

# A Graph Model with Indirect Co-location Links

Md Shahzamal  
Macquarie University  
Sydney, Australia  
md.shahzamal@students.mq.edu.au

Bernard Mans  
Macquarie University  
Sydney, Australia  
bernard.mans@mq.edu.au

Raja Jurdak  
CSIRO  
Brisbane, Australia  
raja.jurdak@data61.csiro.au

Frank de Hoog  
CSIRO  
Canberra, Australia  
frank.dehoog@data61.csiro.au

## ABSTRACT

Graph models are widely used to analyse diffusion processes embedded in social contacts and to develop applications. A range of graph models are available to replicate the underlying social structures and dynamics realistically. However, most of the current graph models can only consider concurrent interactions among individuals in the co-located interaction networks. They do not account for indirect interactions that can transmit spreading items to individuals who visit the same locations at different times but within a certain time limit. The diffusion phenomena occurring through direct and indirect interactions is called same place different time (SPDT) diffusion. This paper introduces a model to synthesize co-located interaction graphs capturing both direct interactions, where individuals meet at a location, and indirect interactions, where individuals visit the same location at different times within a set timeframe. We analyze 60 million location updates made by 2 million users from a social networking application to characterize the graph properties, including the space-time correlations and its time evolving characteristics, such as bursty or ongoing behaviors. The generated synthetic graph reproduces diffusion dynamics of a realistic contact graph, and reduces the prediction error by up to 82% when compared to other contact graph models, thus demonstrating its potential for forecasting epidemic spread.

## CCS CONCEPTS

• **Computing methodologies** → *Model development and analysis*;

## KEYWORDS

Dynamic graphs; graph mining; social networks; disease tracking

### ACM Reference Format:

Md Shahzamal, Raja Jurdak, Bernard Mans, and Frank de Hoog. 2018. A Graph Model with Indirect Co-location Links. In *Proceedings of Workshop on Mining and Learning with Graphs (MLG'2018)*. ACM, New York, NY, USA, 8 pages. <https://doi.org/unknown>

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

*MLG'2018, August 2018, London, UK*

© 2018 Association for Computing Machinery.

ACM ISBN 123-4567-24-567/08/06...\$15.00

<https://doi.org/unknown>

## 1 INTRODUCTION

Modelling diffusion processes driven by social contacts have recently received significant research attention. These processes range from viral marketing in on-line social networks to infectious disease spreading in social contact within populations. Studying these processes in large real scenarios is not possible without detailed information about the contact patterns and their timing behaviors. However, gathering such large scale data is expensive and complex due to collection methods and privacy concerns. The alternative method is to develop synthetic network structures capturing the statistical properties of real contacts. This synthetic structure is represented by graphs where individuals are presented as actors (i.e., nodes) and relationships among them as edges (i.e., links). In the traditional approach, the edges between nodes are static. However, there are several types of contacts that are not permanent over time such as physical contacts between individuals.

To capture the temporal dynamics of interactions, various dynamic contact graphs have been introduced in the literature [1–4]. Dynamic contact graphs are frequently generated using statistical methods. These models provide realistic graphs for hypothesis testing, "what-if" scenarios, and simulations, but are often mathematically untraceable. Dynamic models require to make graph entities and edges evolve over time and simultaneously maintain the underlying social structures [2]. The preferential attachment is used to study dynamic graphs: new nodes join the graph with non-uniform probability [3]. But, it cannot help maintain social structures. This is addressed by the Nearest Neighbor Model which connects two nodes if they have common neighbor nodes. Another model called Homophily [5] connects two nodes having common interests to form community structures. These models are connectivity driven and face difficulties to capture features of real graphs of large size. The authors of [4] proposed activity driven dynamic network models to build interaction graphs where a node activates at a given time, with its potentiality, and starts creating links with other nodes. This basic activity driven network model (BADN) has been upgraded to capture realistic graph properties applying preferential attachments, reinforcement procedures and attractiveness [6]. Current co-located interaction models, however, assume that links between two individuals are created when they are both present at the same location. Thus, the infectious items are transmitted through a link when both infected and susceptible individuals are present [1, 7]. We refer to diffusion due to these individuals to individual level transmissions as same place same time

transmission (SPST) based diffusion and created contact graphs as SPST interactions graphs. This focus on concurrent presence, is not sufficiently representative of a class of diffusion scenarios where transmissions can occur with indirect interactions, i.e. when there is a time gap between the departure of one individual and the arrival of another. Airborne disease transmission is one such example. An infected individual can release infectious particles in the air through coughing or sneezing. These particles are then suspended in the air and an individual arriving after the departure of the infector can still get infected [8, 9]. In this scenario, current graph models that exclusively track concurrent interactions (SPST) can miss significant spreading events during indirect interactions, thus underestimating the diffusion dynamics. There is a need for a novel dynamic contact graph model to study SPDT diffusions.

This paper proposes a temporal graph model that considers both concurrent (direct) interactions and delayed (indirect) interactions among individuals in forming the links. We termed this model as same place different time interactions (SPDT) graph. In this graph, possible disease transmission links are created if a host node visits a location where at least one other node is present. The node's stay duration at a location with potential to spread is termed as an active period. In order to represent both direct and indirect transmission opportunities, we define the concept of an active copy of a node which is created for each active period of a node. In the proposed model, links are created between the active copies and neighbor nodes. The active copy survives for the active period, when the host is present, in addition to an indirect transmission period, when the host leaves yet the spreading items persist at the location. Thus, the SPDT graph evolves according to temporal changes of links and node's status. SPDT graph generation methods are developed using statistical distributions which are fitted with real graphs of 2 millions users. The model is then validated through its ability to reproduce diffusion dynamics of real contact graphs and studying performances compared to SPST and BADN graph models

The SPDT graph model is introduced in Section 2 while the graph generation is explained in Section 3. Model fitting techniques are presented in Section 4. Section 5 describes the validation of proposed model while Section 6 discusses the results and concludes the paper.

## 2 SPDT GRAPH MODEL

In this section, we first analyse the modeling scenarios and then describe the proposed graph model.

