Estimation of Average Degree of Social Network Using Clique, Shortest Path and Cluster Sampling to monitor Network Reliability

Vivek Kumar Gupta[1], Diwakar Shukla[2]

•

Department of Mathematics and Statistics
Dr. Harisingh Gour Vishwavidyalaya
Sagar, M.P., 470003, India
[1]v.vivekgupta@yahoo.com; [2]diwakarshukla@rediffmail.com

Abstract

In recent past, Online Social Networks (OSN) has emerged as a platform for sharing information, thoughts, and activities. In the real-world network, method of considering the appropriate samples is most frequently used for network analysis. Graph sampling is a procedure used for computing unknown parameters. Many sampling algorithms exist in literature such as Random node, Random edge sampling, Rank degree, etc. can be used for estimation. This paper presents a comparison of clique based procedure (CBP) and shortest path based procedure (SPP) to estimate the average degree of a vertex in a social network using an overlapping cluster sampling. A comparative procedure is used to obtain the lower and upper limit of confidence intervals with the help of multiple samples. Ogive based simulation is also used for single value computation of limits of CI. The results, obtained from simulation, show that clique based sampling algorithm (CBP) is more efficient than the shortest path based sampling algorithm (SPP). The estimated confidence intervals can be used for monitoring the reliability of a social network in terms of control over average network degree.

Keywords: Graph, Sampling, Social network, Overlapping cluster, Confidence interval (CI), Shortest path procedure (SPP), Clique based procedure (CBP), Reliability, Percentage relative gain(PRG)

1. Introduction

Online Social Networks (OSN) are used by large numbers of people around the world interacting with each other by forming like minded groups, based on the commonness of characters. Many real-world complex systems can be represented as a collection of vertices and edges — for example, information networks, communication networks, biological networks, etc. Recently evolved a surge of interest for exploring the characteristics of these networks, modeling their structure, develop algorithms for them, and examining systems that govern networks [8]. However, many of the real-world networks are too large to acquire, store or analyze, e.g. 3 billion emails per day worldwide from multiple sources to multiple destinations. The scientific community focuses on developing scalable analytic methods for different size datasets. In order to facilitate the development and testing of systems for network domains, it is often necessary to take a sample (smaller subgraphs) from a large network structure. A sampled subgraph can be used to drive realistic simulations and experimentation. Just to have a precise assessment of the performance of such systems, it is suggested by many scientists to use appropriate sampling methods that can select a good representative of networks. Graph sampling [4] is used to study small subsets of networks along with preserving the main features of the original network [6] [7].

Physical distances are utilized to get interaction between the different system variables. For example, the distance between two atoms or between two galaxies in the universe to evaluate the intensity of force of attraction.
A good sampling [1] algorithm for estimating a parameter must have:

- Cost effectiveness.
- Sample size suitability for unbiased parameter estimation.
- Practical and effective ways of accessing the graph.
- Lesser amount of time and reduction of computational efforts.

In networks, distance is a kind of path linkage used in a different manner. Distance between two web pages or between two unknown individual physical distances [9] [12] is not relevant. A path is a link in the network and distance in network represents the number of links the path contains.

In this paper, a method of cluster sampling for networks is presented using the concept of the shortest path and cliques. The approach has focus to find the shortest path and cliques between several pairs of vertices by selecting random pairs. The degree sequence of vertices in these shortest paths is taken for construction of overlapping clusters [18, 2]. The sampled pair of vertices of the social network contains only a fraction of all possible pairs of vertices.

The aim is to obtain an estimate of average degree which is a valid parameter of real network. Paper is organized as Section 2 contains definitions, overlapping sampling, motivation, and related work described in brief. Section 3 describes a sampling scheme with properties like bias and variance estimate. The performance of the proposed procedures is examined through ogive based simulation whose results are reported and comparision of efficiency are in Section 4-6. Other sections 7 to 9 reveal reliability, discussion, efficiency comparison, and conclusion.

2. Definition and Related Work

A Social network(graph) G(V, E) is represented as a pair of a vertices set V(G) and an edge set E(G), the number of vertices in G is N. Simple Graph(Network) G(V, E) contains undirected, unweighted edges, neither loops nor multiple edges. The neighborhood of u is N(u) = {v: (u, v) $\in$ E(G)}. It forms a set of edges connected to u. The degree is the number of connections that a vertex has, degree (v) = $|N(v)|$.

Average degree of a vertex is the average number of edges per vertex in the graph. It is defined as:

$$\frac{\textbf{Total number of Edges}}{\textbf{Total number of vertices}} = \textbf{Average Degree}$$

2.0.1 Clique

A clique is a subset of a network in which the vertices are more closely and intensely tied to one another than they are to other vertices of the network. The term "Dyad" is the smallest clique composed of two adjacent vertices. The chain of adjacent cliques is used as a tool for forming the community. Community detection allows professionals like election planners, community specialist physicians to understand the characteristics and role within the network and outside the network [16]. Concerned literature of methodologies of community detection [11] [17] have been developed a lot and several methods are in picture. Every algorithm has advantages, disadvantages, and working limitations over others. Many of them fail while dealing with overlapping communities [20] [19]. For example, the community of soldiers and community of drinkers may be overlapping where unique identification is a difficult procedure. One can find out new ways and means to generate community detection in networks. Mathematically clique is a subset of vertices all adjacent to each other. It can be used for community structure detection for large scale networks. The community identification in these methods is defined as a chain of adjacent cliques. Some methods can find the community structure for very large-scale networks. A method proposed in [19] is also useful for such cases.

