A fuzzy logic based estimator for respondent driven sampling of complex networks

Samira Fatemi a, Mostafa Salehi a,b,*, Hadi Veisi a, Mahdi Jalili c

a Faculty of New Sciences and Technologies, University of Tehran, Iran
b School of Computer Science, Institute for Research in Fundamental Science (IPM), P.o.Box 19395-5746, Tehran, Iran
c RMIT University, Australia

HIGHLIGHTS

• We propose a fuzzy based RDS estimator that takes the node degrees as fuzzy number.
• We use various fuzzy membership functions and find the best one.
• Numerical results show that our estimator significantly improves the performance.

ARTICLE INFO

Article history:
Received 3 December 2017
Received in revised form 17 April 2018
Available online xxxx

Keywords:
Respondent driven sampling
Degree
Estimator
Fuzzy logic

ABSTRACT

Respondent Driven Sampling (RDS) is a popular network-based method for sampling from hidden population. This method is a type of chain referral (or snowball) sampling in which an estimator is used to infer the proportion of the population with that property. Existing RDS estimators are asymptotically unbiased based on various underlying assumptions. However, these assumptions are often violated in practice, and little attention has been given to violation of one of these assumptions on accurately reporting the degree by all nodes. In this paper, we address the violation of this assumption and propose a new estimator based on fuzzy computing. In particular, the number of an individual’s contacts can be a fuzzy concept. Using fuzzy functions, we transform the reported degrees to fuzzy numbers and estimate the infection prevalence in the hidden population by the proposed estimator. We simulate RDS method under the condition that all assumptions are satisfied except the one for the degree, and then evaluate the proposed estimator in synthetic and real datasets. Our results show that the fuzzy-based estimator can reduce the sampling bias in average 54% as compared to the existing methods.

© 2018 Elsevier B.V. All rights reserved.

1. Introduction

Measuring the prevalence of infectious diseases such as HIV or flu in hidden populations, such as injection drug users of people who have high-risk sexual behaviors, is one of the challenges facing governments to prevent and control spread of such diseases. Accessing the members of hidden or hard-to-reach populations is difficult and membership in such populations is potentially risky. Such diseases often start from a number of individuals and spread through the population based on the connection network between the individuals [1]. In hidden populations, traditional sampling methods such as random sampling [2] are not efficient and use of such methods is not practical [1,3]. Thus, one has to rely on network-based methods. In these methods, it is assumed that the connection graph is unknown to us and our knowledge about the network...
is local. Thus, for sampling we often start from an initial node of the network, and first visit some or all of its neighbors, and then neighbors of neighbors and so on.

A number of sampling models have been proposed such as convenience sampling [4], snowball sampling [5,6], random walk sampling [7], time and location sampling [8] and Respondent Driven Sampling (RDS) [1,3,5]. Among these methods, RDS is most widely used in practice. RDS has been used by US Centers for Disease Control and Prevention and the World Health Organization to estimate the prevalence of HIV infections [9]. As a broader exposition of potential applications, RDS can also be used for effective vaccination [10], on fast responses in critical life-saving situations [11]. For a useful introduction of vaccination and information cascades in complex networks, the readers can refer to [12,13].

RDS process starts by recruiting a number of individuals as seeds that are typically 5–10 members of target population. Then, each seed node is interviewed and given a fixed number of coupons. Respondents use coupons for passing and recruiting their contacts as new members to study. Successfully recruited contacts are interviewed and delivered the same number of coupons to recruit next wave of respondent from their contacts. In the same way, the samples can grow through many waves, until it reaches the desired sample size or cannot reach any other sample in the population. The respondents are often encouraged with reward when they participate and recruit from their contacts. Fig. 1 graphically illustrate the procedure of RDS.

RDS is popular method because of mainly two advantages over alternative methods: high response rate in hidden population and asymptotically unbiased estimates [14]. However, in order to obtain unbiased estimations, one has to make some strong assumptions. Inevitably, RDS is biased towards individuals with higher degree in the target population. To adjust this selection bias, Volz and Heckathorn [15] proposed an estimator, denoted by V-H estimator, that weights individuals with inverse of their degree, as

\[
\hat{\mu} = \frac{\sum_{i \in S} 1/d_i}{\sum_{i \in S} 1/d_i},
\]