### 2.1 Modeling Scenarios

We first explain the link creation process in SPDT graphs through airborne disease spreading where infected individuals deposit infectious particles at locations they have visited. These particles persist in the environment and can be transferred to susceptible individuals who are currently present nearby (direct transmission) as well as to individuals who visit the location later on (indirect transmission) [8, 9]. Figure 1 illustrates a series of snapshots over time when different individuals visit a location  $L$  (dashed circle) to create these links. Here, an infected individual  $u$  (host individual) arrives at  $L$  at time  $t_1$  followed by a susceptible individual  $v$  at time  $t_2$ . The appearance of  $v$  at  $L$  creates a directed link for transmitting

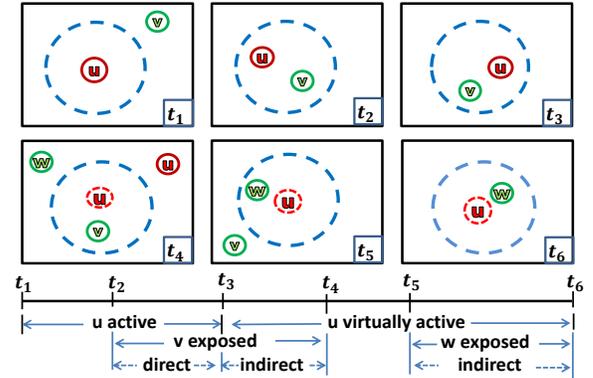


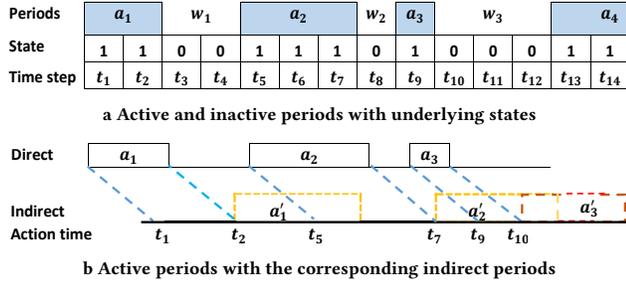
Figure 1: Modeling scenery

disease from  $u$  to  $v$  and lasts until time  $t_4$  including direct contact during  $[t_2, t_3]$  and indirect contact during  $[t_3, t_4]$ . The indirect contact is created as the impact of  $u$  persists (as shown by the dashed circle surrounding  $u$ ), due to the survival of the airborne infected particles in the air. Another susceptible user  $w$  arrives at  $L$  at time  $t_5$  and a link is created from  $u$  to  $w$  through the indirect route due to  $u$ 's infectious particles still being active at  $L$ . These links are called SPDT links and created through space and time. SPDT links may have components: direct and/or indirect transmission links. However, visits of infected individuals to the locations with no susceptible individuals do not lead to transmission of disease.

### 2.2 Model Definition

Individuals may move to various places during their infectious periods while traveling to public places such as office, school, shopping malls and bus stations and visiting tourist places. However, only visits or stays of infected individuals to locations where other susceptible individuals are present create disease transmission links. Our goal is to develop a graph model that is capable of capturing the indirect transmission links along with direct links for co-located interactions among individuals. We focus our modeling of link creation on the time domain to ensure scalability, while abstracting spatial aspects implicitly. Temporal modeling is sufficient to identify the nodes participating in possible disease transmission links. Thus, link creation events in the proposed scenario can be represented as a process where an infected individual activates for a period of time (staying at a location with susceptible individuals) and creates SPDT links. Then, the infected individual becomes inactive for a period of time during which he does not create SPDT links. Inactive periods represent the waiting time between two active periods. Thus, the co-located interaction status of an infected individual can be given by a set  $\{a_1, w_1, a_2, w_2, \dots\}$  where  $a$  is active and  $w$  is inactive period.

We define SPDT graph as  $G = (V, A, E, T)$  to represent all possible disease transmission links among nodes, where  $V$  is the set of nodes. The number of nodes in the graph is constant; however, nodes may have one or more active copies in the graph which captures their ability to spread diseases both at locations they are present and at locations from which they recently departed. The set of active copies for all nodes is represented by  $A$ .  $E$  is the set of links in the graph. The graph is represented over a discrete time set  $T = \{t_1, t_2, \dots, t_z\}$ . Each node in the graph creates a set of active and inactive periods  $\{a_1, w_1, a_2, w_2, \dots\}$ . We define an active copy  $v_i = v(t_s^i, t_f^i)$  for an active period  $a_i$  of node  $v$ , where  $a_i$  starts at



**Figure 2: Proposed model definition**

time step  $t_s^i$  and finishes at  $t_l^i$ . Thus, each node will have several such temporal copies for the observation period. For an active copy  $v_{i+1}$  of a node  $v$ ,  $t_s^{i+1}$  should be greater than  $t_l^i$  of  $v_i$  to capture the requirement that a node should have left the first location before arriving in another location. In this graph, a link  $e_{vu} \in E$  is defined between an active copy  $v_i$  of host node  $v$  and neighbor node  $u$  (node  $u$  visits the current or recent location of node  $v$ ) as  $e_{vu} = (v_i, u, t_s^i, t_l^i)$  where  $t_s^i$  is the joining time and  $t_l^i$  is departure time of  $u$  from the interacted location. The value of  $t_s^i$  should be within  $t_s^i$  and  $t_l^i + \delta$  where  $\delta$  is the time period allowed to create indirect transmission links. Thus, an active copy  $v_i$  of a node  $v$  expires after  $t_l^i + \delta$ , where  $\delta$  captures the decaying probability of infection after  $v$  departs. During the indirect transmission period  $\delta$  node  $v$  can start another active period at another location (see Fig.2b). However, if the infected node  $v$  leaves and returns to the location of  $u$  within a time period  $\delta$ , then there will be two active copies of  $v$ , each with a link to the susceptible node  $u$ . The first copy is due to the persistent of particles from  $v$ 's last visit, while the second copy is due to  $v$ 's current visit.

### 2.3 Graph Evolution

The evolution of the proposed graph is governed by two dynamic processes: 1) switching of nodes between active and inactive states, as in Figure 2a; and 2) link creation and deletion for active copies of nodes. The total number of time steps a node remains in one state determines the current active or inactive period, leading to a set of alternating active and inactive periods  $\{a_1, w_1, a_2, w_2 \dots\}$  for an observation period (see Fig.2a). As stay times at locations are not fixed [10], we define a transition probability  $\rho$  to determine switching from active state to inactive state (modeling stay and departure events of a node at a location). This induces variable lengths of active periods. Similarly, the transition probability  $q$  determines when a node switches from inactive to active state (modeling arrival of a node at location). A similar approach is taken to define link update dynamics in the graph. An active copy of a node creates a link to a newly arriving neighbor node with probability  $p_c$  at each time step until it expires. We define an activation degree probability characterized by  $P(d)$  to model the arrival of multiple new neighbors for an active copy. The created links break (neighbor node leaves interaction area) with probability  $p_b$  at each time step.

