this study, we will focus on both groups by employing an entropy-based approach.

How the spreading process is a particular interest, as there is an assumption that influential nodes are most likely to be infected (Mo & Deng, 2019; Yang, Wang, Lai, Xie, & Wang, 2011). It is critical to understand the structure of various networks and their interconnections (Fei, Zhang, & Deng, 2018; Huang & Yu, 2017; Malliaros, Rossi, & Vazirgiannis, 2016) which has mostly discussed in homogeneous networks. Tang et al. (Tang et al. (2020)) introduce a successful discrete shuffled frog-leaping algorithm (DSFLA) to address the question of maximization of influence in a more constructive way. A new encoding framework and distinct evolutionary rules for the virtual frog population was formulated based on network topology structure. Raychaudhuri and their colleagues (Raychaudhuri, Mallick, Sircar, & Singh, 2020) explore the role of local properties and node position in comparison to the entire network for influential node identification. Ohara et al. (Ohara, Saito, Kimura, & Motoda, 2020) suggested a predictive simulation method focused on the leave-N-out cross-validation technique that approximates well the error of the unknown ground truth for two problems: one to estimate the influence degree of each node, and the other to identify top-K influential nodes. Their suggested method for the first issue estimates the approximation error of the degree of each node's influence, and the method for the second issue estimates the precision of the derived top-K nodes.

In recent years, however, the focus has shifted to heterogeneous networks (Molaei, Zare, & Veisi, 2020). An information network, i.e., G = (V, E) with V nodes and edges, is homogeneous if and only if the edges and nodes are of the same type. Conversely, it is heterogeneous if it contains different nodes and edges. These edges can be indicative of different types of relationships (Deng, Han, Zhao, Yu, & Lin, 2011; Han, 2009; Sun & Han, 2013). In addition to locating influential nodes in homogeneous networks (Liu, Lin, Guo, & Zhou, 2016; Sun, Ma, Zeng, & Wang, 2016; Wang, Wang, & Deng, 2019), numerous studies have been conducted on ranking, classification, clustering, link analysis, semantic analysis and prediction of homogeneous networks (Duan, Wei, Zhou, & Shum, 2012; Eliacik & Erdogan, 2018; Mihalcea & Tarau, 2004; Page, Brin, Motwani, & Winograd, 1999; Ranjbar, Salehi, Jandaghi, & Jalili, 2019; Zhang et al., 2019). In practice, however, most real-world networks are heterogeneous. Moreover, the assumption of network homogeneity may lead to loss of key semantic information and failure to understand the information in its entirety. Therefore, it is necessary to define a new paradigm to study heterogeneous networks and extend the recently developed tools in the field of Network Science (Kivelä et al., 2014).

In heterogeneous networks, numerous studies have also been conducted on semantic parsing (Bordes, Glorot, Weston, & Bengio, 2012), link prediction (Bordes, Usunier, Garcia-Duran, Weston, & Yakhnenko, 2013; Socher, Chen, Manning, & Ng, 2013; Wang, Zhang, Feng, & Chen, 2014) and topic diffusion (Molaei, Babaei, Salehi, & Jalili, 2018). Kuhnle et al. (Kuhnle, Alim, Li, Zhang, & Thai, 2018) investigated multiplex Influence maximization in Social Networks with Heterogeneous diffusion models. They identified a new property, generalized deterministic submodular ensures that the propagation on the multiplex overall is submodular. The proposed algorithm runs in each layer of multiplex network in parallel.

Nonetheless, this paper concentrated on locating influential nodes in heterogeneous networks using the concepts of entropy and meta-path. Entropy was previously employed in Peng, Yang, Cao, Yu, and Xie (2017) to locate influential nodes in a homogeneous network; the current study extends this concept to meta-path in heterogeneous networks.

Meta-path P is defined based on the general scheme of $T_G = (A, G)$ network where A represents the type of nodes, R represents the relationships between the nodes (type of edges), and represents the corresponding meta-path. It is displayed as $A_1R_1A_2R_2 \ldots R_lA_{l+1}$ indicating a $R_1 \cup \ldots \cup R_l$ composition function between $A1$ and A_{l+1}. Moreover, o represents a composition operator on communications (Sun & Han, 2012). For instance, in DBLP (a computer science bibliography), where nodes are the authors connected through different functions, author-paper-author and author-conference-author are considered as separate meta-paths. As shown in Fig. 1, in DBLP network researchers can be connected through opposite meta-path. This can be considered a meta-path instance.

Scholars have not yet fully explored the concept of meta-path. In this paper, the influential nodes were investigated in heterogeneous networks using the concepts of entropy and meta-path. We compute the probability of a node influencing its neighbors in each meta-path and select a subset of nodes so that the number of nodes in the network that are influenced is maximized.

In this paper, a novel entropy-based method is proposed to find the influential nodes in a graph. This proposed method combines local and global information, which are two significant dimensions in a network. Local information considers direct neighbors' information and global information contains indirect neighbors' information via meta-paths. These two kinds of entropy can find influential nodes in a more reasonable and effective way. Based on this foundation, this proposed structure entropy considers not only the local structure but also the global structure, which modifies the previous method which can focus on only one of the structural properties of the network.

Shannon first introduced the entropy theory (Shannon, 1948; Sandoval, 2014). According to Shannon, the main problem of information theory is how to reproduce at one point a message sent from another point. If one considers a set of possible events whose probabilities of occurrence are p_i, $i = 1, \ldots, n$, then a measure $H(p_1, p_2, \ldots, p_n)$ of the uncertainty of the outcome of an event given such distribution of probabilities should have the following three properties:

- $Entropy = -\sum_{i=1}^{n} P_i \log_{10} P_i$;
- $H(p_i)$ should be continuous in p_i;
- if all probabilities are equal, what means that $p_i = \frac{1}{n}$, then H should be a monotonically increasing function of n (if there are more choices of events, and the uncertainty about one outcome should increase).

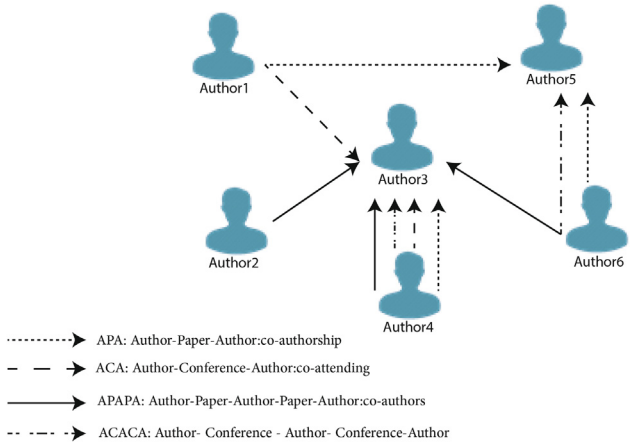


Fig. 1. An Example of a heterogeneous network.

The main contributions of the study are as follows:

1. An entropy-based method was proposed to measure the influence of nodes on each other in order to obtain the influential nodes. This method was extended to heterogeneous network meta-path taking into account meta-path such as APA, APAPA, ACA, and ACACA. The influence of each meta-path was considered separately.
2. The proposed method was based on neighbors, meta-path instances and a combination of both.
3. A criterion was introduced to evaluate and compare the proposed method with other conventional methods. The diffusion method was based on susceptible-infectious (SI).
4. A presentation of an entropy based method that brings improvements to the state-of-the-art methods.

Next, we review the relevant literature and discuss our proposed method, and finally last Section includes the analysis of results and conclusion.

2. Entropy Rank Method (ERM)

The entropy method was adopted to calculate the propagation probability of each node. The entropy method was previously used to maximize the diffusion through influential nodes in single-layer networks (Peng et al., 2017). This study attempted to extend the concept of entropy to meta-paths in heterogeneous networks. Three categories of probability entropies were obtained: (1) based on neighbors in each meta-path, (2) both neighbors and meta-path instances simultaneously, and (3) based on the number of meta-path instances.

Preliminaries:

Single-mediator meta-path: Meta-paths interconnected only through one mediator (such as R) are called single-mediator. For example, in meta-pathA_1RA_2, author A_1 is connected to author A_2 only through paper P.