where \(\hat{\mu}\) is the estimate, \(S\) is the set of sampled individuals, \(A\) is the set of individuals who are infected, and \(d_i\) is the self-reported degree of individual \(i\).
<table>
<thead>
<tr>
<th>Estimator</th>
<th>Reference</th>
<th>Further details</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naive</td>
<td>[1]/1997</td>
<td>First RDS estimator</td>
</tr>
<tr>
<td>S-H</td>
<td>[3]/2004</td>
<td>Estimates by measuring the proportion of within-group and cross-group recruitments</td>
</tr>
<tr>
<td>H</td>
<td>[29]/2007</td>
<td>Tries to adjust differential recruitment</td>
</tr>
<tr>
<td>V-H</td>
<td>[30]/2008</td>
<td>Estimates by weighting nodes with their inverse degrees</td>
</tr>
<tr>
<td>S-S</td>
<td>[31]/2010</td>
<td>Tries to adjust with replacement sampling</td>
</tr>
</tbody>
</table>

It has been proved that V-H estimator provides asymptotically unbiased estimates if some initial assumptions are satisfied [15], however many of such assumptions are not satisfied in practice. Violation of any of these assumptions heavily affects performance of the estimation. One of these assumptions is to accurately report degree that means the respondents should be able to correctly report their degree. Therefore, each respondent must be aware of its neighbors. In many practical cases, the individuals might know their degree approximately, or do not want to report the correct value for any reason. Sometimes, the individuals report a wrong value for the degree. Therefore, in practice these is always some levels of uncertainty in the values reported by the individuals as their degree. It has been frequently shown that violation of degree assumption has significant influence on the outcome of RDS estimator [3,16–18]. In this paper, we propose a novel RDS estimator that takes the node degrees as a fuzzy number and use fuzzy computing to obtain the outcome. We use various fuzzy membership functions and find the one with the best performance. Indeed, fuzzy logic is a powerful tool to handle uncertainty in many systems [19], which is used in this work. Our numerical results on synthetic and real networks show that the proposed fuzzy-based estimator significantly improves the performance of the estimator by reducing its bias caused by violation of degree assumption.

In summary, our main contributions are as follows. We propose a novel RDS estimator that takes the node degrees as a fuzzy number and use fuzzy computing to obtain the outcome. We also use a member of fuzzy membership functions and find the one with the best performance. Numerical results on synthetic and real networks show that our estimator significantly improves the performance.

2. Related works

In this section we give a brief description of RDS assumptions and explain how they are often violated in practice. We then summarize the related works about these assumptions in a table and provide a list of some state-of-the-art RDS estimators.

2.1. RDS assumptions

As mentioned, RDS is a referral-chain technique for collecting data from hidden populations and has been widely used due to providing high response rate in hidden populations and asymptotically unbiased estimation. The existing estimators often make some assumptions, summarized as follows.

- The connection graphs are supposed to be connected [9,20].
- The networks are assumes to have only reciprocal connections, i.e. undirected networks [9,20–22].
- The individuals are allowed to be selected more than once, i.e. sampling with replacement [9,20,23].
- The nodes are required to correctly report their degree [9,24,25].
- During the sampling process, the nodes randomly select among their neighbors [9,20,23,26,27].
- Each node can pass only one coupon [20,28].

However, such assumption might not be hold in many practical applications. For example, many networks of interest might not be connected, and it is a difficult task to check the connectivity with process sampling. Some networks are directed networks and the individuals might have links in only one directions. Some of these assumptions are network assumptions while some are sampling assumptions. There are a number of studies about RDS assumptions that can be categorized in three of classes. In some studies, researchers study how violation on assumptions impact the estimation. In some other studies, researchers try to develop new estimators that are less sensitive to violation of assumptions. The third class of works try to recommend methods to detect violation conditions in practice [9].

2.2. RDS estimators

As mention, due to importance of RDS method, it is necessary to understand sources of bias and develop new estimators which are less sensitive to assumption violation. Table 1 summarized state-of-the-art RDS estimators that have been proposed for this purpose.

V-H estimator is a powerful sampling technique. Accuracy of V-H estimator depends on knowing the exact node degrees and inaccuracy of degree values results in a significant bias in the estimators. Here we address this issue using on innovative solution based on fuzzy logic.
3. Fuzzy logic based estimator for RDS

3.1. Fuzzy logic and membership functions

Fuzzy logic is a programming logic that try to mimic human behavior in reasoning. Unlike binary logic in which we either have 1 or 0. The fuzzy quantities take values between 0 and 1. This value is denoted by fuzzy membership value. For example, a 40 years old person might be 0.2 Young, 0.8 Middle Age and 0.1 Old. Indeed, unlike classic sets, which have crisp bounds and the memberships are either 0 or 1, in fuzzy sets the members belong to the sets based on their membership values.