## 3 GRAPH GENERATION

We turn our attention now to develop methods for generating the proposed graph, which are designed to capture the statistical properties of realistic scenarios, social contact dynamics as well as

temporal dynamics of SPDT interactions. The graph evolves with generating active copies of each node and creating their links.

### 3.1 Node Activation

Active copies of nodes are created over time according to the active periods which are the building blocks of the graph along with nodes. Thus, we need to generate active periods and intervening inactive periods. In our model, determining whether a node will stay in the current state or transit into the other state at the next time step resembles a Bernoulli process of two outcomes. Thus, the number of time steps a node stays in a state can be obtained from a geometric distribution. With the transitional probability  $\rho$  of switching from active to inactive state, the active period durations  $t_a$  can be drawn from the following distribution as:

$$Pr(t_a = t) = \rho(1 - \rho)^{t-1} \quad (1)$$

where  $t = \{1, 2, \dots\}$  are the number of time steps. Similarly, the inactive period durations,  $t_w$ , with the transition probability  $q$  can be drawn from the following distribution as:

$$Pr(t_w = t) = q(1 - q)^{t-1} \quad (2)$$

where  $t = \{1, 2, \dots\}$  are the number of time steps.

Now, we need to define the initial states of nodes to process active copy generation. Our model follows a two state Markov-process with transition matrix

$$P = \begin{bmatrix} q & 1 - q \\ \rho & 1 - \rho \end{bmatrix}$$

for which the equilibrium probabilities that the node is in inactive state and active state are  $\pi_0$  and  $\pi_1$  respectively, where

$$\pi_0 = \frac{\rho}{q + \rho} \quad \text{and} \quad \pi_1 = \frac{q}{q + \rho} \quad (3)$$

If the initial state of node  $v$  is active, the first active copy  $v_1$  is created for the time interval ( $t_s^1 = 0, t_l^1 = t_a$ ). Otherwise,  $v_1$  will be created for the interval ( $t_s^1 = t_w, t_l^1 = t_w + t_a$ ). Active copy creation continues over the observation period and the corresponding interval ( $t_s, t_l$ ) is defined according to the drawn  $t_a$  and  $t_w$ . Active copies are generated for each node independently. The values of  $\rho$  and  $q$  are the same for all nodes which are fitted with real data.

### 3.2 Activation Degree

Now, we need to define interactions of neighbor nodes with an active copy. Multiple neighbor nodes can contact with an active copy. We note the number of neighbor nodes interacting with an active copy as activation degree  $d$ . The value of  $d$  depends on the spatio-temporal dynamics of the graph and are drawn from a geometric distribution (Eq. 4) instead of finding the arrival times of neighbor nodes.

$$Pr(d = k) = (1 - \lambda) \lambda^{k-1} \quad (4)$$

where  $k = \{1, 2, \dots\}$  and scaling parameter  $\lambda$ . However, individuals in reality have heterogeneous accessibility to public places [11] and hence activation degrees vary for individuals. Thus, heterogeneous  $\lambda$  are selected for nodes and are drawn from a power law distribution of Equation 5:

$$f(\lambda_i = x) = \frac{\alpha x^{-(\alpha+1)}}{\xi^{-\alpha} - \psi^{-\alpha}} \quad (5)$$

where  $\alpha$  is the scaling parameter,  $\xi$  is the lower limit of  $\lambda$  and  $\psi$  is the upper limit which is approximately 1. The value of  $\lambda_i$  defines the range of variations of  $d$  for active copies of a node  $i$  and Equation 4 ensures wide ranges for large values of  $\lambda$ . Combining geometry and power law distributions can generate more realistic degree distribution [12] which are shown in the model fitting section.

### 3.3 Link Creation

With the activation neighbor set, we need to define the arrival and departure dynamics of neighbor nodes for each link created with an active copy. We adopt a similar approach to the definition of active and inactive periods created with transition probabilities. We assume that each link is created with probability  $p_c$  at each time step during the life period  $(t_s, t_l + \delta)$  of an active copy and is broken with probability  $p_b$  after creation. For the link creation delay  $t_c$ , time gap between arrivals of host node and neighbor node  $(t_s - t'_s)$ , we use the truncated geometric distribution:

$$P(t_c = t) = \frac{p_c (1 - p_c)^t}{1 - (1 - p_c)^{t_a + \delta}} \quad (6)$$

where  $t = \{0, 1, 2, \dots, t_a + \delta\}$  are the number of time steps and  $t_a$  is the active period duration of corresponding active copy. Truncation ensures that links are created within  $t_l + \delta$ , i.e. before the active copy expires. In contrast, the link duration  $t_d$ , the stay time of neighbor at the interacted location, does not have a specific upper bound and is generated for each link upon creation through a geometric distribution:

$$P(t_d = t) = p_b (1 - p_b)^{t-1} \quad (7)$$

where  $t = \{1, 2, \dots\}$  are the number of time steps. For simplicity, we set  $p_b = \rho$  as both probabilities relate to how long nodes stay at a location. For each link with an active copy, thus, we can find  $t'_s = t_s + t_c$  and  $t'_l = t'_s + t_d$ . A link with  $t'_s \geq t_l$  is an indirect transmission only component. Link can also have indirect component if  $t'_s < t_l > t'_l$ . The above graph generation steps capture the temporal behavior of SPDT links. The social mixing patterns are integrated by selecting the neighboring node, as we describe below.

### 3.4 Social Structure

Social network analysis has shown that the neighbor selection for creating a link follows a memory-based process. Thus, we apply the reinforcement process [6] to realistically capture the repeated interactions between individuals. In this process, a neighbor node from the set of already contacted nodes is selected with probability  $P(n_t + 1) = n_t / (n_t + \eta)$  where  $n_t$  is the number of nodes the host node already contacted up to this time  $t$  and  $\eta$  is the tendency to broaden the contact set size. On the other hand, a new neighbor node is selected with the probability  $1 - P(n_t + 1)$ . Besides, when a node  $j$  is chosen as a new neighbor by node  $i$ , it is selected with the probability proportional to its  $\lambda_j$  as nodes with higher  $\lambda$  will be neighbors to the more nodes [11]. This ensures nodes with higher potential to create links also have higher potential to receive links.

