Nonlinear Dynamics, Psychology, and Life Sciences, Vol. 25, No. 2, pp. 127-155.

Predicting and Correcting Missing Data on Diffusion Processes in Multiplex Networks

Alireza Khosravani, **Mostafa Salehi**[1], *University of Tehran, Tehran, Iran,* **Vahid Ranjbar**, *Yazd University, Yazd, Iran,* **Rajesh Sharma**, *University of Tartu, Tartu, Estonia* and **Shaghayegh Najari**, *University of Tehran, Tehran, Iran*

Abstract: *The diffusion process in networks is studied with the objective of identifying the dynamics and for predicting the behavior of network entities. Social media plays an important role in people's lives. Diffusion processes, as one of the most important branches of social media analysis, have their presence in various domains such as information spreading, diffusion of innovation, idea dissemination, and product acceptance to identify user's pattern and their behavior in social media networks. Users are not limited to one social network and are engaged in multiple social media such as Twitter, Instagram, Telegram, and Facebook. This fact has created new phenomena in social network analysis, called multiplex network analysis. Thus, the scope of diffusion process analysis has been transferred from single layer networks to multiplex networks. Diffusion process analysis can be studied at both infrastructure-level and diffusion-level; at infrastructure-level, the structural network's properties such as clustering coefficient and degree centrality are being studied; and in diffusion-level the diffusion network's properties such as diffusion depth and seed nodes are being studied. On the other hand, a reliable analysis requires complete information on both infrastructure and diffusion networks. However, complete data is not accessible forever, this fact is due to some limitations such as crawling big data, gathering social media policies, and user privacy. Incomplete data can lead to poor analysis, so in this work we, first of all, investigate the impact of missing data in both infrastructure and diffusion networks, the impact of random and non-random missing infrastructure data on nine diffusion network's properties such as number of infected nodes, number of infected edges, diffusion length and number of seed nodes. Secondly, based on the multiplex diffusion tree, we introduce a new model named as MLC-tree for an incomplete diffusion network. Finally, we evaluate our model on both synthetic and real social networks; these results show that the MLC-tree can decrease the relative error more than 50 percent while missing 20 to 80 percent of complete data.*

Key Words: multiplex networks, diffusion networks, missing data

[1] Correspondence address: Mostafa Salehi at Faculty of New Sciences and Technologies, University of Tehran, Tehran, Iran, E-mail: mostafa_salehi@ut.ac.ir

INTRODUCTION

Complex networks in real-world include communication systems, social systems, transportation systems, biological systems, and more. Network science studies the role of network structure in the dynamic behavior of complex networks. For example, dynamic network behaviors include cascade failures in power grids networks, diffusion information or rumor propagation among people, diffusion innovation in the social networks, the spread of epidemics, the spread of computer viruses, and the mutation of genes in biological networks to name a few. Networks are a representation of the description of relationships between system components or community members. Numerous biological, social, economic, and technological systems have subsystems, each of which can be interconnected. The main purpose of network science is to study the behavior of such systems as networks. Understanding, predicting, and controlling network risk has important implications for many real-world issues.

Diffusion of information or rumor propagation, diffusion of innovation, and influence through social networks are examples of diffusion. In the social network, studying the behavioral patterns of people in social networks, modeling of diffusion of innovation, information or rumor propagation, spreading epidemics, etc. are one of the main lines of research. Models of rumor propagation rarely include sociological and psychological aspects of real-world transmission behavior. As a result, these models tend not to incorporate the connections between people as can be seen in real social networks. The GBN-Dialogue model is used to simulate the spread of an out-group negative rumor over the RIT Facebook network (Brooks, DiFonzo, & Ross, 2013). Various models have also been developed to model the diffusion of innovation. Lopes, Navarro, and Silva (2018) used an artificial neural network to connect positive and negative emotions to an individuals' propensity to communicate their opinions through a social network–and find out negative effect spreads faster than the positive. Also Jacobsen and Guastello (2011) presented an overview of diffusion innovation models and examines the adoption S-curve, network theories, diffusion models, influence models, geographical model, and reported that the S-curves can be represented in greater detail as nonlinear dynamical processes.

Information diffusion is affected by various characteristics such as initiator diffusion nodes, the number of nodes who received information, and the path through which information travels has different behavior. Initial studies on diffusion processes have provided an analysis of diffusion on single layer networks, those are predicting the behavior of diffusion and its properties such as depth and width (Clauset, Moore, & Newman, 2008; Eagle, Pentland, & Lazer, 2009; Liben-Nowell & Kleinberg, 2008; Myers & Leskovec, 2010). However, due to the increasing number of social channels, researchers have started to investigate diffusion in multiplex networks (Babaei, Molaei, & Salehi, 2019; Dickison, Havlin, & Stanley, 2012; Dou, Du, & Song, 2016; Najari, Salehi, Ranjbar, & Jalili, 2019; Salehi et al., 2015; Saumell-Mendiola, Serrano, & Boguná, 2012; Zong, Wu, Singh, & Yan, 2012).

In the analysis of the diffusion network, the properties of the diffusion networks such as the diffusion depth, seed nodes, the number of nodes, and edges in the diffusion path are studied. The diffusion study involving multiplex social networks gets hampered sometimes due to the non-availability of complete data. For example, a study on 1.4 million users of Facebook has determined that users are raising profile privacy from 17% to 50% (Dey, Jelveh, & Ross, 2012). As a result, a good percentage of user data is not available as it gets hidden from crawlers. In other words, data collection of networks due to factors such as privacy of individuals, hardware problems and limitations imposed by the application programming interface (API), as well as the collection of information in multiple networks (multiplex networks) and the match of individuals and communication between them in different networks is a difficult task that causes data to be missed (Fatemi, Salehi, & Magnani, 2018; Fatemi, Salehi, Veisi, & Jalili, 2018; Molaei, Farahbakhsh, Salehi, & Respi, 2020; Nikmehr, Salehi, & Jalili, 2019).

Unavailability of complete data of diffusion networks can lead to deviations of the result of analysis from reality. As an example, in a cascade of diffusion, if the information of an effective person is not available, it is possible an important or most central person is accounted for as a less important person. So far, various studies have been done for analyzing missing data on single layer networks (Bliss, Danforth, & Dodds, 2014; Brewer & Webster, 2000; Handcock & Gile, 2010); also, several studies have been accomplished on missing data and their impact on the results of the study on single layer network category (Eslami, Rabiee, & Salehi, 2011; Gomez-Rodriguez, Leskovec, & Krause, 2012; Salehi, Rabiee, & Rajabi, 2012). For example, the models presented in Eslami et al. (2011) attempted to rebuild the complete diffusion network considering incomplete diffusion network, which is explained in the preliminaries section The results of these studies are single layer diffusion network which has similar properties to complete diffusion network.

It is difficult to collect data in multiplex networks and to adapt node (individuals) and edge (communication) between them in different layers (networks), which make missing data possible. Missing data in the multiplex network makes it possible to study the properties of diffusion in multiplex networks with various errors. The main question now is that the missing data affects the properties of diffusion at multiplex networks. In other words, how an error can measure when a multiplex network is associated with missing data. This is an important question that has been studied in this paper. In particular, this paper pursues the following two objectives:

Firstly, we investigate the impact of missing data on the properties of multiplex diffusion networks by taking two random and two non-random missing data strategies. The selection of these four strategies for removing the data is based on Sharma, Magnani, and Montesi (2016), in which it has been shown that these are the four most likely reasons for missing data in multiplex networks. In this article, missing data is applied to the structure of multiplex networks and the impact of missing data in the structure of a multiplex network is evaluated on the diffusion process.

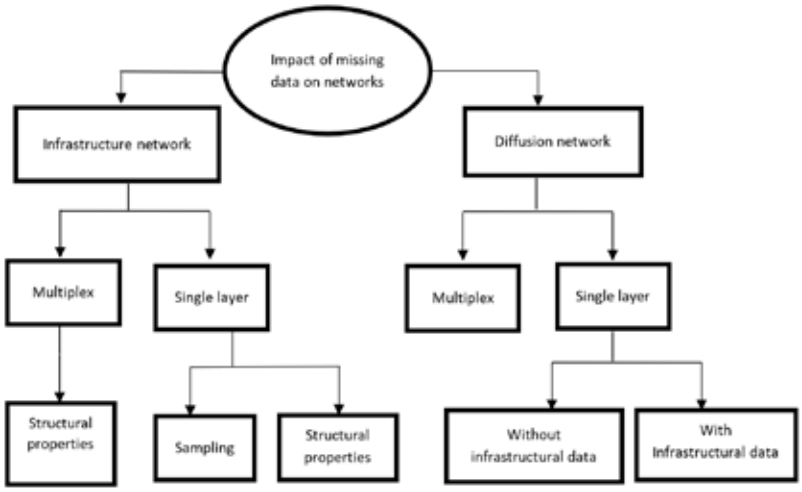

Fig. 1. Impact of missing data on networks' structure.

Secondly, we present an analytical model, MLC-tree (multiplex cascade tree), for estimating properties of a complete diffusion network to decrease the relative error of diffusion properties by only using an incomplete multiplex diffusion network. The model uses seven properties of incomplete diffusion network and extracts three parameters required by it; then we create a multiplex diffusion tree contained properties similar to the complete diffusion network properties. This model estimates the complete diffusion properties only from incomplete diffusion network, without any data from the infrastructure network. In order to evaluate the performance of the model, we compare the estimated properties by our model with the properties of a complete diffusion network. The proposed model can estimate the properties of a complete diffusion network with high precision even when the percentage of missing data from the network is high (90% missing data). The contribution of our work is summarized as:

1. We study the effect of missing data on the properties of multiplex diffusion networks by taking two random and two non-random missing data strategies.
2. We present the MLC-tree model to estimate accurately the complete diffusion network's properties by using an incomplete diffusion multiplex network.
3. We create a multiplex diffusion tree that has similar properties to complete diffusion network properties.
4. We examine the impact of four missing data strategies in the infrastructure network on nine properties of the diffusion network.