In summary, fuzzy reasoning has in general three steps: fuzzification, inference and defuzzification. In order to fuzzify crisp functions, one has to use proper membership functions. The concept of fuzzy functions have important role in the fuzzy sets theory. A membership function assigns a membership value in the range [0,1] to each member. Membership functions can be linear or non-linear and symmetric or asymmetric. There is no standard mathematical approach to choose the best shape for fuzzy membership functions, and it mainly depends on the application at hand. Here we consider three types of functions, as shown in Fig. 2, and choose the best one. In the following we include mathematical details of the considered functions.

3.1.1. Rectangular fuzzy functions

This function is defined by two parameters as follow \((a, b > 0)\):

\[
\text{trn}(x : a, b) = \begin{cases} 
0 & (x < a) \text{ or } (x > b) \\
1 & a \leq x \leq b 
\end{cases}
\]  

(2)

In Eq. (2), \(a, b\) are the beginning and ending of fuzzy intervals in rectangular fuzzy functions and operational rules rectangular numbers used in this paper are given as follow:

- Summing of two rectangular fuzzy numbers:
  \[(a_1, b_1) \pm (a_2, b_2) = (a_1 \pm a_2, b_1 \pm b_2)\]  
  (3)

- Dividing of two rectangular fuzzy numbers:
  \[(a_1, b_1) \div (a_2, b_2) = (a_1/a_2, b_1/b_2)\]  
  (4)

- Reversing of two rectangular fuzzy numbers:
  \[A^{-1} = (a_1, b_1)^{-1} = (1/b_1, 1/a_1)\]  
  (5)

- Defuzzification of a rectangular fuzzy number:
  \[
def(a_1, b_1) = (a_1 + b_1)/2\]  
  (6)
3.1.2. Triangular fuzzy functions
This function is defined by three parameters and its mathematical relation is as follow \((a, b, c > 0)\):

\[
\text{trn}(x : a, b, c) = \begin{cases} 
0 & \text{\((x < a)\) or \((x > c)\)} \\
(x-a) & a \leq x \leq b \\
(b-a) & b \leq x \leq c \\
(c-x) & b \leq x \leq c \\
(c-b) & b \leq x \leq c 
\end{cases}
\] (7)

In Eq. (7), \(a, b, c\) are the beginning, middling and ending of fuzzy intervals in triangular fuzzy functions and the operational rules are:

- Summing of two triangular fuzzy numbers:
  \[(a_1, b_1, c_1) \pm (a_2, b_2, c_2) = (a_1 \pm a_2, b_1 \pm b_2, c_1 \pm c_2)\] (8)

- Dividing of two triangular fuzzy numbers:
  \[(a_1, b_1, c_1) \div (a_2, b_2, c_2) = (a_1/a_2, b_1/b_2, c_1/c_2)\] (9)

- Reversing of two triangular fuzzy numbers:
  \[A^{-1} = (a_1, b_1, c_1)^{-1} = (1/c_1, 1/b_1, 1/a_1)\] (10)

- Defuzzification of a rectangular fuzzy number:
  \[\text{def}(a_1, b_1, c_1) = (a_1 + b_1 + c_1)/3\] (11)

3.1.3. Trapezoidal fuzzy functions
This function is defined by four parameters: a lower limit \(a\), an upper limit \(d\), a lower support limit \(b\), and an upper support limit \(c\), where \(0 < a < b < c < d\). It is mathematically shown by the following equation:

\[
\text{trn}(x : a, b, c, d) = \begin{cases} 
x-a & a \leq x \leq b \\
b-a & b \leq x \leq c \\
1 & c \leq x \leq d \\
d-c & d \leq x \leq c \\
0 & (x > d\text{ or } x < a) 
\end{cases}
\] (12)

The operational rules for this function are as follows:

- Summing of two trapezoidal fuzzy numbers:
  \[(a_1, b_1, c_1, d_1) \pm (a_2, b_2, c_2, d_2) = (a_1 \pm a_2, b_1 \pm b_2, c_1 \pm c_2, d_1 \pm d_2)\] (13)