## 4 MODEL FITTING

We now focus on tuning the model parameters to make them representative of real contact dynamics. While high quality empirical

movement and contact data are difficult to obtain, recent location-based applications create opportunities to gather individual-level geo-tagged updates to serve as a proxy for individual movements. Here, we use the location updates from a social networking application called Momo to estimate model parameters and validate.

### 4.1 Data Set

We analyze 60 million location updates collected over 32 days from 2 million Momo users of two cities (Beijing and Shanghai). The collection system retrieved location updates from the server every 15 minutes. Each update includes spatial coordinates and update times [13]. We build SPDT graphs using these updates where SPDT links are formed assuming airborne disease transmission mechanisms for co-located interactions among users.

Consecutive updates,  $\{(x_1, t_1), (x_2, t_2), \dots\}$  where  $x_i$  are the coordinate values and  $t_i$  are the update times, from a user  $v$  within a radius of 20m (travel distance of airborne infection particles [8, 9]) of the initial update's location  $x_1$  are indicative of the user staying within the same proximity of  $x_1$ . We set the threshold for time difference of any two updates to 30 minutes to remain within the same proximity, as longer gaps may indicate a data gap in the user pattern. For user  $u$ , its visit to the proximity of  $x_1$  will represent an active period that creates an active copy of  $v$  if a susceptible user  $u$  has location updates starting at  $t'_1$  while  $v$  is present, or within  $\delta$  seconds after  $v$  leaves the area. The user  $u$  should have at least two updates within 20m of  $x_1$  to be valid to ensure that it is in fact staying at the same proximity, and therefore can be exposed to the infectious particles, rather than simply passing by. An active period is made with duration  $t_a = t_k - t_1$  which creates an active copy  $v(t_1, t_k)$  of user  $v$ , where  $t_1$  is the starting of active period and  $t_k$  is the end time. If  $u$ 's last update within 20m around  $x_1$  is  $(x'_j, t'_j)$ , an SPDT link  $e_{vu} = (v(t_1, t_k), u, t'_1, t'_j)$  is created with a link creation delay  $t_c = t'_1 - t_1$  and link duration  $t_d = t'_j - t'_1$  for  $v(t_1, t_k)$ . Setting indirect transmission period  $\delta$  to 3 hours (maximum time infectious particles can persist at a location after an infected individual leaves [8]) and processing 60M location updates, we can extract about 3.4M SPDT links that are used for parameter estimation and model validation in the remainder of this section.

### 4.2 Parameter Estimation

In the previous sections, we have defined five co-located interaction parameters (CIP) namely active period ( $t_a$ ), waiting period ( $t_w$ ), activation degree ( $d$ ), link creation delay ( $t_c$ ) and link duration ( $t_d$ ) that construct the SPDT graph  $G = (V, A, E, T)$ . For fitting model parameters with real SPDT graph, we extract CIP using updates collected over 07 days from Momo users of Shanghai city. These location updates made by 126K users form 518K active periods which create a SPDT graph of 1.69M SPDT links. We first use Maximum Likelihood Estimation (MLE) techniques and the sample CIP data for finding model parameters. Then, we generate a synthetic SPDT graph (SG-1) of 126K nodes for 7 days using the estimated model parameters and compare the CIP of synthetic graph with the CIP of real graph (RG) made by 7 days updates of 120K users from Beijing city. To understand the model's response across large graph sizes, we also generate another graph (SG-2) of 0.5M nodes for 7 days.

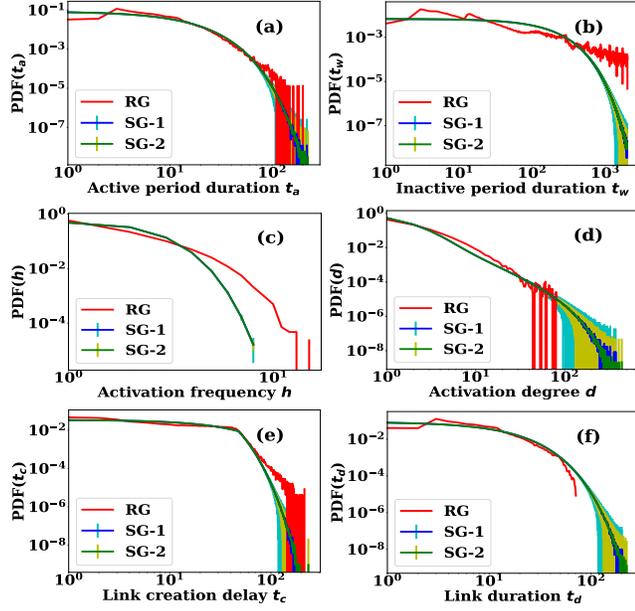


Figure 3: Comparisons of CIP parameters of real graph (RG) and two generated synthetic SPDT graphs: SG-1 and SG-2

The discrete time step of 5 minutes and  $\delta = 3$  hours are used to generate synthetic graphs. Figure 3 shows results of 500 runs for each graph where periods are in time step of 5 minutes.

We find the root squared error (RSE) between the generated and real data distributions of CIP parameters as:

$$RSE = \sqrt{\sum_{i=1}^m (x_i - y_i)^2} \quad (8)$$

where observed values are grouped in the  $m$  bins as they are discrete,  $x_i$  is the proportion of observations for the  $i^{th}$  bin,  $y_i$  is the proportion of empirical dataset values in the  $i^{th}$  bin. As RSE are computed from the proportion values, bins are naturally weighted so that bins representing larger proportions of events have higher contributions in error. We plot the distributions of mean values of CIP with deviations of mean in Fig 3.

**4.2.1 Node Activation Parameters.** In our model, active period durations  $t_a$  are drawn from a geometric distribution with the scaling parameter  $\rho$ . The value of  $\rho$  can be obtained with sample data using the following MLE condition of geometric distribution:

$$\hat{\rho} = \frac{n}{\sum_{k=0}^n t_a^k}$$

where  $n$  is the size of sample set  $t_a = \{t_a^1, \dots, t_a^n\}$ . We apply  $n = 518K$  real active period durations and estimate  $\hat{\rho} = 2.83 \times 10^{-4} s^{-1}$ . The distributions of generated  $t_a$  for both graphs SG-1 and SG-2 are shown in Fig 3a. The RSE errors for both SG-1 and SG-2 is 0.065. The model with fitted parameters consistently generates the active period durations  $t_a$  for both graph sizes.