Note 2.1 : If symbol e_{ij} is used as an edge from vertex vx_i to vx_j then shortest path from one vertex to another is a path sequence of vertices $(vx_1, vx_2, \ldots, vx_n)$ so that overall possible n minimizes the $\sum_{i=1}^{n-1} f(e_{i,i+1})$.

Note 2.2 : If any two vertices are selected in a graph there may exist a shortest path. Also, there may multiple shortest paths of same length d_{ij} between vertices vx_i and vx_j. Note that the shortest path does not consider any loop or any intersect itself. Further, in an undirected graphical network, these lengths of shortest path $d_{ij} = d_{ji}$ holds between any two vertices vx_i and vx_j. But in a directed graph network, it may happen that $d_{ij} \neq d_{ji}$.

2.1. General Computational Algorithm

- Take a network G(V, E), where V = set of vertices $(vx_1, vx_2, \ldots, vx_N)$, E = set of edges $\{e_1, e_2, \ldots, e_m\}$.
- Using random sampling, select K pair of vertices or set of vertices from G(V, E) as the case may be.
- Apply an appropriate procedure of overlapping cluster formation.
- Create a degree sequence of vertices (clusters).
- Estimate average degree of network using overlapping cluster sampling mean estimation method.

2.2. Computational Procedure for Creating Clusters

2.2.1 Shortest Path Procedure(SPP)

In this, non-adjacent pair of vertices are selected in a graphical network, and using Dijkstra's algorithm [10] one can find the shortest path whose degree sequence can be obtained.

2.2.2 Clique Based Procedure (CBP)

In this, the K vertices are selected as source vertices and one can find the clique, where a clique is a complete subgraph whose degree sequence can be calculated.

2.3. Motivation

The Clique Based Procedure (CBP) was used by [16] for computing the average edge length for community detection. This procedure provides the construction of overlapping clusters. The shortest path procedure (SPP) also provides the construction of overlapping clusters. In sampling theory, there exist methodologies to estimate average value of a parameter in the setup of overlapping clusters. This paper presents a comparison of SPP and CBP using the mathematical approach of cluster sampling techniques for network mean degree estimation. Newman and Milgram [11, 14, 13] suggested with evidence of why using the concept of shortest path for sampling social networks. Newman's [11] experiment on scientific collaboration shows that on average 64% scientists collaborator shortest path pass through one's top-ranked collaborator and 17% pass through the second-ranked one. Milgram's [14] [13] experiment of small-world phenomena concludes that delivering a message from one person to another by using shortest path based on local information exist in large social networks and that by using only local information. In general, in social networks, information [5] propagates along the shortest paths of users as a direct and simple way to communicate. For example, smart advertisement of products with minimum cost by maximum influence path. Above discussion motivates to take the cliques and shortest paths [3] as the building blocks to sampled network. By using it one can estimate different network parameters and at the same time can preserve the network functionalities. A clique may be an alternative of shortest path and need to be examined.

3. Estimation of Average Degree using Overlapping Cluster Sampling

Let there are total K clusters, many of them are having overlapping vertices of different degrees formed by appropriate computational method. The i^{th} cluster (i = 1, 2, 3, ..., K) contains N_i units. Suppose the term Y_{ij} denotes degree of j^{th} vertex belonging to i^{th} clusters and F_i be the frequency of j^{th} vertex occurring in K clusters. Total distinct vertices in the network graph are N.

Step 1: Let k out of K (k< K) clusters are selected randomly who are formed either by method SPP or by CBP.

Step 2: From the i^{th} cluster of size N_i, the $n_i(n_i < N_i)$ vertices are selected by SRSWR.

Define $D_{ij} = \frac{MY_{ij}}{NF_j}$, i = 1,2, ..., K and j = 1, 2, ...,N_i,
where, M denotes total vertices in all K cluster (including overlapping, M > N). The D_{ij} indicates that degree values Y_{ij} at K clusters are normalized and converted. Overall average unknown network parameter is:

$$\bar{D} = \frac{1}{K}\sum_{i=1}^{K}\frac{1}{N_i}\sum_{j=1}^{N_i} D_{ij} \tag{1}$$

Theorem 1. A biased estimator of average $\bar{D}$ is given by [2]

$$\bar{d} = \frac{1}{k}\sum_{i=1}^{k}\frac{1}{n_i}\sum_{j=1}^{n_i} d_{ij} \tag{2}$$

where, d_{ij} represents D_{ij} units who are present in sample n_i.