Multi- or (n) mediator meta-paths: Meta-paths interconnected through more than one mediator are called multi-mediator. For example, in meta-path $A_1P_1A_2P_2A_3$, author A_1 connected to author A_3 through two papers (mediators) P_1 and P_2.

In this paper we covered short meta-path lengths because in Sun, Han, Yan, Yu, and Wu (2011), we can see that the meta path with relatively short length is good enough, and a long meta path may even reduce the quality.

Algorithm 1. Finding Influential nodes

Input: Datasets
Output: Top Influential Nodes
P_{ij}^K =The probability of reaching to node j from node i based on neighbors in meta-path k
Pr_{ij}^K =The probability of reaching to node j from node i based on neighbors and meta path instances in meta-path k
Pc_{ij}^K =The probability of reaching to node j from node i based on meta path instances in meta-path k
If_i^K =Entropy of node i based on neighbors in meta-path k
Im_i^K =Entropy of node i based on neighbors and meta path instances in meta-path k
Ic_i^K =Entropy of node i based on meta path instances in meta-path k
C_{ij}^K =Number of instances between nodes i and j in meta-path k
Create Graph G(V,E) from dataset:

(*continued*)

Algorithm 1. Finding Influential nodes

1. Calculate entropy based on Neighbors

For each Meta-path K calculates P_i

$$P^K_{A_1A_2}(t)=\frac{1}{N^K A_1(t)}, P^{K_1}_{A_2}(t)=\sum_{E\in N^{K_1}_{A_2}(t)} P^{K_1}_{EA_2}(t)$$

$$P^{K_1}_{A_1A_2}(t)=\sum_{paths\ from\ A_1\ to\ A-l} P^{K_1}_{A_2}(t)\times\cdots\times P^{K_{l-1}}_{A_l}(t)$$

$$If^K_i(t)=\sum_{j\in N_i} P^K_{ij}(t)\log_{10}P^K_{ij}(t)$$

$$Info^K_{Fi}=\frac{If^K_i(t)}{\sum_K\sum_{q=1}^{N^K_i(t)}}$$

2. Calculate Entropy Based on neighbors and meta path instances

$$P^K_{A_1A_2}(t)=\frac{C^K_{A_1A_2}(t)}{\sum_{E\in N^K_{A_1}A} C^K_{A_1E}(t)}, P^{K_1}_{A_2}(t)=\sum_{E\in N^{K_1}_{A_2}(t)} P^{K_1}_{EA_2}(t)$$

$$P^{K_1}_{A_1A_2}(t)=\sum_{paths\ from\ A_1\ to\ A-l} P^{K_1}_{A_2}(t)\times\ldots\times P^{K_{l-1}}_{A_l}(t)$$

$$Im^K_i(t)=-\sum_{j\in N_i} Pr^k_{ij}(t)\log_{10}Pr^k_{ij}(t)$$

$$Info^K_{mi}=\frac{Im^K_i(t)}{\sum_K\sum_{q=1}^{N^K_i}C^K_{iq}(t)}$$

$$Info_{Mi}=\sum_K Info^k_{mi}(t)$$

3. Calculate Entropy Based on meta path instances

$$Pc^K_{A_1A_2}(t)=\frac{C^K_{A_1A_2}(t)}{\sum_{E\in N^K_{A_1}A} C^K_{A_1E}(t)}$$

$$Ic^K_i(t)=-\sum_{j\in N_i} Pc^k_{ij}(t)\log_{10}Pc^k_{ij}(t)$$

$$Info^K_{ci}=\frac{Ic^K_i(t)}{\sum_K\sum_{q=1}^{N^K_i}C^K_{iq}(t)}$$

$$Info_{ci}=\sum_K Info^k_{ci}(t)$$

4. Calculate Final Entropy for All meta paths

$$PF_i(t)=\alpha Info_{fi}(t)+\beta Info_{ci}(t)+\gamma Info_{mi}(t)$$

2.1. Entropy locating through neighbor nodes

2.1.1. Entropy locating based on neighbors

In Eq. (1), P_i should be replaced with probability of reaching node i neighbors individually. As for single-mediator meta-path, $\frac{1}{N^K_i(t)}$ in which $N^K_i(t)$ is the number of neighboring nodes in K meta-path. In case of multi-mediator meta-paths, however, the probability of reaching node i neighbors is calculated differently; first the node i neighbors must be identified. Generally, A_1 node is considered A_l neighbor in meta-path $A_1R_1A_2R_2A_3\ldots A_{l-1}R_{l-1}A_l$. The probability of each of these neighbors should be calculated and replaced with pi. Therefore, each n-mediator meta-path is considered as n single-mediator meta-path. In fact, we assume that meta-path $A_1PA_2PA_3$ has been broken into A_1PA_2 and A_2PA_3. In other words, meta-path $A_1R_1A_2R_2A_3\ldots A_{l-1}R_{l-1}A_l$ is separated into

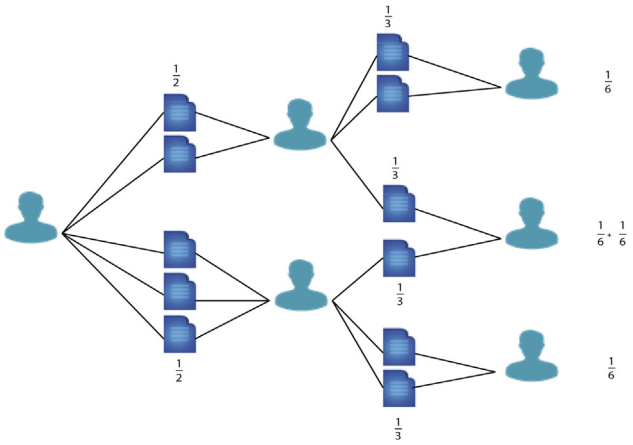


Fig. 2. Calculating probability with respect to neighbors.

$l-1$ single-mediator meta-paths. At each stage, probability is calculated individually for each meta-path until A_l is achieved. These $l-1$ single-mediator meta-paths are respectively called K_1 to K_{l-1}. At each stage, the probability of the two nodes that go to the same node, or have a common neighbor, will be combined. At this stage, the nodes are calculated based on the neighbors and not the number of meta-path instances, as is simply shown in Fig. 2:

As shown in Fig. 2, the probability of reaching the neighbors of each node i is calculated only based on the number of neighbors and not the number of meta-paths. For each general meta-path such as $A_1R_1A_2R_2A_3\ldots A_{l-1}R_{l-1}A_l$, the probability sampling for each node to reach from A_1 to A_l is displayed in Eq. (1):

$$P^K_{A_1A_2}(t)=\frac{1}{N^K A_1(t)}P^{K_1}_{A_2}(t)=\sum_{E\in N^{K_1}_{A_2}(t)}P^{K_1}_{EA_2}(t)P^{K_1}_{A_1A_2}(t)$$
$$=\sum_{paths\ from\ A_1\ to\ A-l}P^{K_1}_{A_2}(t)\times\cdots\times P^{K_{l-1}}_{A_l}(t) \tag{1}$$

As shown in Eq. (2), for each node i in meta-path K, P_i is replaced by $P^K_{iN}(t)$ for node i neighbors in entropy Equation, i.e.N_i.

$$If^K_i(t)=\sum_{j\in N_i}P^K_{ij}(t)\log_{10}P^K_{ij}(t) \tag{2}$$

Here,$Info^K_{Fi}(t)$ is considered at t and in meta-path K to obtain friendship between author i and its neighbors. We obtain the entropy of given single and n-mediator meta-paths and finally insert them in the following Equation. Each meta-path is assigned a weight depending on its paths.

$$Info^K_{Fi}=\frac{If^K_i(t)}{\sum_K\sum_{q=1}^{N^K_i(t)}C^K_{iq}(t)} \tag{3}$$

Normalization is achieved through dividing $If^K_i(t)$ by $\sum_K\sum_{q=1}^{N^K_i(t)}C^K_{iq}(t)$. In Eq. (3), $Info^K_{Fi}(t)$ is considered between author i and its neighbors at t and in meta-path K to calculate the entropy based on the number of neighbors.

$$Info_{Fi}(t)=\sum_K Info^K_{Fi}(t) \tag{4}$$

In Eq. (4) for all the single-mediator $Info^K_{Fi}(t)$ is meta-paths the friendship between author i and its neighbors at t.