The active period durations  $t_a$  have similar patterns for all individuals. However, the inactive period duration of an individual depends on how frequently the individual visits public places. Thus, the distribution of  $t_w$  is determined with the distribution of activation frequencies  $h$  of individuals. To find  $q$ , we fit  $h$  with the real

activation frequencies of momo users. According to Equation 3, the probability of transition event  $0 \rightarrow 1$  at a time step is:

$$p_{01} = \frac{\rho q}{q + \rho}$$

Thus, the number of transition events  $0 \rightarrow 1$  during  $z$  time steps represents the number of activation events  $h$ . The probability of  $h$  activations by the nodes is given by the Binomial distribution as:

$$Pr(h | q) = \binom{z}{h} \left( \frac{\rho q}{q + \rho} \right)^h \left( 1 - \frac{\rho q}{q + \rho} \right)^{z-h}$$

The term  $\frac{\rho q}{q + \rho}$  becomes small as  $\rho = 0.085$ ,  $z$  is usually large and  $q < 1$ . Thus, the above equation can be approximated to a Poisson distribution as:

$$Pr(h | q) = \frac{\left( \frac{z \rho q}{q + \rho} \right)^h e^{-\frac{z \rho q}{q + \rho}}}{h!}$$

The MLE condition for the Poisson distribution is given as

$$\frac{z \rho q}{q + \rho} = \frac{1}{m} \sum_{i=1}^m h_i$$

Using the activation frequencies sample set  $h = \{h_1, h_2, \dots, h_m\}$  of size  $m=126K$  to MLE equation provides  $\hat{q} = 2.23 \times 10^{-5} s^{-1}$ . The activation frequencies represent the number of active periods a user does in a day. The sample data set is collected over 7 days to find the average values. The generated activation frequencies for SG-1 and SG-2 are presented in Figure 3c, with RSE of 0.077 for both SG-1 and SG-2 compared to real one. We also plot the corresponding waiting periods durations distribution in Figure 3b which follows the distribution of real  $t_w$  with RSE around 0.031 in both networks. The  $t_w$  is characterized by the irregularity of using Momo Apps.

**4.2.2 Activation Degree Parameters.** For each active copy, an activation degree  $d$  is assigned following Equation 4. The value of  $d$  depends on the node's public accessibility  $\lambda$  drawn from the Power law distribution in Equation 5. Therefore, the distribution of  $d$  in the network will be given for any  $\lambda$  as:

$$\begin{aligned} Pr(d) &= \frac{\beta}{\xi^{\beta-1}} \int_{\xi}^1 (\lambda^{d-\beta-2} - \lambda^{d-\beta-1}) d\lambda \\ &= \frac{\beta}{\xi^{\beta-1}} \left( \frac{1 - \xi^{d-\beta-1}}{d - \beta - 1} - \frac{1 - \xi^{d-\beta}}{d - \beta} \right) \end{aligned}$$

For estimating the parameters  $\hat{\beta}$  and  $\hat{\xi}$ , we derive the MLE equations and apply the activation degree sample set  $d = \{d_1, \dots, d_n\}$  of size  $n=518K$ . We estimate  $\hat{\beta} = 2.98$ ,  $\hat{\xi} = 0.25$ . We set  $\hat{\psi} = 0.999$  as  $\lambda$  should be below 1. The generated activation degree distributions for SG-1 and SG-2 with real data presented in Figure 3d. The RSE error is 0.055. The fluctuating error at the tail is due to data sparsity.

**4.2.3 Link Creation Parameters.** Recall that the maximum link creation delay for a link has the upper bound of the period  $t_a + \delta$ . We apply real active period durations and link creation delays of Momo users to estimate the link creation probability  $\hat{p}_c$

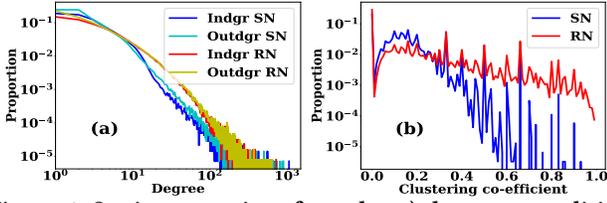


Figure 4: Static properties of graphs: a) degree centralities and b) clustering co-efficients

using the MLE condition:

$$0 = \frac{m}{p_c} - \sum_{k=1}^l \frac{\sum_{j=1}^n (t_c^k - 1)(1 - (1 - p_c)^{t_a^j + \delta}) + (t_a^j + \delta)(1 - p_c)^{t_a^j + \delta}}{\sum_{j=1}^n ((1 - (1 - p_c)^{t_a^j + \delta})^{-1})}$$

where  $t_c^1, t_c^2, \dots, t_c^l$  are sample set of size  $l = 1.2M$  and  $t_a = \{t_a^1, \dots, t_a^n\}$  with  $n = 518K$ . The estimated value of  $\hat{p}_c$  is  $9.33 \times 10^{-5} s^{-1}$ . The generated link creation delays are presented in the Figure 3e, where the generated  $t_c$  have RSE of 0.035 in comparison with the real distribution. The errors were consistent in both SG-1 and SG-2. Then, we set  $p_b = \rho$  for link duration distribution. Figure 3f presents the comparison of generated link durations with real durations which has RSE error of 0.075. The CIP parameters of the generated network are fitted well with the real network parameters made by Momo users. The variations for generated graphs are very low and consistent in both SG-1 and SG-2 networks.