Proof. Let us consider (see [2])
E_2 = The conditional expectation over a given sample of cluster
E_1 = The expectations for over all such sample,

$$\begin{aligned} E(\bar{d}) &= \left(\tfrac{1}{k}\textstyle\sum_{i=1}^{k}\tfrac{1}{n_i}\sum_{j=1}^{n_i} d_{ij}\right) \\ &= E_1E_2\left(\tfrac{1}{k}\textstyle\sum_{i=1}^{k}\tfrac{1}{n_i}\sum_{j=1}^{n_i} d_{ij}\right) \\ &= E_1\left(\tfrac{1}{k}\textstyle\sum_{i=1}^{k} E_2(\bar{d}_{i\bullet})\right) \\ &\text{where, } \bar{d}_{i\bullet} = \text{sample within clusters} \\ &= E_1\left(\tfrac{1}{k}\textstyle\sum_{i=1}^{k} E_2(\bar{D}_{i\bullet})\right) = \bar{D} \neq \bar{Y} \end{aligned}$$

Hence the theorem. ■

Note 3.1 The $\bar{d}$ is a biased estimator of $\bar{Y}$ and its bias is given by $Bias(\bar{d}) = E(\bar{d}) - \bar{Y}$
This bias can be estimated by [2]

$$\widehat{Bias}(\bar{d}) = \frac{K-1}{K\bar{N}(k-1)}\sum_{i=1}^{k}(N_i - \bar{n})(\bar{d}_{i\bullet} - \bar{d}) \tag{3}$$

3.0.1 Estimation of Variance:

Consider average square between cluster averages in the sample is

$$s_b^2 = \frac{1}{k-1}\sum_{i=1}^{k}\left(\bar{d}_{i\bullet} - \bar{d}\right)^2 \tag{4}$$

It can be shown that

$$E(s_b^2) = S_b^2 + \frac{1}{K}\sum_{i=1}^{K}\left(\frac{1}{n_i} - \frac{1}{N_i}\right)S_i^2 \tag{5}$$

Also, one can define

$$s_i^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} \left(d_{ij} - \bar{d}_{i\bullet}\right)^2 \tag{6}$$

Further,

$$E(s_i^2) = S_i^2 = \frac{1}{N_i - 1} \sum_{j=1}^{N_i} \left(d_{ij} - \bar{D}_i\right)^2 \tag{7}$$

So,

$$E\left[\frac{1}{k}\sum_{i=1}^{k}\left(\frac{1}{n_i} - \frac{1}{N_i}\right)s_i^2\right] = \frac{1}{K}\sum_{i=1}^{K}\left(\frac{1}{n_i} - \frac{1}{N_i}\right)S_i^2 \tag{8}$$

Thus, one can express

$$E(s_b^2) = S_b^2 + E\left[\frac{1}{k}\sum_{i=1}^{k}\left(\frac{1}{n_i} - \frac{1}{N_i}\right)s_i^2\right] \tag{9}$$

and an unbiased estimator of S_b^2 is

$$\widehat{S_b^2} = s_b^2 - \frac{1}{k}\sum_{i=1}^{k}\left(\frac{1}{n_i} - \frac{1}{N_i}\right)s_i^2 \tag{10}$$

Also, an estimator of the variance can be obtained by replacing S_b^2 and S_i^2 by their unbiased estimators as:

$$\widehat{Var}(\bar{d}) = \left(\frac{1}{k} - \frac{1}{K}\right)\widehat{S_b^2} + \frac{1}{kK}\sum_{i=1}^{k}\left(\frac{1}{n_i} - \frac{1}{N_i}\right)\widehat{S_i^2} \tag{11}$$

3.0.2 Confidence Interval (CI)

Let a and b are the two real numbers and P(A) denotes the probability of an event A. The 95% confidence interval is defined as $P[a < \theta < b] = 0.95$, where θ is an unknown parameter. As per theory of normal distribution the best choice of a and b is

a = Estimated average -1.96$\sqrt{\text{Estimated variance}}$,
b = Estimated average + 1.96$\sqrt{\text{Estimated variance}}$.

4. Proposed Sampling Scheme And Dataset

To evaluate amd compare the two methods CBP and SPP sampling the well known Zachary's Karate Club [15] network datasets have taken into account. Zachary network are widely used to study the efficiency of different graph sampling techniques.

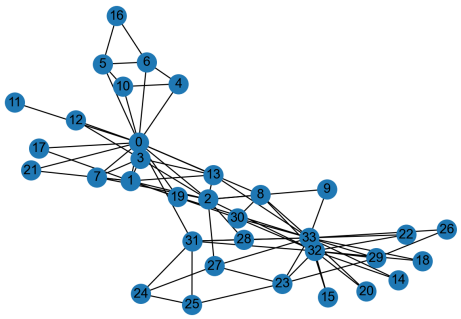

Figure 1: *Karate Club Network [15].*

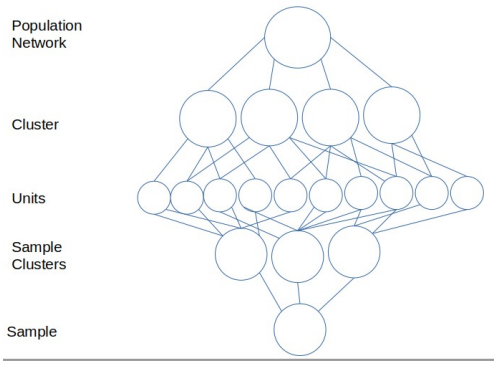

Figure 2: *Overlapping cluster sampling scheme diagram.*

Table 1: *Dataset Description.*

Description of Dataset(network)			
Network	vertex	Edge	Description
Karate	34	78	Zachary Karate Club Network [15]