2.1.2. Locating entropy with respect to neighbors and meta-path instances

At this stage, the procedure is similar to before subSection 2.1.1, except here the probability is calculated based on the number of instances with respect to each meta-path.

As shown in Fig. 3, the probability of reaching from one node to another is calculated based on the number of neighbors as well as the number of connected instances between the two nodes. In each meta-path, K_1 to K_{l-1} for reaching from each A_1 node to A_l node can be calculated as follows:

$$P^K_{A_1A_2}(t)=\frac{C^K_{A_1A_2}(t)}{\sum_{E\in N^K_{A_1}A}C^K_{A_1E}(t)}P^{K_1}_{A_2}(t)=\sum_{E\in N^{K_1}_{A_2}(t)}P^{K_1}_{EA_2}(t)P^{K_1}_{A_1A_2}(t)$$
$$=\sum_{paths\ from\ A_1\ to\ A-l}P^{K_1}_{A_2}(t)\times\cdots\times P^{K_{l-1}}_{A_l}(t) \tag{5}$$

$$Im^K_i(t)=-\sum_{j\in N_i}Pr^k_{ij}(t)\log_{10}Pr^k_{ij}(t) \tag{6}$$

$$Info^K_{mi}=\frac{Im^K_i(t)}{\sum_K\sum_{q=1}^{N^K_i}C^K_{iq}(t)} \tag{7}$$

Fig. 3. Calculating probability with respect to neighbors and meta-path instances.

Table 1
A solved example of proposed method.

–	APA			APAPA		
Nodes of graph based on Fig. 3	P	Pr	C	P	Pr	C
Author1- Author2	1/3	3 × 1/6	3	1/3 × 1/5	3 × (1/6 × 1/8)	3
Author1- Author3	1/3	2 × 1/6	2	1/3 × 1/5	1/6 × 1/8	1
Author1- Author4	1/3	1/6	1	2 × (1/3 × 1/2)	9 × (1/6 × 1/6)+2 × (1/6 × 1/3)	11
Author1- Author5	–	–	–	1/3 × 1/5	2 × (1/6 × 1/8)	2
Author1- Author6	–	–	–	1/3 × 1/5	1/6 × 1/8	1
Author2- Author3	–	–	–	1/2 × (1/3 + 1/5)	6 × (1/6 × 1/6)+ 3 × (1/6 × 1/8)	9
Author2- Author4	1/2	3 × 1/6	3	1/2 × 1/3	3 × (1/6 × 1/6)	3
Author2- Author5	–	–	–	1/2 × 1/5	6 × (1/6 × 1/8)	6
Author2- Author6	–	–	–	1/2 × 1/5	3 × (1/6 × 1/8)	3
Author3- Author4	1/2	1/3	1	1/2 × 1/3	2 × (1/3 × 1/6)	2
Author3- Author5	–	–	–	1/2 × 1/5	2 × (1/3 × 1/8)	2
Author3-Author6	–	–	–	1/2 × 1/5	1/3 × 1/8	1
Author4- Author5	1/5	2 × 1/8	2	1/5 × 1/2	3 × (1/8 × 1/4)	3
Author4- Author6	1/5	1/8	1	1/5 × 1/2	6 × (1/8 × 1/5)	6
Author5- Author6	1/2	3 × 1/5	3	1/2 × 1/5	2 × (1/5 × 1/8)	2

Table 2
Informations of DataSets.

Dataset	#Nodes	#Edges
DBLP	215,222	611,542
ACM	468,114	1,862,602
Yelp	1282	30,838

$$Info_{Mi} = \sum_{K} Info^{k}_{mi}(t) \tag{8}$$

2.1.3. Entropy based on the number of meta-path instances

Here, the number of instances between two nodes in each meta-path plays an important role in determining the entropy. The interaction entropy for each node i is equal to $\frac{C^K_{ij}(t)}{\sum_{q=1}^{N^K_i(t)} C^K_{iq}(t)}$ divided by total passing instances through node i in each meta-path, i.e. is inserted into the entropy Equation. In this stage, the total number of meta-path instances is included in the Equation without separating each meta-path.

$$Pc^K_{A_1A_2}(t) = \frac{C^K_{A_1A_2}(t)}{\sum_{E \in N^K_{A_1}A} C^K_{A_1E}(t)} \tag{9}$$

$$Ic^K_i(t) = -\sum_{j \in N_i} Pc^k_{ij}(t) \log_{10} Pc^k_{ij}(t) \tag{10}$$

These calculations apply to each meta-path. The Info can be used for all meta-paths. The main Equation is as follows:

$$Info^K_{ci} = \frac{Ic^K_i(t)}{\sum_K \sum_{q=1}^{N^K_i} C^K_{iq}(t)} \tag{11}$$

$$Info_{ci} = \sum_{K} Info^{k}_{ci}(t) \tag{12}$$

However, $N^K_i(t)$ should be calculated based on the same meta-path. For example, in meta-path, $A_1PA_2PA_3, A_1$ through one mediator node neighbors A_3 and in A_2PA_3, $A2$ neighbors $A3$. (see Tables 1 and 2).

2.1.4. Final influence on each node

The influence of each node can be either the sum influence of all three methods or a coefficient should be assigned to each:

$$PF_i(t) = \alpha Info_{fi}(t) + \beta Info_{ci}(t) + \gamma Info_{mi}(t) \tag{13}$$

where $\alpha + \beta + \gamma$ should equal 1.

2.2. Solved sample example

This example is solved based on Fig. 3, we considered all of nodes in this figure and then calculated probabilities between all connected nodes due to the certain meta-paths. The calculation steps for node are as follow:

$K = APA$:

$$If^{APA}_{Author1} = -\sum_{j \in N_i} P^{APA}_{Author1,j}(t) \log_{10} P^{APA}_{Author1,j}(t) =$$

$$-\sum_{j \in Author2,Author3,Author4} \tfrac{1}{3} \log_{10} \tfrac{1}{3} = 0.47$$

$$Im^{APA}_{Author1} = -\sum_{j \in N_i} Pr^{APA}_{Author1,j}(t) \log_{10} Pr^{APA}_{Author1,j}(t) =$$

$$\tfrac{3}{6} \log_{10} \tfrac{3}{6} + \tfrac{2}{6} \log_{10} \tfrac{2}{6} + \tfrac{1}{6} \log_{10} \tfrac{1}{6} = 0.43$$

$$Ic^{APA}_{Author1} = -\sum_{j \in N_i} Pc^{APA}_{Author1,j}(t) \log_{10} Pc^{APA}_{Author1,j}(t) =$$

$$\tfrac{3}{6} \log_{10} \tfrac{3}{6} + \tfrac{2}{6} \log_{10} \tfrac{2}{6} + \tfrac{1}{6} \log_{10} \tfrac{1}{6} = 0.43$$

$j \in \{Author2, Author3, Author4, Author5, Author6\}$

$$If^{APAPA}_{Author1} = -\sum_{j} P^{APAPA}_{Author1,j}(t) \log_{10} P^{APAPA}_{Author1,j}(t) =$$

$$P^{APAPA}_{Author1,j}(t) \log_{10} P^{APAPA}_{Author1,j}(t) = -\left(\left(\sum_{1}^{4} \tfrac{1}{15} \log_{10} \tfrac{1}{15}\right) + \tfrac{2}{6} \log_{10} \tfrac{2}{6}\right) = 0.47$$

$$Im^{APAPA}_{Author1} = -\sum_{j} Pr^{APAPA}_{Author1,j}(t) \log_{10} Pr^{APAPA}_{Author1,j}(t) =$$

$$Pr^{APAPA}_{Author1,j}(t) \log_{10} Pr^{APAPA}_{Author1,j}(t) =$$

$$-\left(\tfrac{3}{48} \log_{10} \tfrac{3}{48} + 2\,\tfrac{1}{48} \log_{10} \tfrac{1}{48} + \tfrac{11}{36} \log_{10} \tfrac{11}{36} + \tfrac{2}{48} \log_{10} \tfrac{2}{48}\right) = 0.36$$

$$Ic^{APAPA}_{Author1} = -\sum_{j} Pr^{APAPA}_{Author1,j}(t) \log_{10} Pc^{APAPA}_{Author1,j}(t) =$$

$$Pc^{APAPA}_{Author1,j}(t) \log_{10} Pc^{APAPA}_{Author1,j}(t) =$$

$$-\left(\tfrac{3}{18} \log_{10} \tfrac{3}{18} + 2\,\tfrac{1}{18} \log_{10} \tfrac{1}{18} + \tfrac{11}{18} \log_{10} \tfrac{11}{18} + \tfrac{2}{18} \log_{10} \tfrac{2}{18}\right) = 0.5$$

3. Experimental results

3.1. DataSet

We implemented our methods on three datasets to assure that the results are reliable. These three datasets are usually used in multi-layer networks (Sun et al., 2011; Gui, Sun, Han, & Brova, 2014).