### 4.3 Network Properties

The previous section has addressed the temporal aspects of interactions under SPDT graph fitting CIP parameters to empirical data. This section explores the fitted model's ability to reproduce important static graph properties of empirical networks. The model's parameters have been tuned using the updates from Shanghai. We now utilize the location updates from Beijing to compare the network properties. We generate a synthetic SPDT graph of two weeks with 147K nodes for comparison against an empirical dataset from Momo with the same number of users and duration. We summarize the generated graph by a static graph where a directed edge between two nodes is created if they have at least one SPDT link from host node to neighbor node at any time. We first analyze the degree centrality that quantifies the extent of a node's connectedness to other nodes [14]. In a disease spread context, nodes with higher degree centrality get infected quickly as well as infect a higher number of other nodes [15]. In our model, the growth of the contact set of node  $i$  is determined by  $\lambda_i$  and the neighbor selection process defined by  $p(n_t + 1) = n_t / (n_t + \eta)$ . The value of  $\eta$  controls the degree to which nodes expand their contact set size. We select  $\eta = 1$  to provide reasonable growth in contact set sizes through the influence of  $\lambda$  which varies across nodes, while the selection of an optimal value of  $\eta$  is beyond the scope of this paper. Another desirable feature is that nodes which have more directed links to other nodes also receive more links. We explore this effect in Figure 4a, which plots the distribution of in-degree and out-degree of nodes for both real (RN) and synthetic graphs (SN). The degree distributions are similar in both networks. The correlation between the in-degree and out-degree are about 0.895 for both graphs.

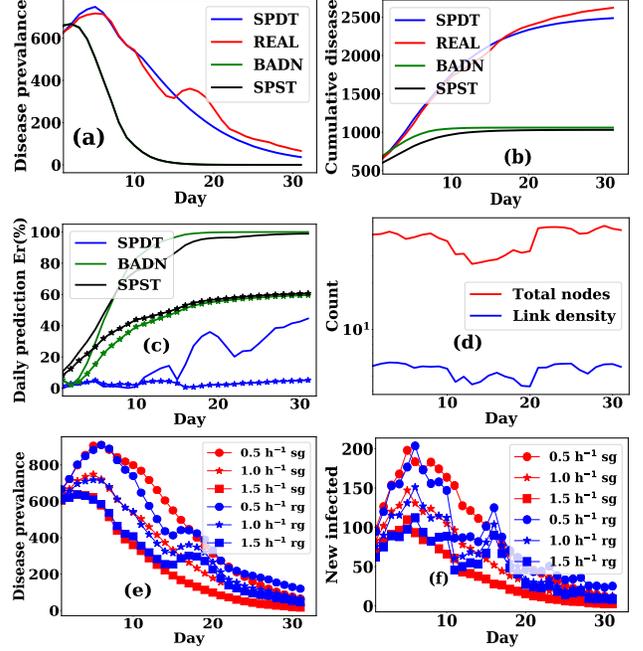


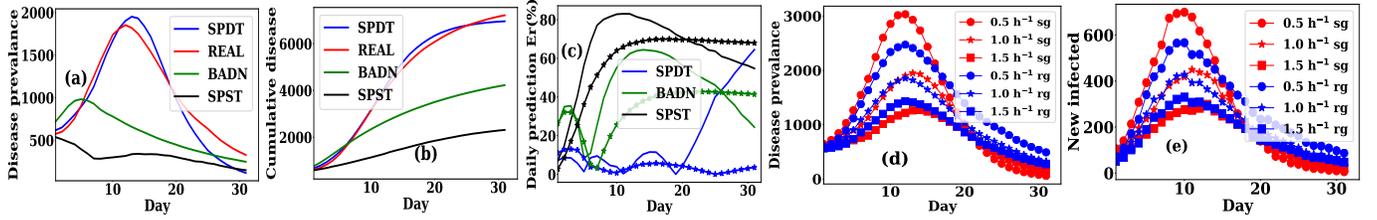
Figure 5: Diffusion dynamics on the sparse graph: (a) disease prevalences  $I_p$  for real graph, SPDT model, BADN graph and SPST graph, (b) final epidemic size, (c) prediction error for  $I_p$ : \* lines for cumulative and other for daily (d) number of Momo users (K) and link densities per user, (e)  $I_p$  for various  $r$  in RG and SG, and (f) new infections for various  $r$

While degree centrality highlights the node connectivity, we use the local clustering coefficient to study the social structure of the network to understand the community structure in the generated graphs [16]. A node selects a new neighbor from its second degree neighbor set, i.e. it's current neighbor's neighbors. For analysis, we convert the directed SPDT links into undirected links and compute the clustering coefficients for each node. We present the results in Figure 4b. The average clustering coefficient in the real graph is 0.11 while the synthetic graph has 0.08. The RSE error between the distributions of clustering co-efficient is 0.0623. We attribute the difference between the two graphs to the distinctions between the randomized links in the synthetic graph and the well-known non-random network structure of empirical social networks, which we leave for future work. Still, our results reflect that the proposed graph model can approach the empirical social structure even with simple methods of neighbor selection.

## 5 MODEL VALIDATION

In this section, we validate the proposed model simulating SPDT process on the generated synthetic graph and real contact graph. Accordingly, airborne disease spreading is simulated on the synthetic and real traces. The simulations are also conducted on another synthetic graph constructed according to the basic activity driven networks (BADN) model [4] to understand how well the proposed model capture diffusion dynamics comparing to the current graph models. In our simulations, the infection probability  $P_I$  for inhaling  $E_T$  dose of infectious particles by a node is

$$P_I = 1 - e^{-\sigma E_T} \quad (9)$$



**Figure 6: Diffusion dynamics on the dense graphs: (a) disease prevalences  $I_p$  for real graph, SPDT model, BADN graph and SPST graph, (b) final epidemic sizes (c) prediction errors, (d)  $I_p$  for various  $r$  in RG and SG, and (e) new infections for various  $r$**

**Table 1: Summary of simulation results**

Graphs	$r/h^{-1}$	Peak $I_p$	Peak Day	Total I	Total Er(%)	Mean Er(%)	STD Er(%)
Sparse graphs							
BADN	1.0	631	3	1069	67	78	33
SPST	1.0	568	2	1030	68	78	27
	0.5	909	7	3462	3	14	15
SPDT	1.0	748	6	2486	5	17	15
	1.5	637	4	1818	16	31	26
Real	0.5	909	7	3355			
	1.0	717	6	2621			
	1.5	637	4	2171			
Dense graphs							
BADN	1.0	1057	6	4225	42	43	17
SPST	1.0	543	2	2310	68	62	22
	0.5	3034	13	9143	4	34	27
SPDT	1.0	1951	14	6956	4	18	19
	1.5	1269	15	5427	7	12	12
Real	0.5	2472	13	9544			
	1.0	1852	13	7313			
	1.5	1429	13	5862			

where  $\sigma$  is the infectiousness of particles [9]. The value of E for a SPDT link  $e_{vu} = (v(t_s, t_l), u, t'_s, t'_l)$  can be calculated as:

$$E = \frac{gp}{Vr^2} \left[ r(t_i - t'_s) + e^{rt_l} (e^{-rt_i} - e^{-rt'_l}) + e^{rt_s} (e^{-rt'_l} - e^{-rt'_s}) \right]$$