4.1. Clique Based Procedure (CBP)

The computational procedure(CBP) prposed by authors is as under:
Step 1: Choose randomly K non-adjacent vertices vx_s vertices.
Step 2: Find cliques using vx_s as source vertex.
Step 3: Take degree of vertices of cliques in overlapping cluster.
Step 4: By SRSWR rule, choose k overlapping clusters from K clusters.
Step 5: By SRSWR rule, select sample of n_i vertices from N_i vertices among k clusters.

Table 2: *Clique of random vertices in Karate Club graph*

Cliques of vertices in Karate Club graph			
Serial No.	Vertices	Cliques	Degree sequence
S_1	vx_1	$[vx_0, vx_1, vx_2, vx_3, vx_7]$	(16, 9, 10, 6, 4)
S_2	vx_3	$[vx_0, vx_1, vx_2, vx_3, vx_{13}]$	(16, 9, 10, 6, 5)
S_3	vx_{15}	$[vx_{33}, vx_{32}, vx_{15}]$	(17, 12, 2)
S_4	vx_{28}	$[vx_{33}, vx_{28}, vx_{31}]$	(17, 3, 6)
S_5	vx_{22}	$[vx_{33}, vx_{32}, vx_{22}]$	(17, 12, 2)
S_6	vx_9	$[vx_2, vx_9]$	(10, 2)
S_7	vx_5	$[vx_5, vx_{16}, vx_6]$	(4, 2, 4)
S_8	vx_{10}	$[vx_0, vx_4, vx_{10}]$	(16, 3, 3)
S_9	vx_{12}	$[vx_0, vx_{12}, vx_3]$	(16, 2, 6)
S_{10}	vx_{30}	$[vx_{33}, vx_{32}, vx_8, vx_{30}]$	(17, 12, 5, 4)
S_{11}	vx_{14}	$[vx_{33}, vx_{32}, vx_{14}]$	(3, 3, 6)
S_{12}	vx_{23}	$[vx_{33}, vx_{27}, vx_{23}]$	(17, 4, 5)
S_{13}	vx_{26}	$[vx_{33}, vx_{26}, vx_{29}]$	(17, 2, 4)
S_{14}	vx_{20}	$[vx_{33}, vx_{32}, vx_{20}]$	(17, 12, 2)
S_{15}	vx_{11}	$[vx_0, vx_{11}]$	(16, 1)
S_{16}	vx_{21}	$[vx_0, vx_1, vx_{21}]$	(16, 9, 2)
S_{17}	vx_{19}	$[vx_0, vx_1, vx_{19}]$	(16, 9, 3)
S_{18}	vx_{31}	$[vx_{24}, vx_{25}, vx_{31}]$	(3, 3, 6)
S_{19}	vx_{17}	$[vx_0, vx_1, vx_{17}]$	(16, 9, 2)
S_{20}	vx_{18}	$[vx_{33}, vx_{32}, vx_{18}]$	(17, 12, 2)

4.2. Shortest Path Based Procedure (SPP)

Computaional procedure(SPP) existing in literature due to Dijkstra's algorithm [10] is as under:

Step 1: Choose randomly K pairs of non-adjacent vertices vx_s as source vertex and vx_d as destination vertex.
Step 2: Find shortest path between K pairs of vertices using shortest path algorithm through Dijkstra's algorithm [10].
Step 3: Degree sequence is formed to each vertex appearing in the computed shortest path.
Step 4: Take degree sequence as overlapping clusters which divide the graph vertices.

Step 5: By SRSWR rule, choose k overlapping clusters from K clusters.
Step 6: By SRSWR rule, choose sample n_i vertices from N_i among k clusters.