DBLP: (Citation Network, 2019) Objects act as authors. Moreover, various meta-paths including Author-Paper-Author-Paper-

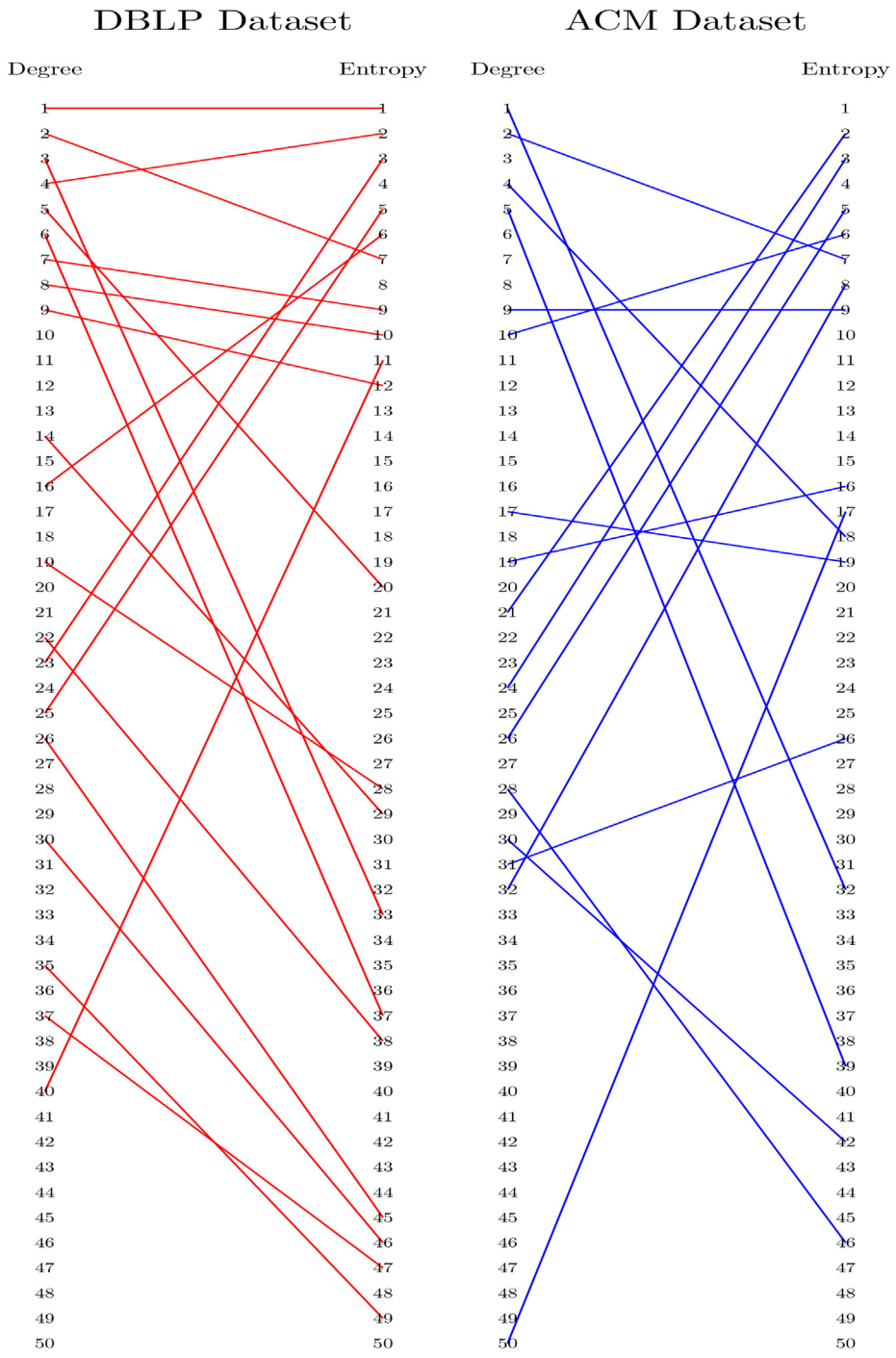


Fig. 4. Position of the top 50 identified influential nodes in the entropy-based and the degree-based methods.

Author (APAPA), Author-Paper-Author (APA), Author-Conference-Author (ACA) and Author-Conference-Author-Conference-Author (ACACA) were considered. The data is up to 2016.

ACM: (Citation Network, 2019) In this network, various meta-paths were taken into account namely APA, ACA, APAPA, and ACACA. The data is up to 2003.

Yelp: (Yelp, 2019) Objects act as users. various meta-paths were used namely User-Business-User (UBU), User-Business-Service (UBS), User-Business-Revervation (UBR), and User-Business-Category (UBC).

3.2. Evaluation criterion

The susceptible-infectious (SI) method was used for evaluation. The algorithm is described in Algorithm 2.

Furthermore, the diffusion rate was calculated using infected scale criterion (Zhang, Chen, Dong, & Zhao, 2016) as $f(t) = \frac{N_i(t)}{N}$.

Where N is the total number of nodes in the network and $N_{i(e)}$ is the number of activated nodes in different epochs indicated by t.

Algorithm 2. SI Model for All Nodes

```
Input  : Datasets and top-k influential node from pseudo code 1
Output: Number of Infected node at time t
Start
Create Graph G(V,E) from dataset
Label top-k influential nodes as infected and the others nodes as
  susceptible
Compute from pseudo code 1
For i in G.nodes:
if i==Infected: then
    For j in G.neighbors(i):
    if If j== susceptible and ≥ Threshold: then
        Label j as infected state;
    else
        Cont;
    end
else
    Cont;
end
```

3.3. Time complexity

If d is considered as the mean of n-mediator neighbors, v as the number of graph nodes, e as the number of graph edges, and k as the number of influential nodes, then the time for formation of the graph related to the given meta-path is $o(v)$, and the time complexity of pseudo-code (1) is equal to $o(ve)$ while pseudo-code 1 is $o(e + vd + vlog(k))$. Therefore, in worst case scenario, the time complexity of the algorithm equals $o(ve)$.

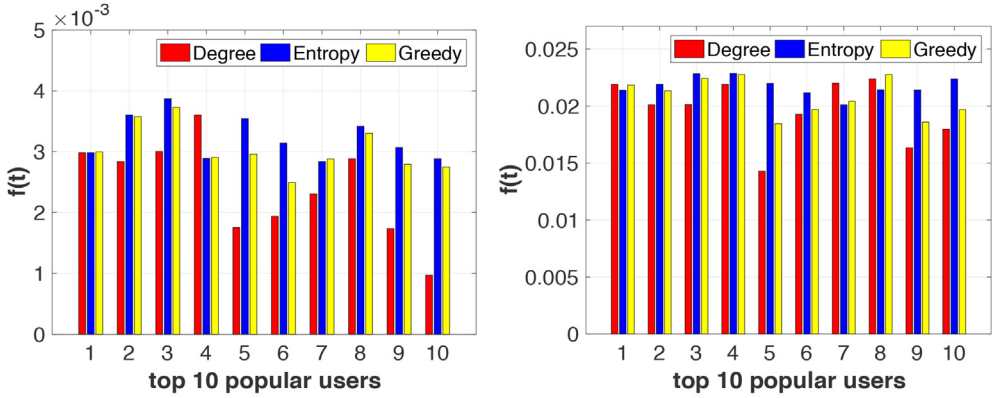
3.4. Analysis of results

The DBLP and ACM both indicated the overlapping of entropy-based and degree-based methods. In this chart, the first 50 authors were separated using both methods and displayed in order (Fig. 4). Both methods indicated overlaps and discrepancies. The information was diffused and measured to show which method outperformed the other, as follows:

3.4.1. Information diffusion from a specific node in DBLP data

The information from each node individually to calculate which node ultimately activated the highest number of nodes. We adopted the *SI* model for that purpose. Fig. 5 displays the ten most influential nodes (in epoch = 100 and 1000).