- Dividing of two trapezoidal fuzzy numbers:
  \[(a_1, b_1, c_1, d_1) \div (a_2, b_2, c_2, d_2) = (a_1/a_2, b_1/b_2, c_1/c_2, d_1/d_2)\] (14)

- Reversing of a trapezoidal fuzzy number:
  \[A^{-1} = (a_1, b_1, c_1, d_1)^{-1} = (1/d_1, 1/c_1, 1/b_1, 1/a_1)\] (15)

- Defuzzification of a trapezoidal fuzzy number:
  \[\text{def}(a_1, b_1, c_1, d_1) = (a_1 + b_1 + c_1 + d_1)/4\] (16)

3.2. Fuzzy estimator

In practice, there is always some levels of uncertainly in the values reported by the individuals as their degree. Since fuzzy logic is a well-known method to handle uncertainly, here we propose a new estimator that takes the node degrees as a fuzzy number. In this estimator, our main goal is to adjust the bias of degree inaccuracy in V-H estimator. We show that our method can successfully adjust the bias caused in the estimation process. The proposed estimator is designed through two steps:

(i) Converting the reported degrees by the individuals to fuzzy numbers: In this step, the reported degrees of sample nodes are converted to fuzzy numbers according to the selected fuzzy membership function.

(ii) Extending V-H estimator to use fuzzy numbers for the estimation: To this end, we replace degrees in the relevant equations with the obtained fuzzy numbers. The final output of the estimator is a fuzzy number that needs to be converted in to a crisp number. This is performed using centroid defuzzification method.
Fig. 3. Estimation of the methods for different sampling rates. “Classical” represent the results of classical RDS estimator when degree condition is fulfilled, “Violated” represent the results when the degree condition is violated, and the results obtained by the proposed method in such conditions is shown as “Fuzzy”. The results show mean values and standard deviations over 100 runs. We apply triangular membership function in various degree intervals in each data test regard with their degree distributions, we select [1,5,10], [9,20,30], [25,40,55], [50,65,80], [75,100,130] for Jazz network, [11,5,4,6,8,7,10,13], [12,15,18] for American Football network, [11,25,20,40,50], [45,65,75], [70,90,100], [95,120,150] for Barabasi–Albert network and [1,1400,380,420,450,440,480,510,500,540,570,560,580,600] for Small-world network. As this figure shows, the proposed Fuzzy-based estimator has much less estimation error as compared to the case with original V-H estimator used for degree-violated situation.

The procedure of the proposed estimator is as follows:

1. The reported degree values by the individuals (for both sample and infected nodes) are fuzzified using Eqs. (2), (7) or (12) depending on the selected fuzzy membership function.
2. Inverse of fuzzy degrees are obtained using Eqs. (5), (10) or (15) depending on the fuzzy membership functions used.
3. The inverted fuzzy degrees of the sample and infected nodes are summed using Eqs. (3), (8) or (13).
4. The fuzzy prevalence estimation is obtained based on Eqs. (4), (9) or (14).
5. The estimated prevalence rate is defuzzified using Eqs. (6), (11) or (16).

It is worth to note that time and memory complexity of the proposed method is similar to V-H estimator that is computed as $O(S)$ where $S$ is the sample size.

4. Performance evaluation

In this section, we provide details on how we evaluate the proposed estimator on test data sets. We also introduce the disease spread model used in this paper.
4.1. Datasets

We use both real-world and synthetic networks to evaluate our proposed estimator.

**Synthetic Networks:** We use two models to generate synthetic networks as follows:

- Watts–Strogatz (WS) networks \([32]\): The model starts from a completely regular network where nodes are connected to their \(k\)-nearest neighbors. Each link is rewired with probability \(p_0 \in (0, 1)\). In simulation, we generate networks with \(|V| = 1000\) nodes, \(k = 5\) and \(p_0 = 0.5\).
- Barabasi–Albert Scale Free (B-ASF) \([33]\): We start with \(m_0 + 1\) all-to-all nodes. At each time step, one node with \(m\) links is added to the network. The endpoints of the new links are randomly selected among all nodes according to the following probability: \(p_1(x) = \beta_1 \frac{d_{in}(x)}{|E|} + \beta_2 \frac{d_{out}(x)}{|E|} + \beta_3 \frac{1}{|V|}\), where \(d_{in}(x)\) is the number of incoming edges to node of \(x\), \(d_{out}(x)\) is the number of outgoing edges from node of \(x\) and \(\beta_1 + \beta_2 + \beta_3 = 1\). We generate networks with \(|V| = 1000\) and average degree of nodes is 50. We leave the number of links, \(|E|\), unconstrained. For attaching probability, we set \(\beta_1 = 0.7\), \(\beta_2 = 0.2\), \(\beta_3 = 0.1\).