The rest of the article is structured as follows: The preliminaries section reviews the previous work. Next, in the research methodology section, we

describe our proposed model (MLC-tree) along with the discussion on the impact of four different missing strategies on the properties of the diffusion network. In the datasets section, the dataset of the multiplex network and the diffusion model in the multiplex network are discussed. In the results section, the results of the MLC-tree model are evaluated, finally, in the conclusion section, conclusions and future works are presented.

PRELIMINARIES

In general, the analysis of networks can be divided into two areas, infrastructure networks and diffusion network, also studies on missing data categorize based this two groups as infrastructure network analysis and diffusion network analysis. An overview of the impact of missing data on the structure of networks appears in Fig. 1.

Missing Data on Infrastructure Network

Initial studies in this category are based on sampling real data and evaluating impact of missing data on infrastructure network (Bliss et al., 2014; Brewer & Webster, 2000; Gjoka, Kurant, & Markopoulou, 2013; Handcock & Gile, 2010). These studies have proposed some sampling methods in those nodes and edges are sampled so that show the properties of infrastructure network. Some of the studies (Eslami et al., 2011; Feng et al., 2018; Gomez-Rodriguez et al., 2012; Zong et al., 2012) specified the impact of missing data on usual network's properties such as the degree distribution, nodes centrality, clustering coefficient and so on. While some other studies (Cai, Wang, Cui, & Stanley, 2018; Sharma et al., 2016) are presenting models that can retrieve complete networks from the incomplete networks. To evaluate the predicting models, they compared the retrieved network with the actual ground truth network.

Missing Data on Diffusion Network

Studies on this topic can be categorized into two groups: inferring incomplete diffusion network with infrastructure network, and inferring diffusion network without infrastructure network.

Incomplete Diffusion Network with Infrastructure Network

Studies in this category are based on infrastructure networks. In these works, accessibility of initial infected nodes (seed) and time of infection are considered, and communications between infected nodes (e.g., in the diffusion information process, the node that information received to it, called an infected node) are not known. In other words, in these studies, the path of diffusion and infected edges (e.g., in the diffusion information process, a connection between two nodes that information transmitted through it, called infected edge) are unknown, thus the goal of these works is about estimating these paths by using available data (Duong, Wellman, & Singh, 2011; Eyal, Rosenfeld, Sina, & Kraus, 2014; Gjoka, Kurant, Butts, & Markopoulou, 2010; Kim & Leskovec, 2011; Maeno, 2007; Sina, Rosenfeld, & Kraus, 2015; Zhao, Wang, Lui, Towsley, & Guan, 2019).

Diffusion Network without Infrastructure Network

Some of these works tried to retrieve the diffusion network without the knowledge of the infrastructure network, only by using the information about diffusion network, such as infected nodes and their transmission time. In these studies, attention has been paid to the difficulty of collecting network data. Researchers in (Gomez-Rodriguez et al., 2012) are used random walk based models for predicting diffusion cascades and time of infection transmission. This is known as the first study on the Inference diffusion network without using the data of the infrastructure network. The basic assumption is that all network nodes and their infection time are available. In this study, the network has been retrieved only by accessing nodes information and cascades that have been spread on the diffusion network. In Eslami et al. (2011), the DNE model is presented; this model, without infrastructure network and only with information of infected nodes and time of infection of each node the diffusion network was retrieved.

Another category of studies related to this area is about diffusion maximization. In these studies, infrastructure data and diffusion data are incomplete. This means that also the number of nodes in the diffusion network and time of infected nodes is not available. In these studies, have attempted to provide models for the estimated properties of the diffusion network. In (Sadikov, Medina, Leskovec, & Garcia-Molina, 2011), the K_tree model and in (Belák, Mashhadi, Sala, & Morrison, 2016) (SiCE) and (ReCE) which are based on the assumption that network infrastructure data is not available, and the diffusion data (number and infection time of nodes) is also incomplete, models are presented to estimate complete diffusion network with properties of incomplete diffusion network.

So far, missing data in diffusion on multiplex networks has not been addressed except in (Sadikov et al., 2011), in which the authors proposed four types of missing data that can occur as follows: (a) Missing random node: some of the nodes are randomly removed from the infrastructure network; (b) missing random edge: some of the edges are randomly removed from the infrastructure network; (c) missing low degree nodes: low degree nodes are removed from the infrastructure network; and (d) missing low betweenness centrality nodes: low betweenness centrality nodes are removed from the infrastructure network.

To better understand the problem, let us first introduce the notions used for multiplex networks and formally define the missing data problem in them using a toy example. Figure 2 illustrates an example of missing nodes by random in a multiplex network and its impact on the average cascade length, seed, and infected node in the diffusion network. Figure (2a) shows a multiplex network with complete data, each node in each layer has a different level of communications with other nodes. For example, node 1 in Layer 1 is associated with nodes 2, 3, 5, and 7, and is connected with only nodes 2 and 5 in the Layer 2. Note this example is small, but in reality, social networks have a large number of users, and missing data rates and errors are much higher.

Figure 2b shows the same network with missing node 5 as in Fig. 2a, and its communications are not available in both layers. Consider the process of information diffusion on this network (Fig. 3). Figure3a and 3b show a particular instance of diffusion in multiplex networks for complete (Fig. 2a) and incomplete network (Fig. 2b). Figure 4a shows the diffusion network of the same networks (Figs. 2 and 3) in the form of a tree. It can be observed, there is only one node as a seed node and the numbers of infected nodes are 8. The average cascade length is 2.66. Figure 4b, shows the diffusion tree due to the loss of node 5 in the diffusion network. The diffusion process starts from node V1 at layer 1 (seed node). The average cascade length is equal to 2.33.

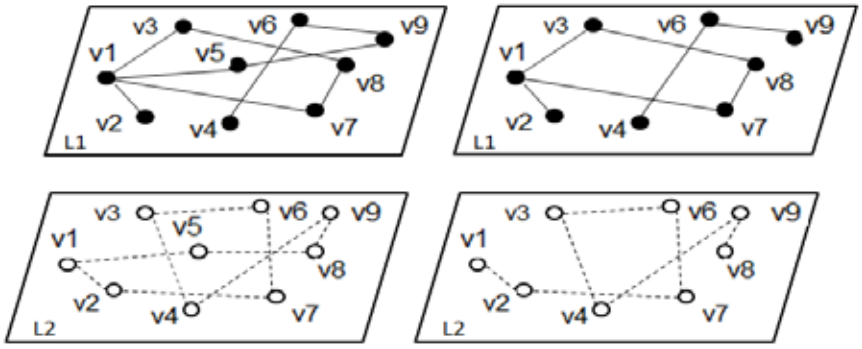

Fig. 2. Multiplex infrastructure network (a) complete network and (b) incomplete network (data of node 5 is missed).

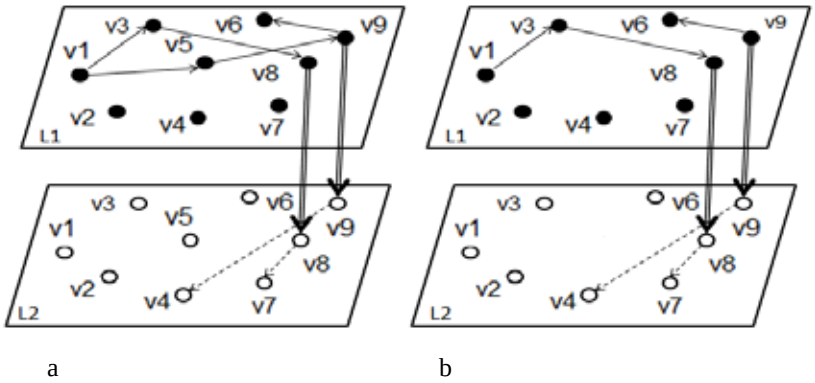

Fig. 3. Diffusion network of multiplex infrastructure networks (a) complete multiplex diffusion network and (b) incomplete multiplex diffusion network (data of node 5 is missing).

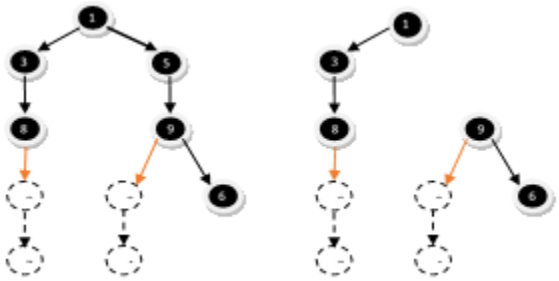

Fig. 4. Diffusion tree of diffusion in multiplex networks a) complete diffusion tree and b) incomplete diffusion tree (data of node 5 is missing).

Research Methodology

In this paper, by using the four different data removal strategies (introduced in the previous section), evaluate the impact of missing data on eight properties of diffusion networks: (a) the number of infected nodes in the first layer 2, (b) the number of infected nodes in the second layer 3, (c) the total number of infected nodes in the multiplex network, (d) the number of infected edges between the layers, (e) the number of infected edges in first layer, (f) the number of infected edges in second layer, (g) the number of non-leaf nodes, and (h) the length of the diffusion network 9. The number of seed nodes of diffusion.