Table 3: *Shortest path of random pair of vertices in Karate Club graph*

Shortest path of random pair of vertices in Karate Club graph			
Serial No.	Pairs of vertices	Shortest Path	Degree sequence
S_1	(vx_{16}, vx_{26})	$[vx_{16}, vx_5, vx_0, vx_8, vx_{33}, vx_{26}]$	(2, 4, 16, 5, 17, 2)
S_2	(vx_{16}, vx_{26})	$[vx_{16}, vx_6, vx_0, vx_{19}, vx_{33}, vx_{26}]$	(2, 4, 16, 3, 17, 2)
S_3	(vx_{29}, vx_{12})	$[vx_{29}, vx_{32}, vx_2, vx_0, vx_{12}]$	(4,12,10,16,2)
S_4	(vx_{28}, vx_{16})	$[vx_{28}, vx_2, vx_0, vx_5, vx_{16}]$	(3, 10, 16, 4, 2)
S_5	(vx_{10}, vx_{15})	$[vx_{10}, vx_0, vx_{19}, vx_{33}, vx_{15}]$	(3,16, 3, 17,2)
S_6	(vx_{26}, vx_2)	$[vx_{26}, vx_{29}, vx_{32}, vx_2]$	(2, 4, 12, 10)
S_7	(vx_{25}, vx_7)	$[vx_{25}, vx_{31}, vx_0, vx_7]$	(3, 6, 16, 4)
S_8	(vx_{23}, vx_4)	$[vx_{23}, vx_{25}, vx_{31}, vx_0, vx_4]$	(5, 3, 6, 16, 3)
S_9	(vx_9, vx_{24})	$[vx_9, vx_2, vx_{27}, vx_{24}]$	(2,10, 4, 3)
S_{10}	(vx_{22}, vx_{24})	$[vx_{22}, vx_{32}, vx_{31}, vx_{24}]$	(2, 12, 6, 3)
S_{11}	(vx_{21}, vx_{27})	$[vx_{21}, vx_0, vx_2, vx_{27}]$	(2, 16, 10, 4)
S_{12}	(vx_{18}, vx_{25})	$[vx_{18}, vx_{32}, vx_{23}, vx_{25}]$	(2, 12, 5, 3)
S_{13}	(vx_{18}, vx_4)	$[vx_{18}, vx_{32}, vx_2, vx_0, vx_4]$	(2, 12, 10, 16, 3)
S_{14}	(vx_{17}, vx_5)	$[vx_{17}, vx_0, vx_2, vx_{32}, vx_1, vx_5]$	(2, 16, 10, 12, 9, 4)
S_{15}	(vx_{30}, vx_{25})	$[vx_{30}, vx_{32}, vx_{23}, vx_{25}]$	(4,12, 5, 3)
S_{16}	(vx_2, vx_{26})	$[vx_2, vx_8, vx_{33}, vx_{26}]$	(10, 5, 17, 2)
S_{17}	(vx_1, vx_{20})	$[vx_1, vx_2, vx_{32}, vx_{20}]$	(9, 10, 12, 2)
S_{18}	(vx_4, vx_8)	$[vx_4, vx_0, vx_2, vx_{32}, vx_1, vx_8]$	(3, 16, 10, 12, 9, 5)
S_{19}	(vx_3, vx_{26})	$[vx_3, vx_{13}, vx_{33}, vx_{26}]$	(6, 5, 17, 2)
S_{20}	(vx_{14}, vx_{11})	$[vx_{14}, vx_{32}, vx_{31}, vx_0, vx_{11}]$	(2,12, 6, 16, 1)

4.3. Frequency table of vertices in overlapping clusters

Table 4: *Frequency of vertices occuring in K-clusters*

Frequency of vertices in K-clusters, N = 34, $M_{sp} = 94, M_{cl} = 63$.											
vx	Y_{ij}	F_{ij}	F'_{ij}	D_{ij}	D'_{ij}	**vx**	Y_{ij}	F_{ij}	F'_{ij}	D_{ij}	D'_{ij}
vx_0	16	12	8	3.69	3.70	vx_{17}	2	1	1	5.53	3.70
vx_1	9	3	5	8.29	3.33	vx_{18}	2	2	1	2.76	3.70
vx_2	10	10	3	2.76	6.18	vx_{19}	3	2	1	4.15	5.56
vx_3	6	1	3	16.59	3.70	vx_{20}	2	1	1	5.53	3.70
vx_4	3	3	1	2.76	5.56	vx_{21}	2	1	1	5.53	3.70
vx_5	4	3	1	3.69	7.41	vx_{22}	2	1	1	5.53	3.70
vx_6	4	1	1	11.06	7.41	vx_{23}	5	3	1	4.61	9.26
vx_7	4	1	1	11.06	7.41	vx_{24}	3	2	1	4.15	5.56
vx_8	5	3	1	4.61	9.26	vx_{25}	3	4	1	2.07	5.56
vx_9	2	1	1	5.53	3.71	vx_{26}	2	5	1	1.11	3.70
vx_{10}	3	1	1	8.29	5.56	vx_{27}	4	2	1	5.53	7.41
vx_{11}	1	1	1	2.76	1.85	vx_{28}	3	1	1	8.29	5.56
vx_{12}	2	1	1	5.53	3.71	vx_{29}	4	2	1	5.53	7.41
vx_{13}	5	1	1	13.82	9.26	vx_{30}	4	1	1	11.06	7.41
vx_{14}	2	1	1	5.53	3.70	vx_{31}	6	4	2	4.15	5.56
vx_{15}	2	1	1	5.53	3.70	vx_{32}	12	10	6	3.32	3.70
vx_{16}	2	3	1	1.84	3.70	vx_{33}	17	5	9	9.4	3.5

In Table-2 and Table-3, overlapping clusters were collected using CBP and SPP. In which many vertices lie in clusters repeatedly. To use overlapping cluster sampling, degree values of vertices are normalized using its frequency by

$$D_{ij} = \frac{MY_{ij}}{NF_j}, \text{ i = 1, 2,..., K and j = 1, 2, ..., } N_i$$

where,
N = Total number of distinct vertices in network.
M = Total number of vertices in overlapping clusters.
M_{sp} = Total number of vertices in overlapping clusters obtained by SPP.
M_{cl} = Total number of vertices in overlapping clusters obtained by CBP.
Y_{ij} = Degree of vertices in network.
F_{ij} = Frequency of vertices in clusters formed by SPP.
F'_{ij} = Frequency of vertices in clusters formed by CBP.
D_{ij} = Normalised degree of vertices in clusters created by SPP.
D'_{ij} = Normalised degree of vertices in clusters created by CBP.