**Real-world Networks:** Although studying dynamical processes on model networks might provide useful information on how they behave in real networks, real networks might have properties that are not fully captured by models. Therefore, we also apply the proposed algorithm on a number of real networks, as follows.

- Jazz Musicians Network (JMN): This is a jazz musicians social network \([34]\) that originates from musician community. The dataset includes 198 users and there are 2742 links among these users. Each link between two users, represents two musicians that play in one group.
- American College Network (ACN): This is a network that originates from soccer players teams for students at University of Los Angeles \([35]\) in fall 2000. Each link between nodes represents two teams that play to another. The dataset includes 115 node and 613 links among these nodes.

4.2. Epidemic spreading model

Epidemic disease is one of the important issues in biological and social networks studies. Epidemic diseases are caused by biological pathogens and can be transmitted from person to person. Examples include HIV and influenza. Epidemic diseases can spread widely in a population or remain at low levels for a long time, depending on the network structure. The distribution pattern of an epidemic disease is mainly determined by characteristics of the pathogen and the structure of the connection network.

To study the spread pattern on the test datasets, we use the Susceptible–Infectious–Recovered (SIR) epidemic spreading model \([36]\). This is an epidemiological model widely used to simulate the spreading of epidemics, i.e. number of people infected with a contagious disease, in a population as a function of time. At every time step, the model assumes a transition rate \(\theta_1\) for a susceptible person to become infected, if the person is in contact with an infected neighbor, and a rate \(\theta_2\) for an infected person to become recovered or die. The recovered person will never be infected again. In the simulations, we use \(\theta_1 = 0.2\) and \(\theta_2 = 0.05\). We run the simulation until steady-state solution is obtained.

4.3. Experimental setup

We evaluate the performance of the proposed estimator in various situations and in terms of different aspects. To this end, we study the effects of sampling rate, infection rate, degree distribution and parameters of fuzzy membership functions. Also, we consider some evaluation measures including Average Estimates (AE), defined by the mean of estimates derived by estimators, Bias, defined by the absolute difference between estimated and true population and \(SD\), standard deviation for each sample size \(j\). For each sampling rate, we repeat the simulations 100 times by selecting random nodes as seed and report the averages over these runs.

4.4. Simulation results

For every dataset, we simulate RDS in ideal conditions and use V-H estimator for prevalence estimation. The results obtained from this implementation is referred to as “Classical-RDS” in this paper which is considered as the ground-truth performance. We also consider the case when the degree condition is violated. We round the degree values up to the nearest multiple of 5, and apply V-H estimator; the results are referred to as “violated” \([24]\) here. The results obtained from the proposed method is referred to as “Fuzzy”. In the evaluations of the proposed method, as there is no real reported degree in the datasets, we used the same violation method (i.e., rounding the real degree values in the datasets up to the nearest multiple of 5) to generate the reported degrees for simulations. In the paper, we use rectangular, triangular and trapezoidal membership functions \(3.1\) and consider various intervals to convert degrees to fuzzy numbers with regard to the degree distribution in each case test. Consequently, triangular fuzzy function is selected as the best membership function according to its average results. Therefore we apply triangular fuzzy function in all simulation scenarios.

In the following we study the performance of the methods in different scenarios and in terms of various evaluation metrics.
Table 2
Bias of estimates of the estimators for different sampling rates. It is seen that the standard RDS estimator has the lowest bias, as there is no violation in the degree assumption. Also, the proposed fuzzy-based estimator has much less bias as compared to V-H estimator with violated degree assumption, indicating efficiency of the fuzzy computing in improving the performance of the estimator.