Figure 5 provides an overview of our methodology to understand the effect of missing data on the diffusion process in a multiplex network. In the first step, synthetic and real networks are considered as an infrastructure network. In step 2, an Independent Cascade Model (ICM) (Goldenberg, Libai, & Muller, 2001) is used to simulate the diffusion process on multiplex networks. In step 3, information about the number of infected nodes, infected edges, seed nodes, and diffusion lengths are extracted from the complete diffusion network. In step 4, one at a time, we apply one of the four missing data strategies, those are random node missing, random edge missing, low degree nodes missing, and low betweenness centrality missing, to multiplex infrastructure network. In step 5, the same properties that were extracted in step 3 from the complete diffusion network are extracted from the incomplete diffusion network at this step. In step 6, by comparing the extracted properties of a complete and incomplete diffusion network, the impact of each missing data on the multiplex diffusion network is measured. In step 7, by using accessible properties of the incomplete network, the required properties of MLC-tree model are extracted. In step 8, we generate an estimated multiplex cascade tree by using MLC-tree model and in the final stage, we compare the properties of the estimated multiplex cascade tree with the complete one.

Proposed Model

The goal of the proposed MLC-tree (multiplex cascade tree) model is to find a complete multiplex cascade tree, which has a set of properties equal to the

properties of a complete diffusion network. For example, number of nodes in each layer, number of edges in each layer, and the depth of diffusion. However, we do not have access to a complete diffusion network, we only have access to an incomplete diffusion network. Thus, we can only compute the properties of the incomplete network. Note that properties of incomplete diffusion network can be very different from the properties of a complete diffusion network. Figure 6 shows our approach. To estimate the properties of the complete diffusion network, we first propose a complete multiplex cascade tree. The box labeled "Estimated multiplex cascade tree from MLC-tree Model" represents a parameterized family multiplex parameter. Multiplex cascade tree sampled from complete multiplex cascade tree is also shown in the box labeled "Incomplete Estimated multiplex cascade tree from MLC-tree Model". We can compute properties of incomplete multiplex cascade tree as well as properties of sampled complete multiplex cascade tree.

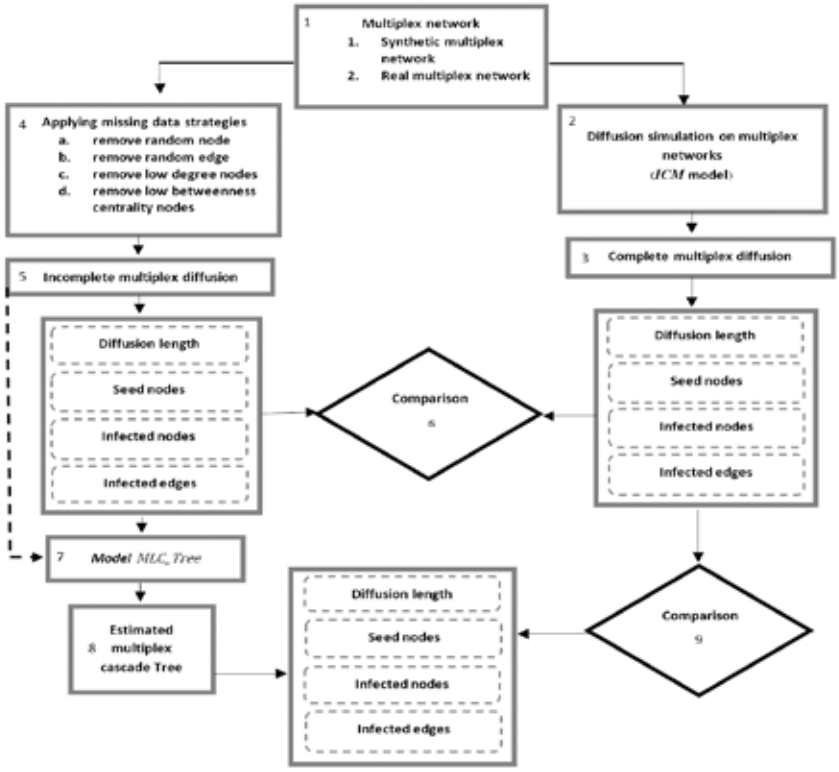

Fig. 5. Flowchart of the impact of missing data on the diffusion process in multiplex networks.

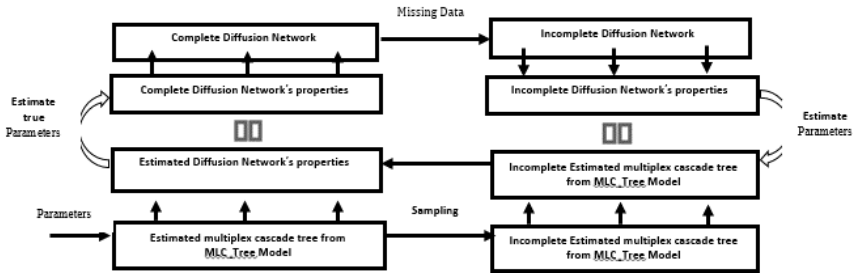

Fig. 6. MLC-tree model Implementation methodology.

Our strategy is next to find a sampled multiplex cascade tree with properties similar to the properties of an incomplete diffusion network. For example, we can find a complete multiplex cascade tree (its parameters) by finding a sampled multiplex cascade tree, with the expected number of nodes in each layer equal to the number of nodes in each layer in incomplete diffusion network. Once we find such a complete multiplex cascade tree, we can approximate the properties of a complete diffusion network by the properties of a complete multiplex cascade tree.

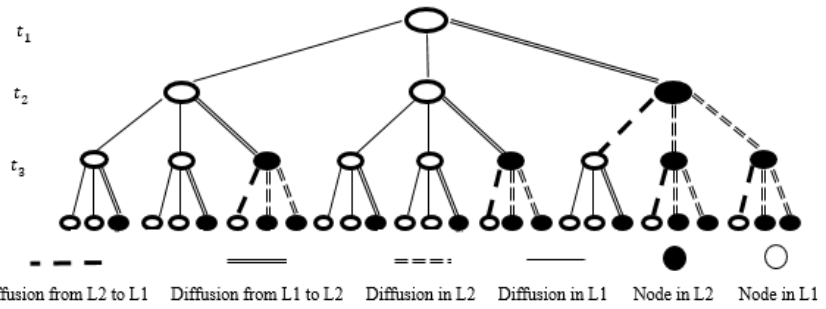

Fig. 7. Multiplex cascade tree, L1 and L2 Respectively are Layer 1 and Layer 2.

MLC-tree is a mathematical analytic model. The basic assumption is that the diffusion network has finite depth and an infected node. To generate a multiplex cascade tree, some essential parameters are needed that are described in Table 1, multiplex cascade tree with height, and a balanced branch factor will be created. To create a multiplex cascade tree, it is assumed that each infected node will only be infected through some other node, and once infected, a node will remain infected during the whole diffusion process, so that in the creation of the model, the value of parameter k=1. With k=1, a balanced multiplex cascade tree will be created.

In the next step, nodes who have received information, share information with their neighbors. Multiplex cascade tree created by MLC-tree can well model this fact. At first, diffusion will begin with the seed node, and in the next step

number of infected nodes will increase, and after several steps, the process of information diffusion stops.

Table 1. Table of symbols.

Symbol	**Definition**
n_1	number of nodes in first layer of complete multiplex cascade tree
n_2	number of nodes in second layer of complete multiplex cascade tree
m_1	number of nodes in first layer of incomplete multiplex network
m_2	number of nodes in second layer of incomplete multiplex network
b_1	number of edges between layer
b_2	number of intra layer edges
p	probability of observing a node in multiplex cascade tree
h	height of multiplex network
k	number of infected nodes' parent

When a multiplex cascade tree missing some of its information, the incomplete multiplex network can be generated, in which we can introduce properties of this new network by using Main Tree information. For example, the parameter p is the probability of observing a node in the incomplete multiplex cascade tree. We introduce other properties in Table 2.

Theorem 1. X_1, The number of infected nodes in the first layer: the number of infected nodes in the first layer of an incomplete multiplex cascade tree is obtained from Eq. 1.

$$X_1 = \frac{p}{2}\left(\frac{((b_1 + b_2)^{h+1}) - 1}{(b_2 - b_1) + 1} + \frac{((b_2 - b_1)^{h+1}) - 1}{(b_2 - b_1) - 1}\right) \quad (1)$$

Proof. The number of nodes in the first layer in a complete multiplex cascade tree (n_1) is obtained. by geometric series $n_1 = \sum_{i=1}^{h} \frac{(b_i+b_2)^i+(b_2-b_1)^i}{2}$, number of nodes in the first layer of incomplete multiplex cascade tree is $m_1 = pn_1$ which is equal to the mathematical expectation of binary random variable with parameters (n_1, p), given as Eq. 2.

$$X_1 = \frac{p}{2}\left(\frac{((b_1 + b_2)^{h+1}) - 1}{(b_2 - b_1) + 1} + \frac{((b_2 - b_1)^{h+1}) - 1}{(b_2 - b_1) - 1}\right) \quad (2)$$

Theorem 2. X_2, The number of infected nodes in the second layer: the number of infected nodes in the second layer of incomplete multiplex cascade tree is obtained from Eq. 3.

$$X_2 = \frac{p}{2}\left(\frac{((b_1 + b_2)^{h+1}) - 1}{(b_2 - b_1) + 1} - \frac{((b_2 - b_1)^{h+1}) - 1}{(b_2 - b_1) - 1}\right) \quad (3)$$