The proposed method is compared with other methods using the *infected scale* Equation. Evidently, for part a, p is considered $p = 0.001$, in which p represents the ratio and the number of spread starters, and different Ps are displayed as $epoch = 100$ for part b. Fig. 6 shows that starting from nodes obtained by entropy



(a) Methods in epoch 100 (b) Methods in epoch 1000

Fig. 5. Top-10 influential nodes on DBLP dataset.

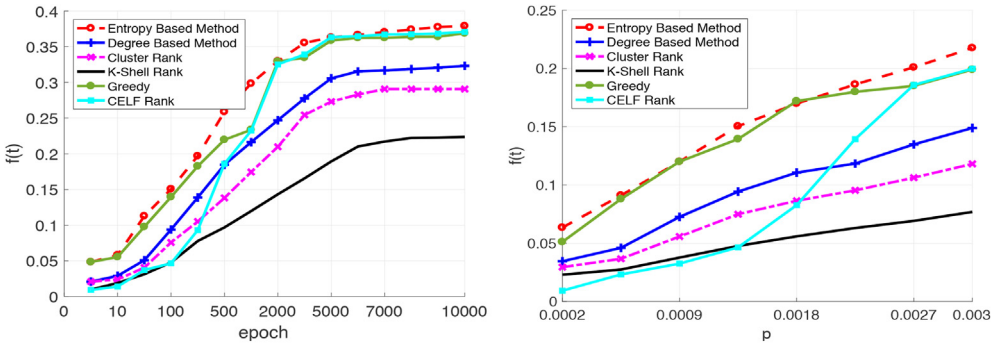


Fig. 6. Comparing of proposed method with other methods using infected scale Equation on DBLP dtataset.

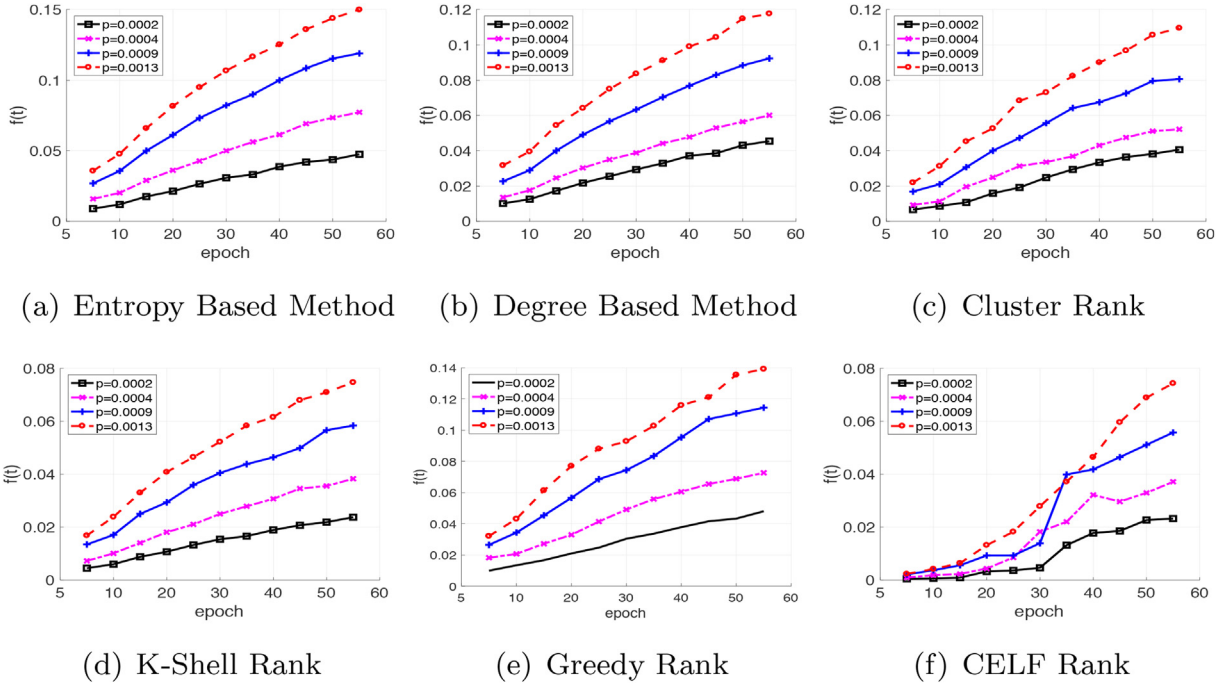


Fig. 7. Trend of the proposed Entropy diffusion method in compare to existing methods. We considered different Ps and epochs on DBLP dataset for showing diffusion speed.

results in faster diffusion and a greater number of active nodes compared to other methods.

The diffusion was compared to that of five other methods: the degree-based method, cluster rank, K-shell rank, CELF (Leskovec et al., 2007) and Greedy method (Abebe, Kleinberg, Parkes, & Tsourakakis, 2018) using different Ps and taking into account epochs 5 to 60 (Fig. 7). The proposed method outperformed the previous alternatives.

3.4.2. Information diffusion from a specific node in ACM data

In ACM data, we similarly adopted SI for information diffusion. Fig. 8 shows the nodes obtained via the proposed method versus the degree and greedy based methods. Information was diffused through each node to find out how many nodes will ultimately be activated.

As shown in Figs. 6–10 , our method improved slightly at lower epochs. At higher epochs, however, the difference intensified. The multi-mediator meta-paths included in this paper covered direct nodes as well as indirectly influenced nodes which resulted in the section of nodes with higher direct and indirect influence on neighbors. In fact, both direct and indirect neighbors were investigated. In the meantime, if a lower degree node is connected to higher influential nodes, it will be considered more significant.

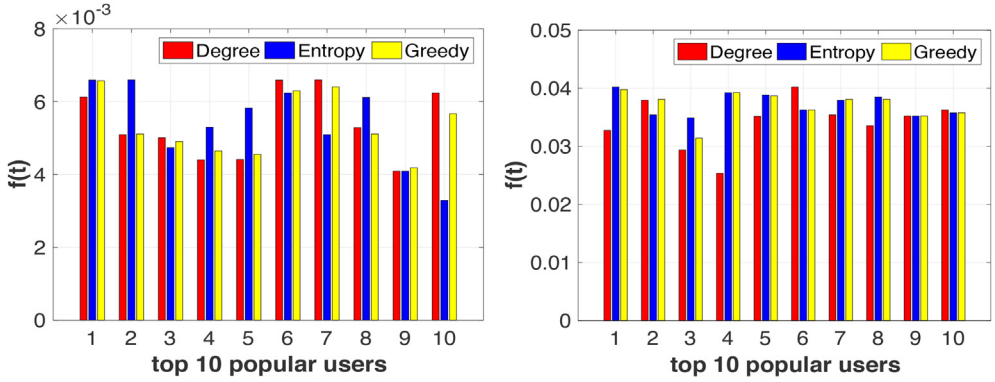
Fig. 9 shows that starting from nodes obtained by entropy, the number of active nodes is higher than that of other methods. As a result, these nodes tend to be more influential. The diffusion speed is clearly indicated in the figure.

The diffusion speed was examined using different Ps and employing different methods (Fig. 10).