5. Experinmental Results

5.1. Ogive Based Simulation Procedure

Step 1: Draw sample of k clusters by SRSWR from K clusters(k < K).

Step 2: Draw sample of n_i vertices of second stage units from N_i among k clusters.

Step 3: Calculate lower limit and upper limit of confidence interval(CI).

Step 4: Repeat step I, II and III for P times(P is positive integer).

Step 5: Draw two ogive curves separately for lower limit and upper limit of confidence intervals.

Step 6: Draw a perpendicular from point of intersections of two ogive curves to find lower and upper limits of CI.

5.2. Numerical Illustration

Consider the Karate Club Network datasets (figure-1) which has N=34 identifiable distinct units(vertices). In Table-2 and Table-3 using method CBP and SPP overlapping clusters based on cliques and shortest path are obtained which contain each unit of the network. The objective is to estimate average degree of network and relative efficiency of estimate using confidence interval size. For numerical evaluation, one can take sample in two stages (figure-2). In the first stage sample of size k = 15 clusters are taken from K = 20 clusters. In the second stage sample of vertices from each clusters are taken randomly. Further, one can use ogive simulation for P times.

5.3. Shortest Path Based Procedure for Parameter Estimation [10]

A sample of cluster of size k = 15 of size K = 20 (Table-3) is taken by SRSWR and in each overlapping sampled cluster, a percentage of sample vertices are chosen randomly to calculate confidence intervals and average degree.

Table 5: *Sample cluster units (by SPP)*

Sampled vertices and degree sequence using SPP			
Serial No.	Sample pair of vertices	Shortest Path	Normalised Degree sequence
S_1	(vx_{16}, vx_{26})	$[vx_5, vx_0, vx_8, vx_{26}]$	(3.69, 3.69, 4.61, 1.11)
S_2	(vx_{16}, vx_{26})	$[vx_{16}, vx_6, vx_{19}, vx_{33}]$	(1.84, 11.06, 4.15, 9.4)
S_3	(vx_{29}, vx_{12})	$[vx_{29}, vx_2, vx_0, vx_{12}]$	(5.53, 2.76, 3.69, 5.53)
S_4	(vx_{10}, vx_{15})	$[vx_{10}, vx_{19}, vx_{33}, vx_{15}]$	(8.29, 4.15, 9.4, 5.53)
S_5	(vx_{26}, vx_2)	$[vx_{29}, vx_{32}, vx_2]$	(5.53, 3.32, 2.76)
S_6	(vx_{25}, vx_7)	$[vx_{25}, vx_{31}, vx_7]$	(2.07, 4.15, 11.06)
S_7	(vx_{23}, vx_4)	$[vx_{23}, vx_{25}, vx_0, vx_4]$	(4.61, 2.07, 3.69, 2.76)
S_8	(vx_9, vx_{24})	$[vx_9, vx_{27}, vx_{24}]$	(5.53, 5.53, 4.15)
S_9	(vx_{22}, vx_{24})	$[vx_{22}, vx_{32}, vx_{24}]$	(5.53, 3.32, 4.15)
S_{10}	(vx_{21}, vx_{27})	$[vx_{21}, vx_2, vx_{27}]$	(5.53, 2.76, 5.53)
S_{11}	(vx_{18}, vx_{25})	$[vx_{18}, vx_{32}, vx_{25}]$	(2.76, 3.32, 2.07)
S_{12}	(vx_{17}, vx_5)	$[vx_{17}, vx_0, vx_2, vx_5]$	(5.53, 3.69, 2.76, 3.69)
S_{13}	(vx_1, vx_{20})	$[vx_2, vx_{32}, vx_{20}]$	(2.76, 3.32, 5.53)
S_{14}	(vx_4, vx_8)	$[vx_4, vx_{32}, vx_1, vx_8]$	(2.76, 3.32, 8.29, 4.61)
S_{15}	(vx_{14}, vx_{11})	$[vx_{14}, vx_{32}, vx_0, vx_{11}]$	(5.53, 3.32, 3.69, 2.76)