Proof. The number of nodes in the second layer in a complete multiplex cascade tree is obtained. by geometric series $n_2 = \sum_{i=1}^{h} \frac{(b_1+b_2)^i-(b_2-b_1)^2}{2}$, number of nodes in second layer of an incomplete multiplex cascade tree is $m_2 = pn_2$ which is equal to the mathematical expectation of binary random variable with parameters (n_2, p), given as Eq. 4.

$$m_2 = \frac{p}{2}\left(\frac{((b_1 + b_2)^{h+1}) - 1}{(b_2 - b_1) + 1} - \frac{((b_2 - b_1)^{h+1}) - 1}{(b_2 - b_1) - 1}\right) \quad (4)$$

Theorem 3. X_3, The number of infected edges in the first layer: In each regular multiplex cascade tree, if $k = 1$, the total number of edges is exactly equal to $(n_1 + n_2 + 1)$, which means that each non-root node has one incoming edge. In the diffusion process, each node in each layer will receive infection from one node. Thus, the number of edges in the first layer of the incomplete tree is obtained from Eq. 5.

$$X_3 = \frac{p^2 b_2}{2}\left(\frac{((b_1 + b_2)^h) - 1}{(b_2 - b_1) + 1} + \frac{((b_2 - b_1)^h) - 1}{(b_2 - b_1) - 1}\right) \quad (5)$$

Proof. If z_i represents a random variable, the number of nodes at level i in the first layer of the multiplex cascade tree and w_i represents a random variable, the number of parent nodes at level i in the first layer. The random variables (z_i and w_i) need to be independent. The number of edges in the first layer is obtained from the series $\sum_{i=1}^{h} z_i w_i$ because the mathematical expectation is a linear series (Eq. 6).

$$E\left[\sum_{i=1}^{h} z_i . w_i\right] = \sum_{i=1}^{h} E[z_i . w_i] = \sum_{i=1}^{h} E[z_i] . E[w_i] \quad (6)$$

Since z_i is a binary random variable with two parameters (p, b_2^i) and w_i is also a binary random variable with $(\min(i, k = 1), p)$ parameters. As a result, the number of edges in the first layer is obtained from Eq. 7.

$$E\left[\sum_{i=1}^{h} z_i w_i\right] = \sum_{i=1}^{h} p^2 b_2^i \min(i.k) = \frac{p^2 b_2}{2}\left(\frac{((b_1 + b_2)^h) - 1}{(b_2 - b_1) + 1} + \frac{((b_2 - b_1)^h) - 1}{(b_2 - b_1) - 1}\right) \quad (7)$$

Theorem 4. X_4, The number of infected edges in the second layer: the number of infected edges in the second layer of a complete multiplex cascade tree is obtained from Eq. 8.

$$X_4 = \frac{p^2 b_2}{2}\left(\frac{((b_1 + b_2)^h) - 1}{(b_2 - b_1) + 1} - \frac{((b_2 - b_1)^h) - 1}{(b_2 - b_1) - 1}\right) \quad (8)$$

Proof. Similar to the proof of X_3, with the difference that the role of the layer must be considered differently .In the proof of X_3, the number of nodes and parent in each level and layer i was considered for obtaining the number of edges in the first layer. But here, to prove Eq. 8, the number of nodes and parent in each level i in second layer must be considered.

Theorem 5. X_5, The number of infected edges of between layer and number of edges between layer in multiplex cascade tree obtained from Eq. 9.

$$X_5 = \frac{p^2 b_1}{2}\left(\frac{((b_1 + b_2)^h) - 1}{(b_2 - b_1) + 1}\right) \quad (9)$$

Proof. At level i, the number of edges in multiplex cascade tree is equal to $(b_1 + b_2)^i$. If z_i represents the random variable, the number of nodes in multiplex cascade tree at level $i - 1$. Number of edges between layers at level i is $b_1(b_1 + b_2)^{i-1}$ and w_i represents the random variable number of edges between layers at level i. Number of edges between layers in multiplex cascade tree is obtained from the series $\sum_{i=1}^{h} z_i . w_i$. The mathematical expectation series is linear, so $E[z_i]E[w_i] = E\left[\sum_{i=1}^{h} z_i w_i\right]$. z_i is a random variable with two parameters $((b_1 + b_2)^{i-1}, p)$ and w_i is a random variable with two parameters (b_1, p). Number of edges between layers obtained from Eq. 10.

$$E\left[\sum_{i=1}^{h} z_i . w_i\right] = \frac{p^2 b_1}{2}\left(\frac{((b_1 + b_2)^h) - 1}{(b_2 - b_1) + 1}\right) \quad (10)$$

Theorem 6. X_6, The out-degree of non-leaf nodes: out-degree of non-leaf nodes in multiplex cascade tree is obtained from Eq. 11.

$$X_6 = \frac{X_3 + X_4 + X_5}{(X_1 + X_2) - (X_L)} \quad (11)$$

Out-degree of non-leaf nodes is obtained from independent properties $X_1, X_2, \ldots, X_5$, that in this case where k is assumed to equal one, the value $(b_1 + b_2)$ is independent of h. Value $(b_1 + b_2)$ can only be calculated using this property and makes it easier to compute other properties. Out-degree of non-leaf node in multiplex cascade tree with = 1 , obtained from Eq. 12.

$$X_6 = \frac{p(b_1 + b_2)}{1 - (1 - p)^{b_1 + b_2}} \quad (12)$$

Proof. If it is assumed that properties $X_1, X_2, \ldots, X_5$ are independent, the output-degree of non-leaf nodes obtained from Eq. 11. X_L Indicates the number of leaf nodes, leaf node can be in the first or second layer.

Theorem 7. X_7, The average nodes degree: average nodes degree in multiplex cascade tree is obtained from:

$$averagedegreenode = \frac{number\ total\ edges}{number\ total\ nodes} = \frac{number\ ofedges\ in\ first\ layer + number\ of\ edges\ in\ second\ layer + number\ of\ edges\ in\ between\ layer}{number\ of\ nodes\ in\ first\ layer + number\ of\ nodes\ in\ second\ layer}$$

With assuming $X_1, X_2, \dots, X_5$ is independent, it can be approximated to the average node degree (the average degree node is equal to the number of nodes infected by each node). In a multiplex cascade tree, if assume $h \gg k$, the average degree nodes are equal to pk.

Proof. The average degree nodes in a multiplex cascade tree can be obtained from an approximation (Eq. 13), assuming the number of total edges and total number of nodes are independent. The properties of the formulae are summarized in Table 2.

$$X_7 = \frac{X_3 + X_4 + X_5}{X_1 + X_2} \approx p\ if\ h \gg k \tag{13}$$

Estimating the MLC-tree Model's Parameters

If the available data on the diffusion network is incomplete. The first goal is to find the parameters of an incomplete tree based on its properties. If the properties of the incomplete diffusion network are equal to the incomplete multiplex cascade tree, it can be concluded that properties of the complete diffusion network and properties of the complete multiplex cascade tree are equal. The parameters are estimated in two steps. In the first step, the value of the parameter p is calculated, and in the next step, parameters b_2, b_1 and h are obtained.

The first step is to calculate the parameter$p$: only the Eq. 7 can be used to compute the value of this parameter. When k=1, then the average node degree is equal to the probability of observing the node in the multiplex cascade tree.

The second step is to compute parametersb_2, b_1 and h: By putting parameter$p$ in the properties of$X_1 - X_5$, four equations (Eqs. 1-4) with three variables b_2, b_1 and h are created. Since equations are nonlinear, the system of equations cannot be solved directly. Therefore, to find the parameters, the least squares error method is used to solve each of the equations (Eqs. 1-4). Specifically, if X' is the value of each property obtained from the incomplete diffusion network and X'_m is the same as the estimated values of the model, the squared error is equal to$(X' - X'_m)^2$.

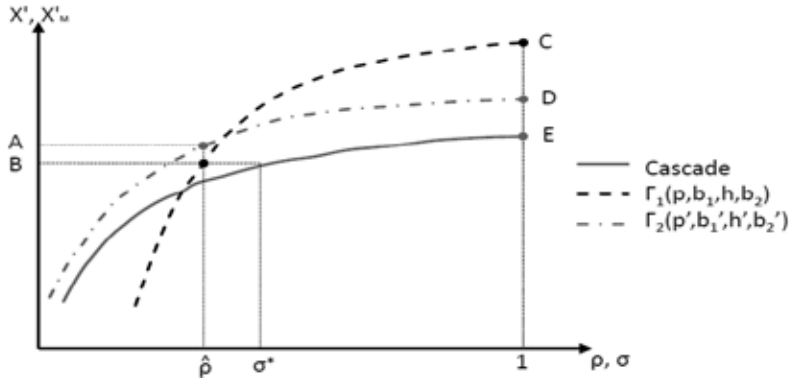

Fig. 8. Fitting a MLC-tree model. Property in two different multiplex trees made with MLC-tree model with respect to the complete diffusion.

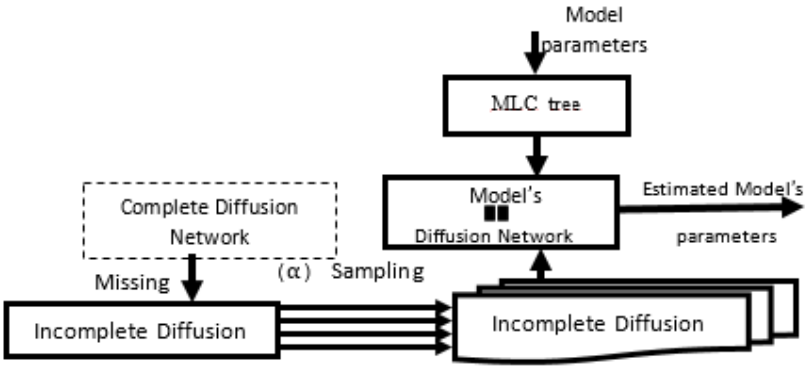

Fig. 9. Estimation of parameters of the MLC-tree model, re-sampling to find the parameters with greater precision.