<table>
<thead>
<tr>
<th>Networks</th>
<th>Sampling rate</th>
<th>Bias</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Classical</td>
</tr>
<tr>
<td>Jazz Musicians</td>
<td>0.02</td>
<td>0.0394</td>
</tr>
<tr>
<td></td>
<td>0.04</td>
<td>0.0477</td>
</tr>
<tr>
<td></td>
<td>0.06</td>
<td>0.0167</td>
</tr>
<tr>
<td></td>
<td>0.08</td>
<td>0.0001</td>
</tr>
<tr>
<td>Barabasi–Albert</td>
<td>0.02</td>
<td>0.0129</td>
</tr>
<tr>
<td></td>
<td>0.04</td>
<td>0.0051</td>
</tr>
<tr>
<td></td>
<td>0.06</td>
<td>0.0036</td>
</tr>
<tr>
<td></td>
<td>0.08</td>
<td>0.0001</td>
</tr>
<tr>
<td>American Football</td>
<td>0.02</td>
<td>0.0576</td>
</tr>
<tr>
<td></td>
<td>0.04</td>
<td>0.0121</td>
</tr>
<tr>
<td></td>
<td>0.06</td>
<td>0.0196</td>
</tr>
<tr>
<td></td>
<td>0.08</td>
<td>0.0001</td>
</tr>
<tr>
<td>Small-World</td>
<td>0.02</td>
<td>0.0330</td>
</tr>
<tr>
<td></td>
<td>0.04</td>
<td>0.0401</td>
</tr>
<tr>
<td></td>
<td>0.06</td>
<td>0.0182</td>
</tr>
<tr>
<td></td>
<td>0.08</td>
<td>0.0188</td>
</tr>
</tbody>
</table>

4.4.1. Effect of sampling rate

Here, we study the performance of the estimators in test networks with various sampling rates, in the range of 2%, 4%, 6%, 8% of the network size. Networks are infected with rate of 0.2 by SIR model. Fig. 3 shows the results for real and model networks. As this figure shows, the proposed Fuzzy-based estimator has much less estimation error as compared to the case with original V-H estimator used for degree-violated situation. Fuzzy-based method indeed results in performance close to the ideal case, when all nodes report their degree and the degree condition is fully met. Expecting, the classical RDS estimator with full satisfaction of degree condition has the best performance in all cases. These results indicate that considering degree as a fuzzy number and using Fuzzy-specific computations substantially improves the performance of the estimator. It is also worth nothing that as sampling rate increases, the standard deviation of the results decrease, which is an expecting phenomenon. Since the infection rate is fixed, as the sampling rate increases, less individuals are selected as sample, leading to increase in the standard deviation.

Table 2 shows the bias values of the estimator for different sampling rates. Clearly, the standard RDS estimator has the lowest bias, as there is no violation in the degree assumption. These results also confirm that the proposed fuzzy-based estimator has much less bias as compared to V-H estimator with violated degree assumption, indicating efficiency of the fuzzy computing in improving the performance of the estimator.

4.4.2. Effect of infection rate

Fig. 4 shows results for different values of infection rate of SIR model. We consider sample size as 5% of each network size. According to these results, as the infection rate increases, the estimators perform worse. However, the proposed fuzzy-based estimator still performs better than the original one, and has close performance to the standard RDS estimator for which the degree assumption is satisfied. Table 3 shows the bias values of the estimator for different infection rates. Again, the fuzzy-based estimator results in lower bias then the original estimator, revealing efficiency of the proposed strategy.
The effect of infection rate on the performance of Fuzzy-Estimators to adjust bias due to reported degree inaccurately. We consider the sample size as 5% of each network size. The results show that as the infection rate increases, the estimators perform worse. However, the proposed fuzzy-based estimator still performs better than the original one, and has close performance to the standard RDS estimator for which the degree assumption is satisfied.

5. Conclusion

In this paper, we proposed a novel method based on fuzzy logic to improve performance of Respondent Driven Sampling (RDS) of complex networked systems. Many RDS algorithms make some assumption, which might not be hold in realistic cases. Some of the original algorithms in this field are based on the assumption that all nodes are aware of their degree and can report it. Here we showed that when such an assumption is violated, one effectively use fuzzy computing to improve the performance of the estimator. In the proposed method we considered degree as a fuzzy number and used fuzzy computation to compute the outcome of the estimator. Our numerical simulations on both real and model networks showed that the proposed fuzzy-based estimator significantly outperforms the original algorithm, and results in close estimation as the estimator that meets the degree condition. The results show that our fuzzy based estimator can reduce the sampling bias in average 54% as compared to the existing methods. Moreover, the behavior of fuzzy estimator is close to well-known V-H estimator when all assumptions are satisfied. As future work, one can study on some theoretical justifications (such as error bounds) to show how the proposed method reduces the sampling bias.

Acknowledgments

This research was in part supported by a grant from IPM (No. CS1397-4-143). Mahdi Jalili is supported by Australian Research Council through project No. DP170102303.
References