3.4.3. Information diffusion from a specific node in Yelp data

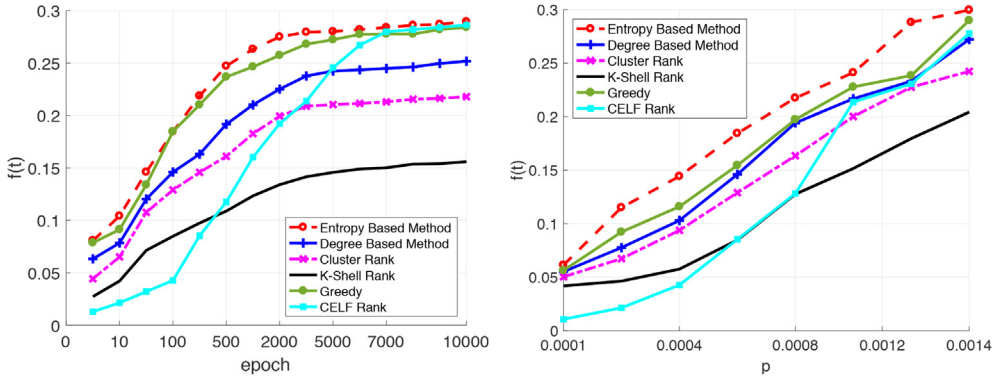
We initially show the influence of the top 5 influential nodes with using of the SI model as before to evaluate the influence on the Entropy-based method (Fig. 11).

To assess the performance of our proposed method (Entropy-based method) effectively, we compare it with four ranking methods, including Degree based, Cluster rank, k-shell, CELF, and Greedy



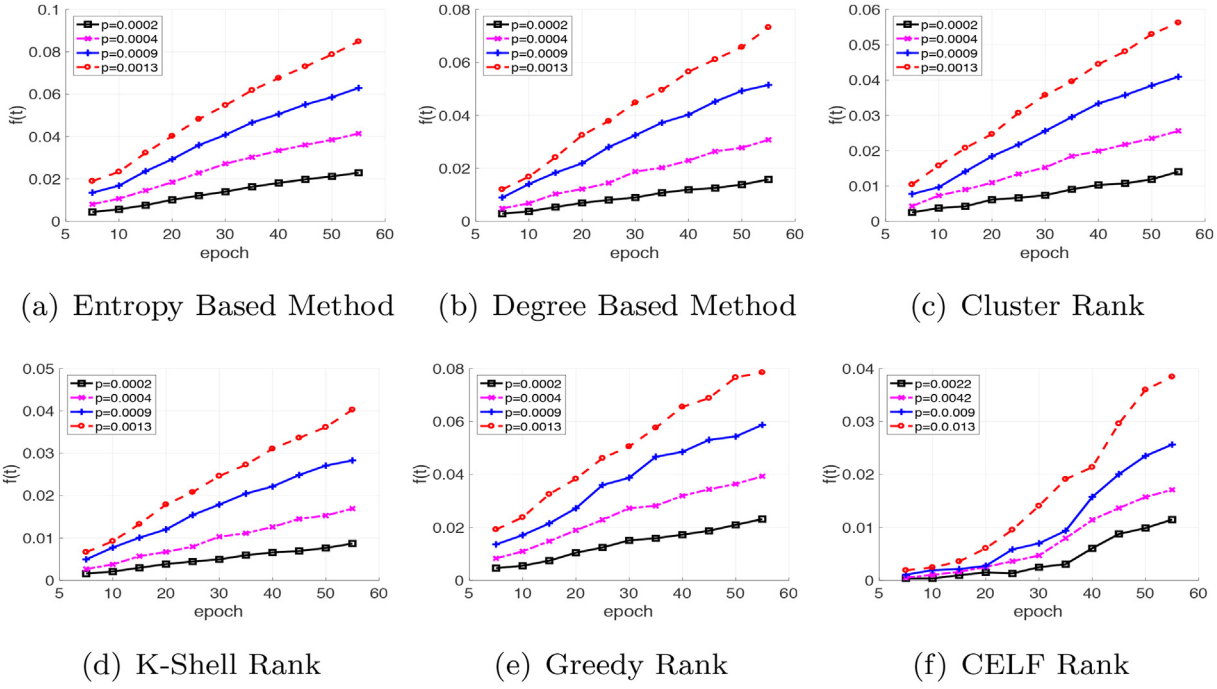
(a) Methods in epoch 100 (b) Methods in epoch 1000

Fig. 8. Top-10 influential nodes on ACM dataset.



(a) p=0.0006 (b) epoch=100

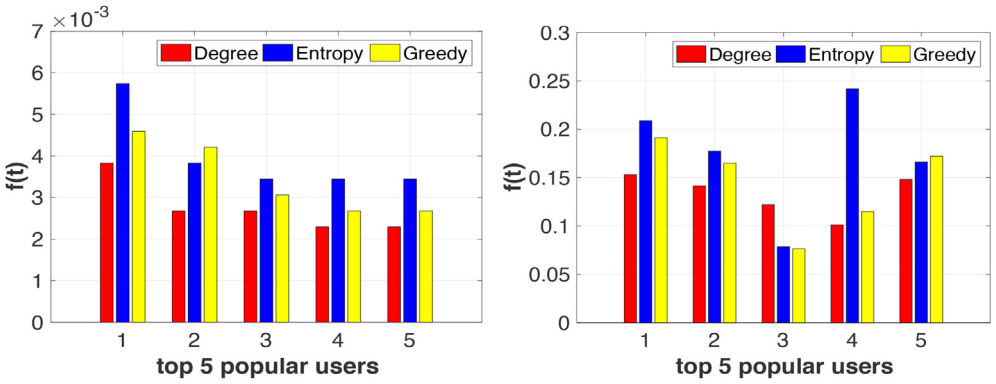
Fig. 9. Comparing the proposed method with other methods using infected scale equation on ACM dataset.



(a) Entropy Based Method (b) Degree Based Method (c) Cluster Rank

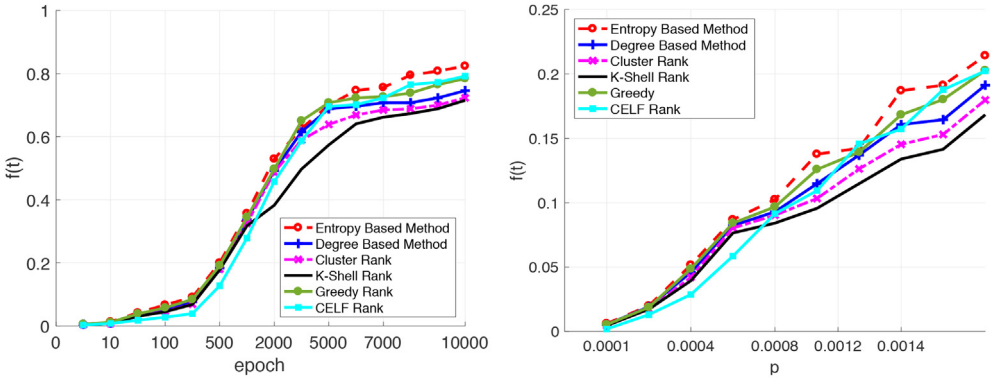
(d) K-Shell Rank (e) Greedy Rank (f) CELF Rank

Fig. 10. Trend of the proposed Entropy diffusion method in comparison to other existing methods. We considered different Ps and epochs on ACM dataset for showing diffusion speed of methods.