Figure 8 shows the amount of a property in the diffusion network with various missing values and two alternative MLC-tree models to the observed diffusion network. The X-axis shows the missing value and Y-axis shows property value. The red curve shows the value $X'_m(\sigma)$of some property X, for example, the number of nodes in layer 1 in incomplete diffusion and black and blue (gray) line show the same property in two estimated incomplete multiplex cascade trees. If at a point such as $\hat{p}$, which indicates missing rate, an estimated incomplete diffusion tree $\Gamma 1(p, b_1, h, b_2)$ is selected as a multiplex cascade tree that has the same parameters as incomplete diffusion network. But this curve at point 1, which indicates a complete diffusion network, has a high error rate. If the curve $\Gamma 2(p, b_1, h, b_2)$, in general gives a closer approximation to the property of

the diffusion network. To obtain the extraction parameters with a higher accuracy, so that complete multiplex cascade tree with same properties as properties of complete diffusion network is created, the re-sampling method is used.

Figure 9 shows process of re-sampling and parameters estimation. At first, ten incomplete diffusion networks are sampled at intervals of [0, 1], then in each of the ten incomplete diffusion networks from the sampling, each of properties is calculated and the average measured for each parameter obtained. For each obtained sample, generate an error equation, with the squared error between the mean values calculated and considered as the values of parameters of the model estimation MLC-tree. Then, apply the least squares method to find parameters b_2, b_1 and h, that minimizes the sum of the errors.

In the compare stage, to evaluate the impact of four random and non-random missing data on nine defined properties in MLC-tree, different percentages of missing are considered, α is a missing percentage which their values are 0.1,0.2, ... ,0.9.

Independent Cascade Model (ICM)

The independent cascade model (ICM) (Goldenberg et al., 2001) is presented for simulating the information diffusion process in social networks. In this model, nodes of the network are divided into active and passive categories, the diffusion process starts with a set of active nodes. At each step, if node *u* at time *T* is activated then at time *T*+1 it has only one chance of activating its neighbor's passive nodes like *v*. Activating node *u* by node *v* considered with probability *p(u,v)*. But if node *u* can't active node *v* in time *T*+1, in the next step, then node u never has the chance to activate the node *v*.

In this work, in our ICM version, seed nodes are selected randomly and are in equal proportions in both layers. For the results to be reliable the diffusion model has been applied 40 times and every time on a new synthetic network and real network, and the average results are considered. The parameters considered for simulation of the independent cascade model are presented in Table 2.

METHOD

Datasets

We used two types of synthetic and real infrastructure networks to simulate the diffusion process. A multiplex network is considered with two layers in the simulations.

For the synthetic dataset that we used to evaluate our model, we use two-layered, three multiplex networks namely (a) Barabasi Albert - Barabasi Albert (BA-BA), (b) ErdosRenyi - ErdosRenyi (ER-ER) and (c) ErdosRenyi–Barabasi Albert (ER-BA). Information of these networks is provided in Tables 2-5.

For the real dataset, we used the Twitter-Instagram dataset presented in Solé-Ribalta, De Domenico, Gómez, and Arenas (2014) for our analysis. This dataset contains two layers of direct networks. The multiplex network of data ag-

gregation comes from two online social networks, Twitter and Instagram. The number of users (nodes) of this network is 13297 per layer. To build the network, users who had more than 21 followers on the Twitter network and 10 followers on the Instagram network have been collected. Users in the first layer (Instagram network) had fewer connections than Twitter. The real network information is shown in Table 6.

Table 2. The Parameters of the IDM used to simulate the diffusion process in four multiple networks in which L1 and L2 respectively are first layer and second layer.

Multiplex network	*BA-BA*	*ER-ER*	*ER-BA*	*Tw-INS*
Total infected nodes	8570	9370	9060	19070
Infected nodes in L1	4220	4610	4320	10950
Infected nodes in L2	4350	4760	4320	8120
Repeat diffusion	40	40	40	40
Probability of transition in L2	0.2	0.2	0.2	0.2
Probability of between layer transition	0.1	0.1	0.1	0.1

Note: BA-BA: Barabasi Albert-Barabasi Albert, ER-ER: ErdosRenyi-ErdosRenyi, ER-BA: ErdosRenyi-Barabasi Albert, TW-INS: Twitter-Instagram.

Table 3. Information for the Directed Barabasi Albert - Barabasi Albert (BA-BA) Multiplex Network.

BA-BA network	*Simulation Component*	*# Nodes*	*# Edges*
First layer (BA)	$m_0 = 6, m = 7$	5000	69640
Second layer (BA)	$m_0 = 7, m = 8$	5000	79520
Multiplex network	Connected layers	5000	149160

Table 4. Information for the Directed ErdosRenyi - ErdosRenyi (ER-ER) Multiplex Network.

ER-ER network	*Simulation Component*	*# Nodes*	*# Edges*
First layer (ER)	$p = 0.005$	5000	125250
Second layer (ER)	$p = 0.0055$	5000	24740
Multiplex network	Connected layers	5000	399690

Table 5. Information for the Directed ErdosRenyi–BarabasiAlbert (ER-BA) Multiplex Network.

ER-BA Network	*Simulation Component*	*# Nodes*	*# Edges*
First layer (ER)	$p = 0.005$	5000	27340
Second layer (BA)	$m_0 = 7, m = 8$	5000	69650
Multiplex network	Connected layers	5000	343110

Table 6. Information for the Directed Online Social Network (Instagram and Twitter) Multiplex Network.

Online social network	*# nodes*	*# edges*
Instagram	13297	52668
Twitter	13297	254388
Multiplex network	26594	307056

MLC-tree Model's Evaluation

One of the assumptions of the MLC-tree model is that missing data has an identical effect on error in the estimated multiplex diffusion tree of the MLC-tree model and complete diffusion network. This hypothesis is valid for the function evaluation and correctness of the proposed MLC-tree model. Moreover, it explains that the estimated multiplex diffusion tree of the MLC-tree model is a true model for the estimation of complete diffusion network features.

To evaluate the correctness of the MLC-tree model, the properties of incomplete multiplex cascade tree are compared with properties of observed diffusion network to determine whether the model has been able to estimates correctly properties of observed diffusion network based on the incomplete diffusion network. As shown in Fig. 10, to evaluate the accuracy of the proposed model (MLC-tree), four properties, namely, (a) infected nodes in the first layer, (b) infected nodes in the second layer, (c) infected edges in the first layer, (d) infected edges in the second layer in multiplex cascade tree of MLC-tree model, in random percentages node missing (α = 0.1, 0.2, ..., 0.9) is compared with observed diffusion network at the same percentage of missing data. The first row is related to ErdosRenyi - ErdosRenyi multiplex network and second row is based on the real multiplex network. The multiplex cascade tree created by the MLC-tree model in different random missing nodes has very close properties with observed diffusion network properties at the same amount of missing data.

Diffusion Network's Properties Estimation

To evaluate error reduction in properties of diffusion network due to missing data, in an estimated tree, 8 properties are extracted and compared with properties of observed diffusion network. As stated, eight properties are: (a) number of infected nodes in the first layer, (b) number of infected nodes in the second layer, (c) total number of infected nodes in the multiplex network, (d) number of infected edges between Layer, (e) number of infected edges in the first layer, (f) number of infected edges in the second layer, (g) number of non-leaf nodes, (h) diffusion network depth. The performance of the MLC-tree model is measured by examining error rate reduction properties of the observed diffusion network. To evaluate the relative error, the relative error of the proposed MLC-tree model is obtained from the relation $\hat{e} = |\frac{x_M - x}{x}|$ was used where x is the value of the property in its observed diffusion network and x_M is estimated value of the same property in the MLC-tree model.

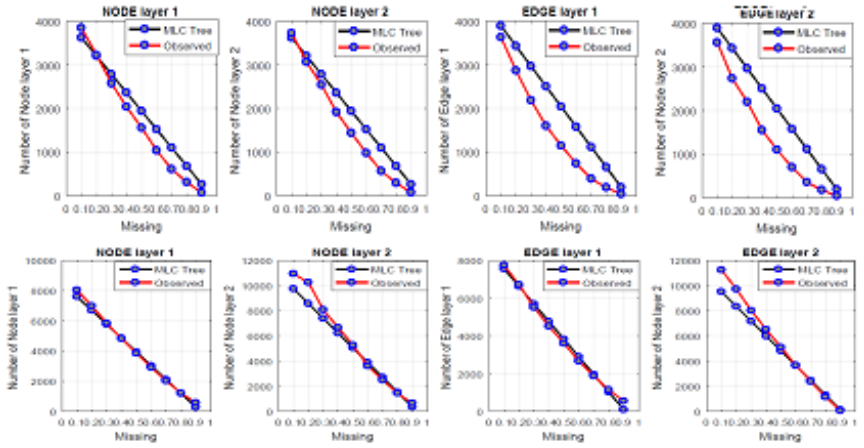

Fig. 10. Four properties of the estimated multiplex cascade tree are compared with same properties of observed diffusion network.