$g$  is the particle generation rate of an infected individual,  $p$  is the pulmonary rate of a susceptible individual,  $V$  is the volume of the interaction proximity,  $r$  is the particle removal rate from the proximity and  $t_i$  is the time depends on the link components:  $t_i = t'_l$  if link has only direct component,  $t_i = t'_s$  if link has only indirect component and  $t_i = t_l$  for both components. In simulations, disease propagates according to the Susceptible-Infected-Recovered (SIR) epidemic model. A susceptible node's daily exposure for all links from infected nodes is calculated and its infection probability  $P_I$  is determined. Based on this probability, the node's status is updated at one day intervals. Once infected, a node continues infecting others for a number of days selected uniformly from range of 3 to 5 days, after the node recovers [17]. For simulation,  $r$  is taken randomly in the range  $[0.25, 8]h^{-1}$  around a median,  $g = 0.304PFU/s$ , pulmonary rate  $q = 7.5L/min$  and  $V = 2512M^3$  assuming 20m radius and 2m height of interaction area [9, 18]. The required exposure to induce disease for 50% susceptible individual is found in the range 0.69 to 3.5 PFU. If we take the mid range value of PFU for causing 50% infection, the value of  $\sigma$  will be 0.33 [19]. All simulations start by randomly selecting 500 seed nodes and continues for 32 days.

As SPDT diffusion changes with  $r$ , we verify how well the model captures the changes in diffusion dynamics due to changes in  $r$ . We run simulations on various graphs for values of 0.5, 1, and  $1.5 h^{-1}$ . The summary of the results is presented in the Table 1. First, we chose the real SPDT graphs made by Momo users of Beijing. This graph has total 297K users that create 6.9 million links through 2.2M active periods. We generate a similar SPDT graph with the

fitted model parameters. We run 500 simulations for both graphs choosing the median  $1.0h^{-1}$  of  $r$  and disease prevalences  $I_p$  over simulation days are presented in Fig 5a. The prevalence  $I_p$  shows closely matching trends between the real and synthetic graphs with a concurrent peak at day 7. However, the intensities of  $I_p$  vary for some days after day 15. The variations of  $I_p$  for the real graph arise from fluctuations in the number of users and link densities (see Fig 5d), whereas the SPDT model assumes a constant number of users in the network. Having increased daily prediction error, our model still maintains a cumulative prediction error around 5% as the new infection rates are relatively small after day 10.

We also compare the model performance simulating diffusion on a BADN graph where nodes activate at each time step with a probability  $b$  and generate  $m$  links to others nodes. Analysis of Momo users reveal that their average stay periods  $\Delta t = 1/\rho$  are 50 minutes and on average create  $m = 2$  links during an activation. Thus, we define the activation potential for a node as  $p = f\Delta t/T$ , where  $f$  is the activation frequency=3 per day, and generate a BADN graph for 297K nodes. We run simulations for 500 times and results are presented in Fig 5a. It has infected 1069 nodes in total which is one third of the real graph. Then, the simulation is run on the SPST graph which is obtained by removing indirect paths from SPDT model. This provides similar output of BADN. We calculate the absolute percentage error (APE) for infection events as:

$$APE = 100 \times \frac{I_r - I_o}{I_r}$$

where  $I_r$  is the number of infection events in the real network and  $I_o$  is the infection event in the corresponding observed graph. We calculate the mean APE (MAPE) for the disease prevalence and APE for cumulative disease cases for all graphs. The daily prediction errors in APE of disease prevalence for different models are shown in Fig 5(c). BADN and SPST have high MAPE error 78% with standard deviation 33% while SPDT graph shows 17% with standard deviation of 15% at particle removal rate  $r = 1h^{-1}$ , which reduces the daily prediction error by nearly 82%. The prevalence trends for BADN and SPST do not capture the real dynamics, confirming the superiority of the proposed SPDT graph model.

We next evaluate the model's sensitivity to diffusion parameters. We simulate disease diffusion for three different values of  $r = \{0.5, 1.0, 1.5\}h^{-1}$  on the real graphs (RG) and SPDT graph (SG) while other parameters are kept the same. Results for 500 simulations are presented in Fig 5e for disease prevalence and Fig 5f is for new daily infections. Similar to Fig 5a, link density variations impact the intensities of  $I_p$  in the real graph. However, at higher values of  $r$  the difference shrinks. This is because both graphs approach SPST as  $r$  increases, which reduces the indirect transmission

period. The results for  $r = 1.5h^{-1}$  is more representative for such situations. The prevalence  $I_p$  increases during the later days in the real network as many neighbors are infected early in the SPDT graph, reducing new infection rates at the tail. The similar situation is found in Fig 5(d). The new infection rate in Fig 5f is more random for the real network and strongly follows the link densities.

To understand the response of the graph model for larger scale simulation with higher link densities, we reconstruct a large and dense graph from the real graph dynamics. In the reconstructed graph, we populate any days during which a user has no Momo updates with contacts selected from other days during which the user does have updates. We randomly copy a day from the users' available days to an absence day without changing link properties [20]. We consider a corresponding synthetic SPDT graph of 297K nodes and a BADN graph fitting with the reconstructed graph. The simulation results presented in Figure 6 show that the SPDT graph reduces the daily prediction error by 58% and 71% compared to BADN and SPST respectively, and it reduces the cumulative prediction error by 90% and 95% to BADN and SPST graphs respectively. The daily prediction error of SPDT graph grows sharply after day 20 compared to the real graph, with SPDT graph underestimating the number of cases. This deviation can be explained as follows. When the real SPDT graph is reconstructed copying available day links to missing, same links of the nodes who have links for only one day is copied to other 31 days. If one of these nodes is infected, they transmit disease to the same neighbor nodes for whole infectious period. In the proposed model, however, nodes always have some probability to connect to a new neighbor and hence the the contact set size grows, which is a higher chance to cause more infections. Thus, new infection rate in the SPDT graph model before touching peak is higher than the infection rate in real graph. As susceptible reduces quickly in SPDT model at early days of simulation, new infection rate falls faster in the later days. The MAPE error reduces from 34% to 12% as  $r$  increases. This is because the indirect paths dominate more at lower  $r$  causing more new infections. Still, our model maintains a cumulative prediction error of 4-7% across all configurations, further highlighting the applicability of our model.