Table 6: *Sample based Computation (for SPP)*

Sample based computation for confidence interval(using SPP)						
S. No.	Degree sequence	$\bar{d}_{(sp)i\bullet}$	$(\bar{d}_{i\bullet} - \bar{d})^2$	$s^2_{(sp)i}$	95% C.I.	CI size
S_1	(3.69, 3.69, 4.61, 1.11)	3.275	1.437	2.2713	[1.798, 4.752]	2.954
S_2	(1.84, 11.06, 4.15, 9.4)	6.6125	4.575	18.797	[2.364, 10.861]	8.497
S_3	(5.53, 2.76, 3.69, 5.53)	4.3775	0.009	1.915	[3.021, 5.734]	2.713
S_4	(8.29, 4.15, 9.4, 5.53)	6.8425	5.612	5.87	[4.468, 9.217]	4.749
S_5	(5.53, 3.32, 2.76)	3.87	0.364	2.145	[2.213, 5.527]	3.314
S_6	(2.07, 4.15, 11.06)	5.76	1.654	22.15	[0.434, 11.086]	10.652
S_7	(4.61, 2.07, 3.69, 2.76)	3.2825	1.419	1.224	[0.434, 11.086]	10.652
S_8	(5.53, 5.53, 4.15)	5.07	0.356	0.6348	[4.168, 5.972]	1.804
S_9	(5.53, 3.32, 4.15)	4.33	0.021	1.246	[3.067 ,5.593]	2.526
S_{10}	(5.53, 2.76, 5.53)	4.607	0.018	2.557	[2.798, 6.416]	3.618
S_{11}	(2.76, 3.32, 2.07)	2.717	3.086	0.392	[2.009, 3.425]	1.416
S_{12}	(5.53, 3.69, 2.76, 3.69)	3.9175	0.309	1.348	[2.780, 5.055]	2.275
S_{13}	(2.76, 3.32, 5.53)	3.87	0.364	2.145	[2.213, 5.527]	3.314
S_{14}	(2.76, 3.32, 8.29, 4.61)	4.7475	0.075	6.185	[2.308, 7.182]	4.874
S_{15}	(5.53, 3.32, 3.69, 2.76)	3.825	0.421	1.438	[2.650, 5.000]	2.35
	Average Value	$\bar{d}_{sp} = 4.4736$	$s^2_{sp} = 1.408$	$\hat{s}^2_i = 4.688$	[2.4483,6.8288]	4.3805

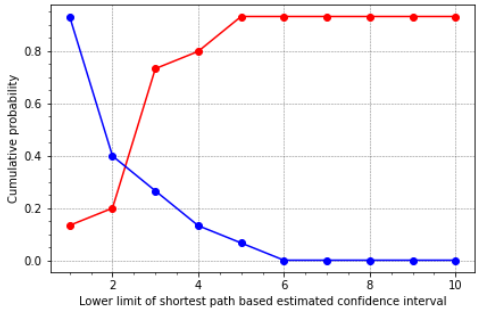

Figure 3: *Ogive for lower limit of CI for SPP.*

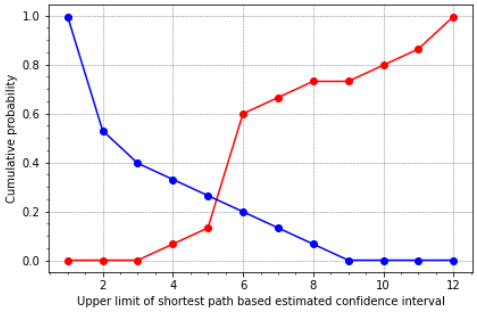

Figure 4: *Ogive for upper limit of CI for SPP.*

$$\widehat{Var}(\bar{d}_{(sp)}) = \left(\frac{1}{k} - \frac{1}{K}\right) \hat{S}_{sp}^2 + \frac{1}{kK} \sum_{i=1}^{k} \left(\frac{1}{n_i} - \frac{1}{N_i}\right) \hat{S}_i^2$$

Estimated average degree = $\bar{\mathbf{d}}_{\mathbf{sp}} = \mathbf{4.47}$
Estimated variance for average degree = $\widehat{Var}(\bar{d}_{(sp)})$ = 1.06
Average CI size = 4.38
A 95% confidence interval estimate using SPP for average degree is $[\mathbf{2.4483}, \mathbf{6.8288}]$.
Through ogive based simulation(figure 3 & 4) for average degree the confidence interval is $[\mathbf{2.38}, \mathbf{5.42}]$

5.4. Clique Based Procedure (CBP) for Parameter Estimation

Consider sample of cluster clique having size k = 15 of size K = 20(Table-2) by SRSWR in each overlapping sampled cluster. Herein a percentage is used to calculate confidence intervals and average degree of network.