RESULTS

In this section, we provide the impact of four missing data strategies in an infrastructure network on nine properties of the diffusion network. Result access from an average of 40 multiplex networks and diffusion simulated. Figures 10 to 19 show the results of our analysis, that in each of them the X-axis shows the missing data percentage (α) and the Y-axis shows the relative error percentage. In all group results, the two-layer infrastructure networks are considered as multiplex networks, so that in any of them, the results in the first row are for BarabasiAlbert-BarabasiAlbert (BA-BA) as a synthetic network and the second

row for the real multiplex network Instagram-Twitter. Also, we got our results for two more synthetic multiplex networks ErdosReyni-ErdosReyni (ER-ER) and ErdosReyni-BarabasiAlbert (ER-BA); those were similar to BA-BA results so we ignore them to prevent unnecessary information. The missing random edges' only effects on the number of diffusion edges in the first and second layers, the length of the diffusion network, and the number of seed nodes.

Results indicated that, in general, all the missing data strategies have an almost linear relative error on properties of diffusion network, except for the two properties of diffusion length and number of seed nodes. Among the missing data strategies, missing low betweenness centrality nodes has a little more impact than other missing strategies, except for the number of seed nodes. Figure 19 shows that missing low betweenness centrality nodes has had less impact on the number of seed nodes than the other missing strategies. This means that when diffusion arrives at high betweenness centrality nodes, these nodes transmit diffusion to more adjacent nodes. Inversely, low betweenness centrality nodes transmit diffusion to fewer adjacent nodes, so if the information of high betweenness centrality nodes exists, the diffusion process does not split into several diffusions, and there are fewer mistakes to identify seed nodes. It can be concluded that high betweenness centrality nodes have more impact on the diffusion process.

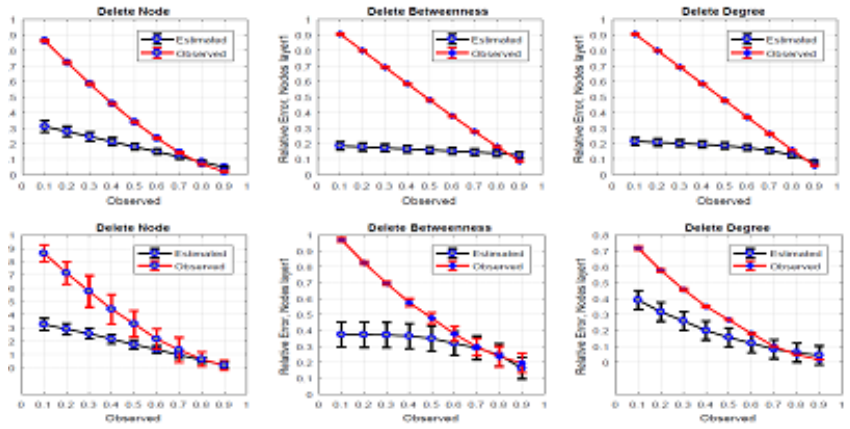

Fig. 11. Impact of the missing number of infected nodes in the first layers.

The impact of missing data on the number of infected nodes in the first layers and estimation of MLC-tree model is shown in Fig. 11. For missing both the random low degree nodes and the low betweenness centrality nodes, the relative error is approximately linear in relation to the percentage of infected nodes in the first layer. On these properties of diffusion network, the missing random node is more effective than missing low degree nodes and missing low betweenness centrality nodes, the missing of low degree nodes also has the least effect on this property

than the missing random node and missing low betweenness centrality nodes. In real data, the missing low centrality nodes are most effective and the missing low degree nodes have the least effect on the number of infected nodes in the first layers. With relative error obtained from the MLC-tree model, it is determined that model at a high missing rate has been able to reduce about 70% of the relative error in the number of nodes infected in the first layer.

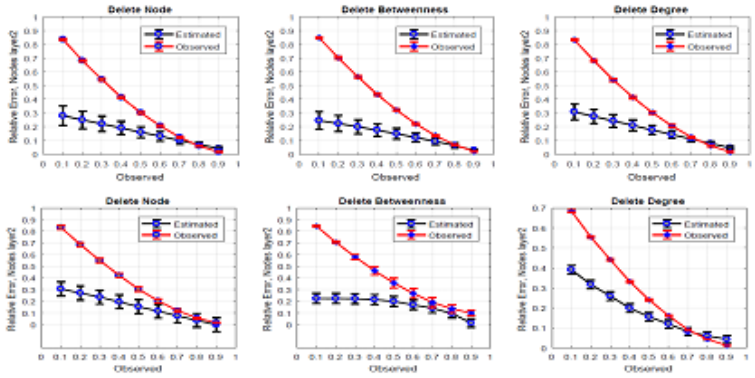

Fig. 12. Impact of the missing number of infected nodes in second layers.

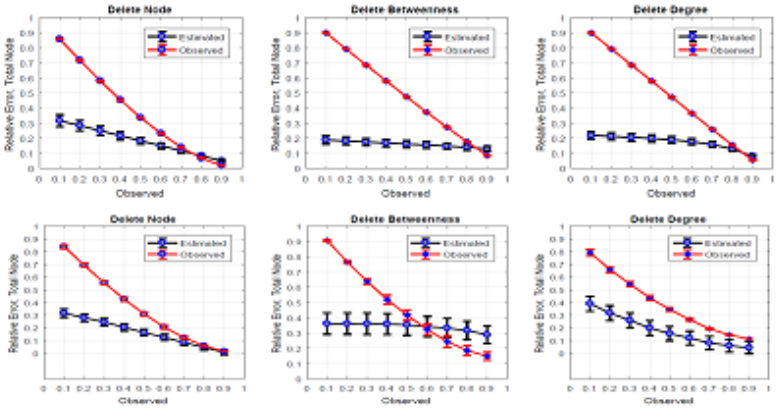

Fig. 13. Impact of the missing total number of infected nodes in the multiplex network.

The impact of missing data on the number of infected nodes in the second layers and estimation of MLC-tree model is shown in Fig. 12. It is clear that the missing data have almost linear effects on the number of infected nodes in the second layer. For example, in all three missing strategies, the BA-BA network for different percentages of the missing have the relative error equal to the percentage

of the missing. Regarding relative error obtained from the MLC-tree model, the model has been able to reduce error in the number of infected nodes in the second layer by a significant percentage. It should be noted that at 10% or 20% of missing data, the model has a relative error close to the relative error in the incomplete diffusion network.

The impact of missing data on the total number of infected nodes in the multiplex network and estimation of the MLC-tree model is shown in Fig. 13. The total number of infected nodes is the sum of the number of infected nodes in the first and second layers: Figure 13 illustrates that the errors in the number of infected nodes in multiplex networks cause the relative error of a random node missing is greater than the relative error of another missing factor for this property. The model also reduces relative error about 70% at high missing rates.

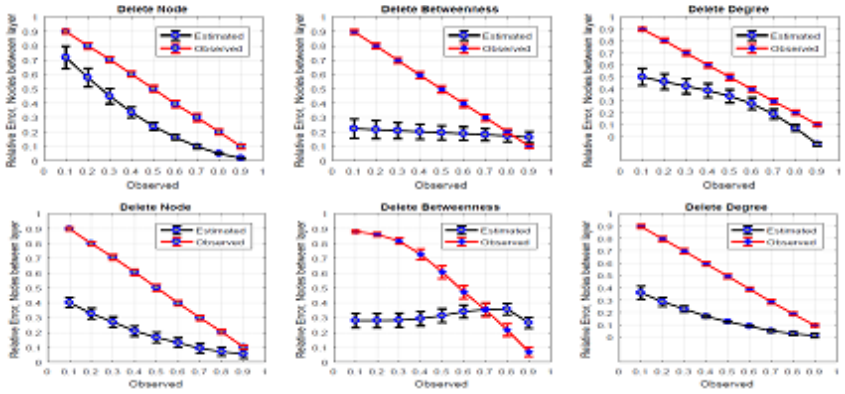

Fig. 14. Impact of the missing number of infected edges between layers.

The impact of missing data on the number of infected between layer edges and estimation of MLC-tree model is shown in Fig. 14. All missing data (random node, low degree node, low betweenness node) have almost linear effects on the number of diffusion edges between layers. Although the relative error of the model is less than the relative error of the incomplete diffusion network, the model has been able to reduce the relative error rate. However, the relative error of the model for the number of infected edges between layers is higher than the relative error of infected edges in first and second layers, especially in missing random nodes. This means that model has a better estimate of the number of infected edges in first and second layers than the number of infected edges between layers.

The impact of missing data on the number of infected edges in first layer and estimation of MLC-tree model is shown in Fig. 15. Missing random edges of the infrastructure network in different percentages is a relatively linear relationship with the relative error of first layer edge missing, missing random edge is an exactly linear relation to relative error from the number of missing edges in the

first layer in diffusion network. The MLC-tree model significantly reduces the relative error of missing random edge. At a high percentage missing, the model has reduced up to 90 percent relative error.

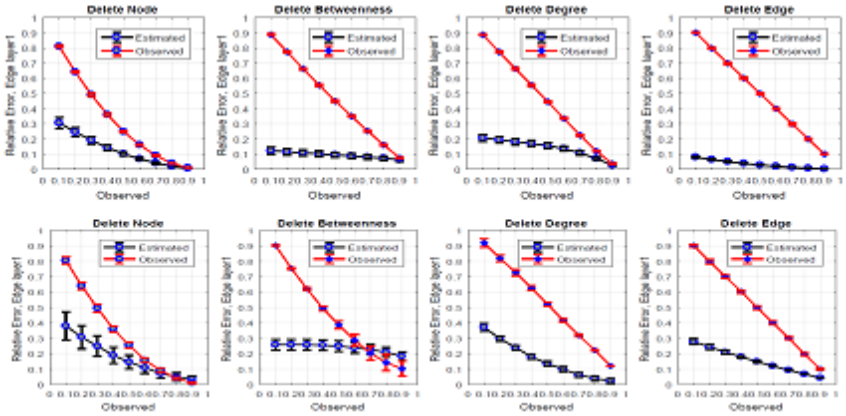

Fig. 15. Impact of the missing number of infected edges in the first layer.