## 6 CONCLUSION

We have introduced a SPDT graph model and demonstrated its utility for a case study of airborne disease diffusion. The SPDT graph captures both direct and indirect contacts for simulating diffusion process. The proposed graph model is capable of capturing contact dynamics, applying reinforcement for capturing repetitive interactions and public accessibility-based attractiveness to engage in interactions. We demonstrated how the model can be fitted to empirical geo-tag data from social networking App and its ability to reproduce both graph structural properties and diffusion dynamics of empirical graphs. The model generates co-located interaction parameters with considerably low RSE error and reduces prediction error of cumulative spread by up to 95% over existing graph models. The graph model shows similar response of real graphs to environmental conditions for changing diffusion dynamics.

The significance of our SPDT model lies in its capture of indirect interactions for diffusion phenomena, which accounts for previously disregarded pathways for transmission. We expect the model

to be useful for forecasting infectious disease diffusion within a population given contacts data of a population. More importantly, the model can be used to simulate what-if scenarios to aid health managers and authorities in planning for possible outbreaks and allocating resources for targeted responses. The SPDT graph model can be applied to study diffusion phenomena in on-line social network (OSN) such as on-line post can be seen by current active users instantly while inactive users sees them later on [21]. This model can also be used to model online and offline activities of users in OSN [22]. There are several interesting directions for future developments of our model. Nodes in the current model activate with the same frequency. Thus, it would be interesting to study graph properties with heterogeneous frequencies and find correlation with public accessibility. Another interesting direction is to study the contact set size growth and temporal properties like betweenness. The sensitivity of model parameters to graph stability and model performances against other graph models will also be studied.

## REFERENCES

- [1] P. Holme, "Modern temporal network theory: a colloquium," *The European Physical Journal B*, vol. 88, no. 9, p. 234, 2015.
- [2] M. Kim and J. Leskoved, "Modeling social networks with node attributes using the multiplicative attribute graph model," STANFORD UNIV CA, Tech. Rep., 2011.
- [3] A. Vázquez, "Growing network with local rules: Preferential attachment, clustering hierarchy, and degree correlations," *Physical Review E*, vol. 67, 2003.
- [4] N. Perra, B. Gonçalves, R. Pastor-Satorras, and A. Vespignani, "Activity driven modeling of time varying networks," *Scientific reports*, vol. 2, p. 469, 2012.
- [5] M. McPherson, L. Smith-Lovin, and J. M. Cook, "Birds of a feather: Homophily in social networks," *Annual review of sociology*, vol. 27, no. 1, pp. 415-444, 2001.
- [6] M. Karsai, N. Perra, and A. Vespignani, "Time varying networks and the weakness of strong ties," *Scientific reports*, vol. 4, p. srep04001, 2014.
- [7] E. Baccelli, P. Jacquet, B. Mans, and G. Rodolakis, "Multi-lane vehicle-to-vehicle networks with time-varying radio ranges: Information propagation speed properties," in *Information Theory Proceedings (ISIT)*. IEEE, 2013, pp. 809-813.
- [8] A. Fernstrom and M. Goldblatt, "Aerobiology and its role in the transmission of infectious diseases," *Journal of pathogens*, vol. 2013, 2013.
- [9] M. Shahzamal, R. Jurdak, R. Arablouei, M. Kim, K. Thilakarathna, and B. Mans, "Airborne disease propagation on large scale social contact networks," in *Proceedings of the 2nd Int. Workshop on Social Sensing*. ACM, 2017, pp. 35-40.
- [10] T. Cheng and J. Shen, "Grouping people in cities: From space-time to place-time based profiling," in *Human Dynamics Research in Smart and Connected Communities*. Springer, 2018, pp. 181-201.
- [11] L. Alessandretti, K. S., A. B., and N. P., "Random walks on activity-driven networks with attractiveness," *Physical Review E*, vol. 95, no. 5, p. 052318, 2017.
- [12] S. Chattopadhyay, C. Murthy, and S. K. Pal, "Fitting truncated geometric distributions in large scale real world networks," *Theoretical Computer Science*, 2014.
- [13] T. Chen, M. A. K., and R. Boreli, "The where and when of finding new friends: Analysis of a location-based social discovery network," in *ICWSM*, 2013.
- [14] L. C. Freeman, "Centrality in social networks conceptual clarification," *Social networks*, vol. 1, no. 3, pp. 215-239, 1978.
- [15] G. F. De Arruda, A. L. B., P. M. Rodríguez, F. A. R., Y. M., and L. da F. C., "Role of centrality for the identification of influential spreaders in complex networks," *Physical Review E*, vol. 90, no. 3, p. 032812, 2014.
- [16] G. Laurent, J. S.äki, and M. Karsai, "From calls to communities: a model for time-varying social networks," *The European Physical Journal B*, vol. 88, 2015.
- [17] K. Spricer and P. T., "Characterizing the initial phase of epidemic growth on some empirical networks," *arXiv preprint arXiv:1709.00973*, 2017.
- [18] S. Shi, C. Chen, and B. Zhao, "Air infiltration rate distributions of residences in beijing," *Building and Environment*, vol. 92, pp. 528-537, 2015.
- [19] R. H. Alford, J. A. Kassel, P. J. Gerone, and V. Knight, "Human influenza resulting from aerosol inhalation," *Proceedings of the Society for Experimental Biology and Medicine*, vol. 122, no. 3, pp. 800-804, 1966.
- [20] A. Machens, F. G., C. R., A. E. T., A. B., and C. C., "An infectious disease model on empirical networks of human contact: bridging the gap between dynamic network data and contact matrices," *BMC infectious diseases*, vol. 13, no. 1, 2013.
- [21] S. Gao, J. Ma, and Z. Chen, "Modeling and predicting retweeting dynamics on microblogging platforms," in *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*. ACM, 2015, pp. 107-116.
- [22] F. Probst, L. Grosswiele, and R. Pflieger, "Who will lead and who will follow: Identifying influential users in online social networks," *Business & Information Systems Engineering*, vol. 5, no. 3, pp. 179-193, 2013.