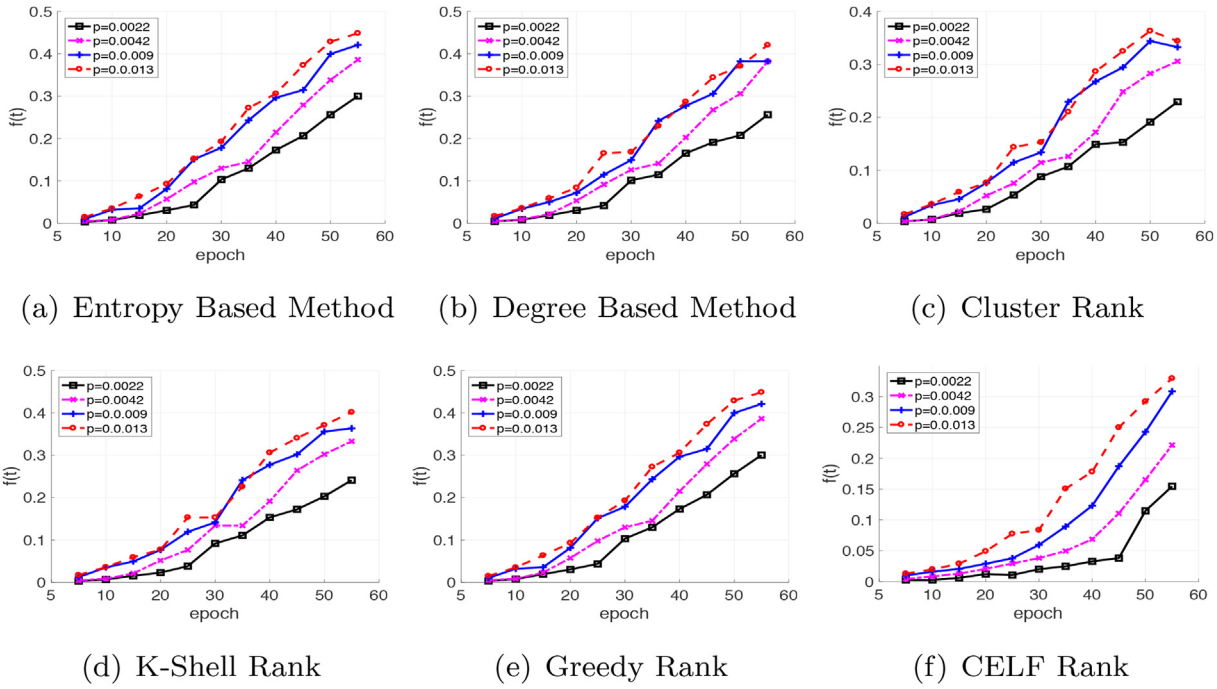
(a) Methods in epoch 100 (b) Methods in epoch 1000

Fig. 11. Top-5 influential nodes on Yelp dataset.



(a) p=0.003 (b) epoch=100

Fig. 12. Comparing the proposed method with other methods using infected scale equation on Yelp dataset.



(a) Entropy Based Method (b) Degree Based Method (c) Cluster Rank

(d) K-Shell Rank (e) Greedy Rank (f) CELF Rank

Fig. 13. Trend of the proposed Entropy diffusion method in comparison to other existing methods. We considered different Ps and epochs on Yelp dataset for showing diffusion speed of methods.

rank in Fig. 12. In this experiment, the value of p and *epoch* are 0.003 and 100 respectively.

The infected scale, $f(t)$, on the Yelp network under different methods with different infected rates are shown in Fig. 13. As the size of the initial nodes increases, the final infected scale of these methods ascends step by step basically. In both three datasets, the advantage of Entropy-based method is obvious. Our proposed method emphasizes the role of the meta-path in the network.

4. Conclusion

In this paper, a new method for locating influential nodes in heterogeneous networks is proposed based on meta-paths concept. The impact of individuals was calculated relying on meta-path instances. These relationships were calculated based on entropy analysis of direct neighbors, meta-path instances, and the combination of both. Analysis demonstrated that the proposed method offers a more effective solution than previously implemented methods.

On the basis of the real social network datasets, the experimental results have shown that the Entropy-based network performs better than Degree-based, Cluster rank, k-shell, CELF, and Greedy rank methods. The spreading power of our method is the strongest in the experiment, however, its time complexity is $O(ve)$ which in huge networks requires much time. As a future direction of this research, we aim to use heuristics methods to find the most influential nodes in the network. In fact, during feature engineering, we will be utilizing context details such as profile and text. For future endeavors, it is recommended that the characteristics of each node in addition to the graph can be used for locating influential nodes. Also, graph summarization and feature extraction can be combined with the meta-path concept. Beside them, the meta-paths and graphs have the potential to investigated by representational learning. By learning embeddings that encode graph structure, we can capture structural information about the graph that can then be used to find influential nodes in the graph.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

CRediT authorship contribution statement

Soheila Molaei: Data curation, Methodology, Software, Visualization, Writing - original draft. **Reza Farahbakhsh:** Conceptualization, Investigation, Visualization, Writing - original draft. **Mostafa Salehi:** Conceptualization, Formal analysis, Methodology, Supervision. **Noel Crespi:** Methodology, Writing - review & editing.

Acknowledgment

Mostafa Salehi was supported in part by a grant from IPM, Iran through project No. CS1399-4-162.

References

Abebe, R., Kleinberg, J., Parkes, D., & Tsourakakis, C. E. (2018). Opinion dynamics with varying susceptibility to persuasion. In *Proceedings of the 24th ACM sigkdd international conference on knowledge discovery & data mining* (pp. 1089–1098).

Althoff, T., Jindal, P., & Leskovec, J. (2017). Online actions with offline impacts. In *Proceedings of the 10th ACM international conference on web search and data mining - WSDM'17* (pp. 537–546). New York, NY, USA: ACM Press.

Bordes, A., Glorot, X., Weston, J. & Bengio, Y. (2012). Joint learning of words and meaning representations for open-text semantic parsing. In Artificial intelligence and statistics (pp. 127–135).

Bordes, A., Usunier, N., Garcia-Duran, A., Weston, J. & Yakhnenko, O. (2013). Translating embeddings for modeling multi-relational data. In *Advances in neural information processing systems* (pp. 2787–2795).

Borge-Holthoefer, J., & Moreno, Y. (2012). Absence of influential spreaders in rumor dynamics. *Physical Review E, 85*(2). 026116.

Chen, W., Wang, C., & Wang, Y. (2010). Scalable influence maximization for prevalent viral marketing in large-scale social networks. In *Proceedings of the 16th ACM sigkdd international conference on knowledge discovery and data mining - KDD'10* (pp. 1029). New York, NY, USA: ACM Press.

Citation Network Dataset. (2019). Available at: https://aminer.org/billboard/citation.

Deng, H., Han, J., Zhao, B., Yu, Y., & Lin, C. X. (2011). Probabilistic topic models with biased propagation on heterogeneous information networks. In *Proceedings of the 17th ACM sigkdd international conference on knowledge discovery and data mining* (pp. 1271–1279).

Dinh, T. N., & Thai, M. T. (2011). Precise structural vulnerability assessment via mathematical programming. In *2011-milcom 2011 military communications conference* (pp. 1351–1356).

Duan, Y., Wei, F., Zhou, M., & Shum, H.-Y. (2012). Graph-based collective classification for tweets. In *Proceedings of the 21st ACM international conference on information and knowledge management* (pp. 2323–2326).

Eesley, C., & Wang, Y. (2017). Social influence in career choice: Evidence from a randomized field experiment on entrepreneurial mentorship. *Research Policy, 46*(3), 636–650.

Eliacik, A. B., & Erdogan, N. (2018). Influential user weighted sentiment analysis on topic based microblogging community. *Expert Systems with Applications, 92*, 403–418.

Fei, L., Zhang, Q., & Deng, Y. (2018). Identifying influential nodes in complex networks based on the inverse-square law. *Physica A: Statistical Mechanics and its Applications, 512*, 1044–1059.

Flanagin, A. J. (2017). Online social influence and the convergence of mass and interpersonal communication. *Human Communication Research, 43*(4).

Fraigniaud, P., & Natale, E. (2019). Noisy rumor spreading and plurality consensus. *Distributed Computing, 32*(4), 257–276.

Gu, J., Lee, S., Saramäki, J. & Holme, P. (2017). *Ranking influential spreaders is an ill-defined problem*.

Gui, H., Sun, Y., Han, J., & Brova, G. (2014). Modeling topic diffusion in multi-relational bibliographic information networks. In *Proceedings of the 23rd ACM international conference on information and knowledge management* (pp. 649–658).

Han, J. (2009). Mining heterogeneous information networks by exploring the power of links. In *Discovery science* (pp. 13–30).

Huang, D.-W., & Yu, Z.-G. (2017). Dynamic-Sensitive centrality of nodes in temporal networks. *Scientific Reports, 7*, 41454.