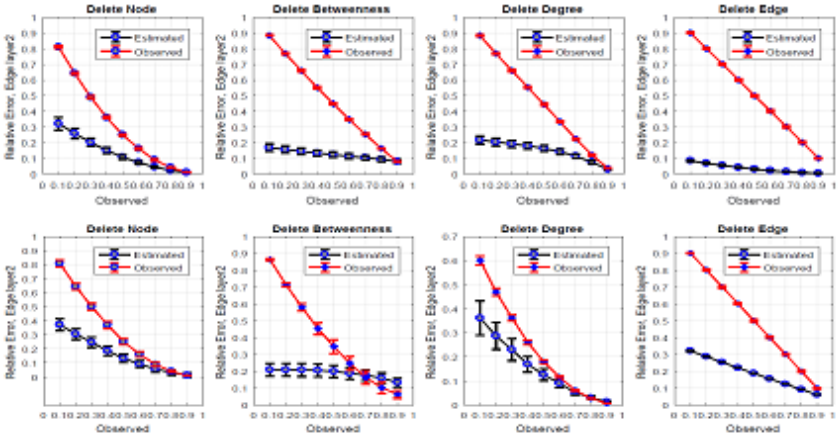

Fig. 16. Impact of the missing number of infected edges in the second layer.

The impact of missing data on the number of infected edges in the second layer and estimation of MLC-tree model is shown in Fig. 16. The relative error of missing edges in the second layer is close to the relative error of missing edges in the first layer. The MLC-tree model also reduces relative error by approximately same relative error of missing edges in the first layer.

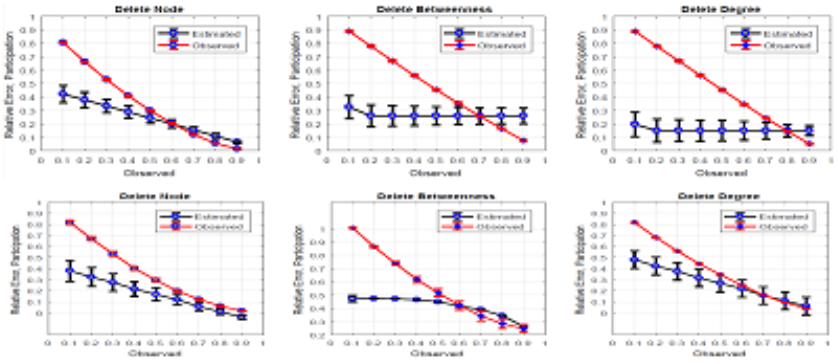

Fig. 17. Impact of missing number of non-leaf nodes.

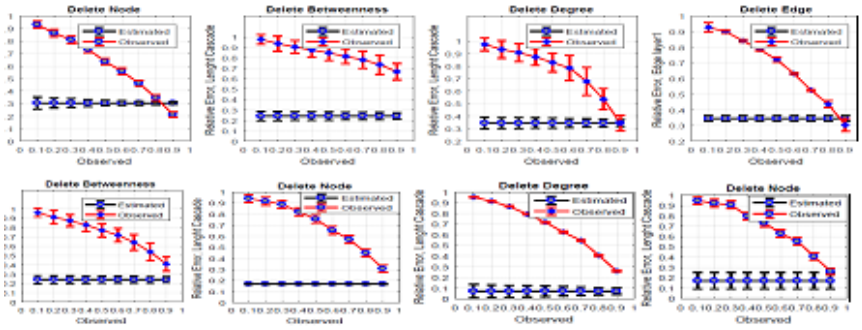

Fig. 18. Impact of 4 missing data on the cascade's length.

The impact of missing data on the number of non-leaf nodes and estimation of MLC-tree model is shown in Fig. 17. A node is a leaf if it has no children; this means that diffusion does not continue when reaches these nodes. In some applications, such as detecting node influence in the diffusion process, leaf nodes are not as important, and non-leaf nodes are more important to detect influencers in the diffusion process. Measuring missing non-leaf nodes can be useful in correcting the results of studies of this kind. Figure 17 shows that missing random nodes have the highest relative error and missing low centrality nodes has the least relative error for missing non-leaf nodes. The MLC-tree model reduces relative error in high missing up to 70%, and at low percentage missing, relative error is close to or greater than incomplete diffusion network.

The impact of missing data on the length of a diffusion network and estimation of MLC-tree model is shown in Fig. 18. Diffusion length is a factor in determining influence rate (Solé-Ribalta et al., 2014) and spreading content as well as the effect of seed nodes in the diffusion network. Figure 18 illustrates the

effect of missing data on the length of diffusion. According to the results of relative errors, at a high rate of missing data, the length of the diffusion network has a high error rate. This means that missing data causes the cascades of diffusion to be broken and by the mistake, the number of cascades is considered to be more than actual, there were fewer cascades in the diffusion network. Short diffusion length can be interpreted as having spread in small and scattered societies. The high relative error in this property in an incomplete diffusion network involves a lot of doubts about the results of an analysis on spreading content (information) and influencer in the diffusion network. Figure 18 shows that the MLC-tree model significantly reduces the relative error of diffusion length in all missing data. In missing 10% or 20% data, the MLC-tree model relative error is close to a relative error in incomplete diffusion network.

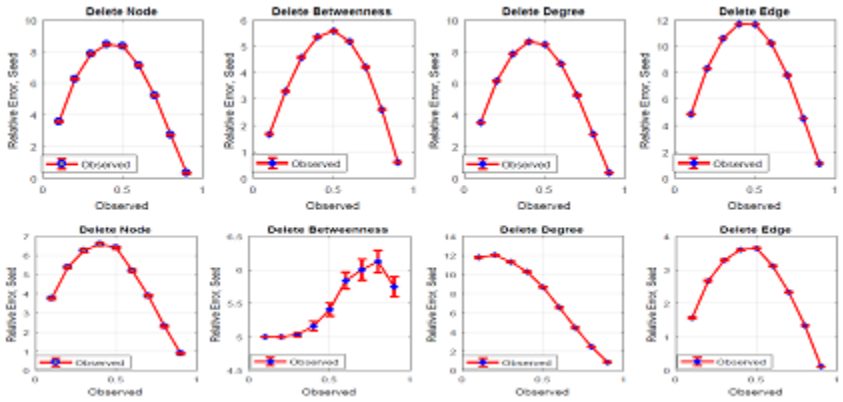

Fig. 19. Impact of the missing number of seed nodes on diffusion.

The impact of missing data on the number of seed nodes and estimation of MLC-tree model is shown in Fig. 19. The number of seed nodes has been shown in different missing strategies. In all missing, the number of seed nodes in an incomplete diffusion network was several times higher than the number of seed nodes in the complete diffusion network. Missing nodes or edges in the diffusion network causes cascade broken into several cascades, which to cause interface nodes, which only play a part in diffusion, as seed nodes. In (Jankowski, Michalski, & Kazienko, 2013), it has tried to detect effective nodes that start diffusion, with a history of nodes in the diffusion. If the diffusion network is associated with missing data, nodes are considered to be incorrectly seeds that are interface nodes. In such studies, a node may be considered as a seed node, which has no effect on the start of diffusion, or node that initiated diffusion and has a significant impact on spreading was considered to be a mistake as a weak seed node. In a low percentage of missing, due to the existence of most nodes in the diffusion network, as well as in high percentage of missing, due to the low number

of diffusion nodes that have existed, the number of observed seed nodes has less error. In missing 30-60%, the diffusion network is beaked to several small-scale networks, and thus the number of seed nodes is higher.

DISCUSSION

One of the criteria in studying the diffusion process is the path and length of the diffusion cascade. The model shows that diffusion cascades can actually be deeper and wider. For example, in diffusion innovation, understanding how many people have been activated in a campaign and through whom, can help campaign managers and marketing companies make better decisions. The MLC-tree model can also help-managers and market analysts to have a more accurate understanding of the scope of the campaign in the community and to allocate their resources more efficiently.

Another important factor in studying network behavior is obtaining important network points for the diffusion process. For example, in digital marketing, identifying influential people on the network and targeting them to start promoting a product leads to a wider promotion of that product. Missing data can also cause the over presence of diffusion cascades than reality, thus causing an error in identification of correct number of influential people. The MLC-tree model helps digital marketing executives select fewer and more suitable people to start an advertising campaign on different networks.

Conclusion

In this paper, we studied the effect of four types of missing data on the diffusion process in multiplex networks. Results indicate that missing data has the highest impact on diffusion length and seed nodes. By analyzing results, it became clear that missing data cause changing to diffusion network into several smaller diffusion networks, and each smaller diffusion network has diffusion length less than the diffusion length of the complete diffusion network. Results show that number of seed nodes in an incomplete diffusion network is far more than in a complete diffusion network. Error at diffusion length in an incomplete diffusion network causes it difficult to identify effective nodes of the diffusion process. Also, the number of seed nodes in an incomplete diffusion network is greater than in a complete diffusion network that leads to a mistake in identifying diffusion's seeds. It may be considered that node in diffusion process has a role of intermediate in transmission of diffusion as one of the important nodes of initiator in diffusion, or the inverse seed node in diffusion process due to missing data of its transmission path as a primitive start node. The analytical model of MLC-tree is presented to obtain some important properties of the diffusion process in the multiplex network. The MLC-tree model uses 7 properties of incomplete diffusion network and extracts three parameters required by it, creating a multiplex diffusion tree that has the same properties as a complete diffusion network. The proposed model can accurately estimate properties of the complete diffusion network even when 90% of the data is missing. Results show that MLC-tree

model is an effective model to reduce the error in important properties of diffusion in multiplex network. The model is also used to estimate the properties of the complete diffusion network without network infrastructure data.