Kivelä, M., Arenas, A., Barthelemy, M., Gleeson, J. P., Moreno, Y., & Porter, M. A. (2014). Multilayer networks. *Journal of Complex Networks, 2*(3), 203–271.

Kuhnle, A., Alim, M. A., Li, X., Zhang, H., & Thai, M. T. (2018). Multiplex influence maximization in online social networks with heterogeneous diffusion models. *IEEE Transactions on Computational Social Systems, 5*(2), 418–429.

Leskovec, J., Krause, A., Guestrin, C., Faloutsos, C., VanBriesen, J., & Glance, N. (2007). Cost-effective outbreak detection in networks. In *Proceedings of the 13th ACM sigkdd international conference on knowledge discovery and data mining* (pp. 420–429).

Liu, D., Jing, Y., Zhao, J., Wang, W., & Song, G. (2017). A fast and efficient algorithm for mining Top-k nodes in complex networks. *Scientific Reports, 7*, 43330.

Liu, J.-G., Lin, J.-H., Guo, Q. & Zhou, T. (2016). Locating influential nodes via dynamics-sensitive centrality. *Scientific Reports*, ***6***.

Malliaros, F. D., Rossi, M.-E. G., & Vazirgiannis, M. (2016). Locating influential nodes in complex networks. *Scientific Reports, 6*(1), 19307.

Mihalcea, R., & Tarau, P. (2004). Textrank: Bringing order into text. In *Proceedings of the 2004 conference on empirical methods in natural language processing.* .

Mo, H., & Deng, Y. (2019). Identifying node importance based on evidence theory in complex networks. *Physica A: Statistical Mechanics and its Applications, 529*. 121538.

Mohammadinejad, A., Farahbakhsh, R., & Crespi, N. (2018). Opiu: Opinion propagation in online social networks using influential users impact. In *2018 IEEE international conference on communications (ICC)* (pp. 1–7).

Molaei, S., Babaei, S., Salehi, M., & Jalili, M. (2018). Information spread and topic diffusion in heterogeneous information networks. *Scientific Reports, 8*(1), 9549.

Molaei, S., Khansari, M., Veisi, H., & Salehi, M. (2019). Predicting the spread of influenza epidemics by analyzing twitter messages. *Health and Technology*, 1–16.

Molaei, S., Zare, H., & Veisi, H. (2020). Deep learning approach on information diffusion in heterogeneous networks. *Knowledge-Based Systems, 189*. 105153.

Ohara, K., Saito, K., Kimura, M., & Motoda, H. (2020). Resampling-based predictive simulation framework of stochastic diffusion model for identifying top-k influential nodes. *International Journal of Data Science and Analytics, 9*(2), 175–195.

Page, L., Brin, S., Motwani, R. & Winograd, T. (1999). *The pagerank citation ranking: Bringing order to the web*. (Tech. Rep.). Stanford InfoLab.

Pei, S., Teng, X., Shaman, J., Morone, F., & Makse, H. A. (2017). Efficient collective influence maximization in cascading processes with first-order transitions. *Scientific Reports, 7*, 45240.

Peng, S., Yang, A., Cao, L., Yu, S., & Xie, D. (2017). Social influence modeling using information theory in mobile social networks. *Information Sciences, 379*, 146–159.

Ranjbar, V., Salehi, M., Jandaghi, P., & Jalili, M. (2019). Qanet: Tensor decomposition approach for query-based anomaly detection in heterogeneous information networks. *IEEE Transactions on Knowledge and Data Engineering, 31*(11), 2178–2189.

Raychaudhuri, A., Mallick, S., Sircar, A., & Singh, S. (2020). Identifying influential nodes based on network topology: A comparative study. In *Information, Photonics and Communication* (pp. 65–76). Springer.

Sandoval, L. (2014). Structure of a global network of financial companies based on transfer entropy. *Entropy, 16*(8), 4443–4482.

Shakya, H. B., Stafford, D. & Hughes. (2017). Exploiting social influence to magnify population-level behaviour change in maternal and child health: Study protocol for a randomised controlled trial of network targeting algorithms in rural Honduras. *BMJ Open*, **7**(3).

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal, 27*, 379–423.

Shen, Y., Nguyen, N. P., Xuan, Y., & Thai, M. T. (2012). On the discovery of critical links and nodes for assessing network vulnerability. *IEEE/ACM Transactions on Networking, 21*(3), 963–973.

Socher, R., Chen, D., Manning, C. D. & Ng, A. (2013). Reasoning with neural tensor networks for knowledge base completion. In *Advances in neural information processing systems* (pp. 926–934).

Sun, Y., & Han, J. (2012). Mining heterogeneous information networks: Principles and methodologies. *Synthesis Lectures on Data Mining and Knowledge Discovery, 3*(2), 1–159.

Sun, Y., & Han, J. (2013). Mining heterogeneous information networks: A structural analysis approach. *ACM SIGKDD Explorations Newsletter, 14*(2), 20–28.

Sun, Y., Han, J., Yan, X., Yu, P. S., & Wu, T. (2011). Pathsim: Meta path-based top-k similarity search in heterogeneous information networks. *Proceedings of the VLDB Endowment, 4*(11), 992–1003.

Sun, Y., Ma, L., Zeng, A. & Wang, W.-X. (2016). Spreading to localized targets in complex networks. *Scientific Reports*, **6**(1).

Tang, J., Zhang, R., Wang, P., Zhao, Z., Fan, L., & Liu, X. (2020). A discrete shuffled frog-leaping algorithm to identify influential nodes for influence maximization in social networks. *Knowledge-Based Systems, 187*. 104833.

Tselykh, A., Tselykh, L., Vasilev, V., & Barkovskii, S. (2018). Knowledge discovery using maximization of the spread of influence in an expert system. *Expert Systems, 35*(6). e12312.

Tselykh, A., Vasilev, V., & Tselykh, L. (2019). Managing influence in complex systems to ensure safety of their operation. In *Proceedings of the 12th international conference on security of information and networks* (pp. 1–6).

Tselykh, A., Vasilev, V., & Tselykh, L. (2020). Assessment of influence productivity in cognitive models. *Artificial Intelligence Review*, 1–27.

Vitak, J., Zube, P., Smock, A., Carr, C. T., Ellison, N., & Lampe, C. (2011). It's complicated: Facebook users' political participation in the 2008 election. *CyberPsychology, Behavior, and Social Networking, 14*(3), 107–114.

Wang, Y., Wang, S., & Deng, Y. (2019). A modified efficiency centrality to identify influential nodes in weighted networks. *Pramana, 92*(4), 68.

Wang, Z., Zhang, J., Feng, J., & Chen, Z. (2014). Knowledge graph and text jointly embedding. In *EMNLP* (pp. 1591–1601).

Wei, B., & Deng, Y. (2019). A cluster-growing dimension of complex networks: From the view of node closeness centrality. *Physica A: Statistical Mechanics and its Applications, 522*, 80–87.

Wen, T., & Deng, Y. (2020). Identification of influencers in complex networks by local information dimensionality. *Information Sciences, 512*, 549–562.

Yang, H.-X., Wang, W.-X., Lai, Y.-C., Xie, Y.-B., & Wang, B.-H. (2011). Control of epidemic spreading on complex networks by local traffic dynamics. *Physical Review E, 84*(4). 045101.

Zhang, J.-X., Chen, D.-B., Dong, Q., & Zhao, Z.-D. (2016). Identifying a set of influential spreaders in complex networks. *Scientific Reports, 6*(1), 27823.

Zhang, W., Yang, J., Ding, X.-Y., Zou, X.-M., Han, H.-Y., & Zhao, Q.-C. (2019). Groups make nodes powerful: Identifying influential nodes in social networks based on social conformity theory and community features. *Expert Systems with Applications, 125*, 249–258.

Zhu, Q., Li, L., & Gan, C. (2018). Modeling and analysis of the impact of adaptive defense strategy on virus spreading. *IAENG International Journal of Applied Mathematics, 48*(2), 1–6.

Yelp. (2019). Available at: https://github.com/librahu/HIN-Datasets-for-Recommendation-and-Network-Embedding.