Future Work

In this study, the multiplex network is considered static structure, but at reality in multiplex networks, nodes create new connections or interrupt their old connections. So it's better to consider the dynamic multiplex network. Also, if the infection time of nodes is available, the edges and branches in the diffusion network can be accurately determined.

ACKNOWLEDGMENT

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

REFERENCES

Babaei, S., Molaei, S., & Salehi, M. (2019). Modeling Information Diffusion in Bibliographic Multilayer Networks. *Tabriz Journal of Electrical Engineering, 49*, 503-515.

Belák, V., Mashhadi, A., Sala, A., & Morrison, D. (2016). Phantom cascades: The effect of hidden nodes on information diffusion. *Computer Communications, 73*, 12-21.

Bliss, C. A., Danforth, C. M., & Dodds, P. S. (2014). Estimation of global network statistics from incomplete data. *PloS one, 9*(10), e108471. doi: 10.1371/journal. pone.0108471.

Brewer, D. D., & Webster, C. M. (2000). Forgetting of friends and its effects on measuring friendship networks. *Social networks, 21*(4), 361-373.

Brooks, B. P., DiFonzo, N., & Ross, D. S. (2013). The GBN-dialogue model of outgroup-negative rumor transmission: Group membership, belief, and novelty. *Nonlinear Dynamics, Psychology, and Life Sciences, 17*, 269-293.

Cai, M., Wang, W., Cui, Y., & Stanley, H. E. (2018). Multiplex network analysis of employee performance and employee social relationships. *Physica A: Statistical Mechanics and its Applications, 490*, 1-12. doi: 10.1016/j.physa.2017.08.008

Clauset, A., Moore, C., & Newman, M. E. (2008). Hierarchical structure and the prediction of missing links in networks. *Nature, 45*, 98-101.

Dey, R., Jelveh, Z., & Ross, K. (2012, March). *Facebook users have become much more private: A large-scale study*. Paper presented at the 2012 IEEE International Conference on Pervasive Computing and Communications Workshops. Lugano, Switzerland.

Dickison, M., Havlin, S., & Stanley, H. E. (2012). Epidemics on interconnected networks. *Physical Review E, 85*(6), 066109.

Dou, P., Du, S., & Song, G. (2016, June). *Inferring diffusion network on incomplete cascade data*. Paper presented at the International Conference on Web-Age Information Management, Nanjing, China.

Duong, Q., Wellman, M. P., & Singh, S. (2011, October). *Modeling information diffusion in networks with unobserved links*. Paper presented at the 2011 IEEE Third

International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing, Boston, MA.

Eagle, N., Pentland, A. S., & Lazer, D. (2009). Inferring friendship network structure by using mobile phone data. *Proceedings of the National Academy of Sciences*, *106*(36), 15274-15278.

Eslami, M., Rabiee, H. R., & Salehi, M. (2011,October). *Dne: A method for extracting cascaded diffusion networks from social networks*. Paper presented at the 2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing, Boston, MA.

Eyal, R., Rosenfeld, A., Sina, S., & Kraus, S. (2014). Predicting and identifying missing node information in social networks. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, *8*(3), 14.

Fatemi, S., Salehi, M., Veisi, H., Jalili, M. (2018). A fuzzy logic based estimator for respondent driven sampling of complex networks. *Physica A, June 2018. DOI: 10.1016/j.physa.2018.06.094.*

Fatemi, Z., Salehi, M., & Magnani, M. (2018). A simple multiforce layout for multiplex sociograms. *Proceedings of Social Informatics (SocInfo2018), Saint Petersburg, Russia, September 2018.*

Feng, S., Cong, G., Khan, A., Li, X., Liu, Y., & Chee, Y. M. (2018, April). *Inf2vec: Latent representation model for social influence embedding.* Paper presented at the 2018 IEEE 34th International Conference on Data Engineering (ICDE), Paris, France.

Gjoka, M., Kurant, M., Butts, C. T., & Markopoulou, A. (2010). *Walking in facebook: A case study of unbiased sampling of osns*. Paper presented at the 2010 Proceedings IEEE Infocom.

Gjoka, M., Kurant, M., & Markopoulou, A. (2013). *2.5 k-graphs: From sampling to generation.* Paper presented at the 2013 Proceedings IEEE INFOCOM.

Goldenberg, J., Libai, B., & Muller, E. (2001). Talk of the network: A complex systems look at the underlying process of word-of-mouth. *Marketing letters*, *12*(3), 211-223.

Gomez-Rodriguez, M., Leskovec, J., & Krause, A. (2012). Inferring networks of diffusion and influence. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, *5*(4), 21.

Handcock, M. S., & Gile, K. J. (2010). Modeling social networks from sampled data. *The Annals of Applied Statistics*, *4*(1), 5. doi: 10.1214/08-AOAS221

Jacobsen, J. J., & Guastello, S. J. (2011). Diffusion models for innovation: s-curves, networks, power laws, catastrophes, and entropy. *Nonlinear Dynamics, Psychology, and Life Sciences*, *15*, 307-333.

Jankowski, J., Michalski, R., & Kazienko, P. (2013, August). *Compensatory seeding in networks with varying avaliability of nodes*. Paper presented at the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2013), Niagra, Ontario, Canada.

Kim, M., & Leskovec, J. (2011, April). *The network completion problem: Inferring missing nodes and edges in networks*. Paper presented at the Proceedings of the 2011 SIAM International Conference on Data Mining, Mesa, AZ.

Liben-Nowell, D., & Kleinberg, J. (2008). Tracing information flow on a global scale using Internet chain-letter data. *Proceedings of the National Academy of Sciences*, *105*, 4633-4638.

Lopes, R. R., Navarro, J., & Silva, A. J. (2018). Emotions as proximal causes of word of mouth: A nonlinear approach. *Nonlinear Dynamics, Psychology and Life Sciences, 22*, 103-125.

Maeno, Y. (2007). Node discovery problem for a social network. *arXiv preprint arXiv:0710.4975*, 16375180

Molaei, S., Farahbakhsh, R., Salehi, M., Crespi, N. (2020). Identifying influential nodes in heterogeneous networks. *Expert Systems with Applications, 160 (December), 113580*. doi: 10.1016/j.eswa.2020.113580

Myers, S., & Leskovec, J. (2010, December). *On the convexity of latent social network inference.* Paper presented at the Advances in neural information processing systems, Location: Vancouver, British Columbia, Canada.

Najari, S., Salehi, M., Ranjbar, V., & Jalili, M. (2019). Link prediction in multiplex networks based on interlayer similarity. *Physica A: Statistical Mechanics and its Applications*, 536, 120978. doi: 10.1016/j.physa.2019.04.214

Nikmehr, G., Salehi, M., Jalili, M. (2019). TSS: Temporal similarity search measure for heterogeneous information networks. *Physica A, 524, 696-707*. doi: 10.1016/j.physa.2019.04.207.

Sadikov, E., Medina, M., Leskovec, J., & Garcia-Molina, H. (2011, February). *Correcting for missing data in information cascades*. Paper presented at the Proceedings of the fourth ACM international conference on Web Search and Data Mining, Hong Kong, China.

Salehi, M., Rabiee, H. R., & Rajabi, A. (2012). Sampling from complex networks with high community structures. *Chaos: An Interdisciplinary Journal of Nonlinear Science, 22*, 023126. doi: 10.1063/1,4712602

Salehi, M., Sharma, R., Marzolla, M., Magnani, M., Siyari, P., & Montesi, D. (2015). Spreading processes in multilayer networks. *IEEE Transactions on Network Science and Engineering, 2*(2), 65-83.

Saumell-Mendiola, A., Serrano, M. Á., & Boguná, M. (2012). Epidemic spreading on interconnected networks. *Physical Review E, 86*(2), 026106. doi: 10.1103/PhysRevE.86.026106

Sharma, R., Magnani, M., & Montesi, D. (2016). Effects of missing data in multilayer networks. *Social Network Analysis and Mining, 6*(1), 69. doi: 10.1007/s13278-016-0384-3

Sina, S., Rosenfeld, A., & Kraus, S. (2015). Sami: An algorithm for solving the missing node problem using structure and attribute information. *Social Network Analysis and Mining, 5*(1), 54. doi: 10.1007/s13278-015-0296.7

Solé-Ribalta, A., De Domenico, M., Gómez, S., & Arenas, A. (2014, June). *Centrality rankings in multiplex networks*. Paper presented at the Proceedings of the 2014 ACM conference on Web Science, Bloomington, IN.

Zhao, J., Wang, P., Lui, J. C., Towsley, D., & Guan, X. (2019). Sampling online social networks by random walk with indirect jumps. *Data Mining and Knowledge Discovery, 33*, 24-57.

Zong, B., Wu, Y., Singh, A. K., & Yan, X. (2012, December). *Inferring the underlying structure of information cascades*. Paper presented at the 2012 IEEE 12th International Conference on Data Mining. Brussels, Belgium.