Table 7: *Clique of random vertex in Karate Club Graph (by CBP)*

Sampled vertices of clique using CBP			
Serial No.	Sample vertices	Sample cliques	Normalised degree sequence
S_1	vx_3	$[vx_0, vx_1, vx_3, vx_{13}]$	(3.70, 3.33, 3.70, 9.26)
S_2	vx_{15}	$[vx_{32}, vx_{15}]$	(3.70, 3.70)
S_3	vx_{28}	$[vx_{33}, vx_{28}]$	(3.5, 5.56)
S_4	vx_9	$[vx_2, vx_9]$	(6.18, 3.71)
S_5	vx_5	$[vx_5, vx_{16}]$	(7.41, 3.70)
S_6	vx_{10}	$[vx_0, vx_4]$	(3.70, 5.56)
S_7	vx_{12}	$[vx_{12}, vx_3]$	(3.71, 3.70)
S_8	vx_{30}	$[vx_{33}, vx_{32}, vx_{30}]$	(3.5, 3.70, 7.41)
S_9	vx_{23}	$[vx_{33}, vx_{27}]$	(3.5, 7.41)
S_{10}	vx_{26}	$[vx_{33}, vx_{26}]$	(3.5, 3.70)
S_{11}	vx_{20}	$[vx_{33}, vx_{20}]$	(3.5, 3.70)
S_{12}	vx_{11}	$[vx_0, vx_{11}]$	(3.70, 1.85)
S_{13}	vx_{21}	$[vx_0, vx_1]$	(3.70, 3.33)
S_{14}	vx_{31}	$[vx_{24}, vx_{31}]$	(5.56, 5.56)
S_{15}	vx_{18}	$[vx_{33}, vx_{18}]$	(3.5, 3.70)

Table 8: *Sample based Computation (for CBP)*

Sample based computation for confidence interval(Using CBP)						
S. No.	Degree sequence	$\bar{d}_{(cl)i\bullet}$	$(\bar{d}_{i\bullet} - \bar{d})^2$	$s^2_{(cl)i}$	95% C.I.	CI size
S_1	(3.70, 3.33, 3.70, 9.26)	4.99	0.436	2.2713	[2.207, 7.788]	5.581
S_2	(3.70, 3.70)	3.7	0.397	0	[3.700, 3.700]	0
S_3	(3.5, 5.56)	4.53	0.04	2.122	[2.511, 6.549]	4.038
S_4	(6.18, 3.71)	4.94	0.325	3.05	[2.525, 7.365]	4.84
S_5	(7.41, 3.70)	5.55	1.488	6.882	[1.914, 9.186]	7.272
S_6	(3.70, 5.56)	4.63	0.09	1.73	[2.808, 6.452]	3.644
S_7	(3.71, 3.70)	3.70	0.397	0.0001	[3.695, 3.715]	0.02
S_8	(3.5, 3.70, 7.41)	4.87	0.292	4.849	[2.378, 7.362]	4.984
S_9	(3.5, 7.41)	5.45	1.254	7.644	[1.623, 9.287]	7.664
S_{10}	(3.5, 3.70)	3.6	0.533	0.02	[3.404, 3.796]	0.392
S_{11}	(3.5, 3.70)	3.6	0.533	0.02	[3.404, 3.796]	0.392
S_{12}	(3.70, 1.85)	2.77	2.434	1.7113	[0.962, 4.588]	3.626
S_{13}	(3.70, 3.33)	3.51	0.672	0.0685	[3.152, 3.878]	0.726
S_{14}	(5.56, 5.56)	5.56	1.513	0	[5.560, 5.560]	0
S_{15}	(3.5, 3.70)	3.6	0.533	0.02	[3.404, 3.796]	0.392
	Average value	$\bar{d}_{cl} = 4.33$	$s^2_{cl} = 0.781$	$\hat{s}^2_i = 0.4876$	[2.883, 5.787]	2.9047

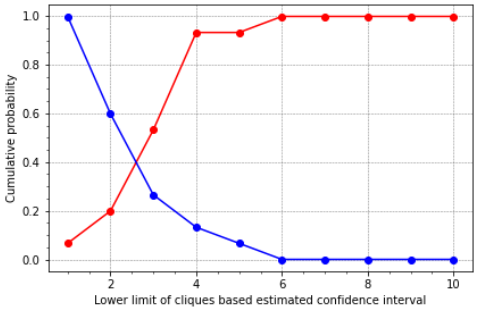

Figure 5: *Ogive for lower limit of CI by CBP.*

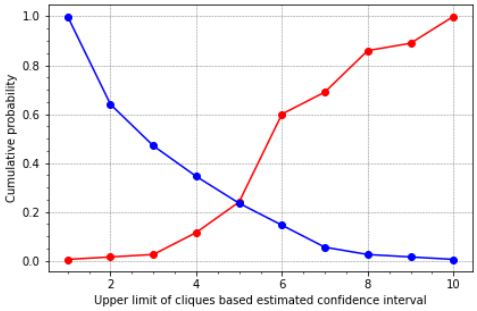

Figure 6: *Ogive for upper limit of CI by CBP.*

$$\widehat{Var}(\bar{d}) = \left(\frac{1}{k} - \frac{1}{K}\right) \hat{S}^2_{cl} + \frac{1}{kK} \sum_{i=1}^{k} \left(\frac{1}{n_i} - \frac{1}{N_i}\right) \hat{S}^2_i$$

Estimated average degree = $\bar{d}_{cl} = 4.33$
Estimated variance for average degree = $\widehat{Var}(\bar{d}_{cl}) =$ **0.022813**
Average confidence interval(CI) size = $[5.787 - 2.883] =$ **2.90**
The 95% confidence interval estimate using CBP is [**2.8831, 5.7878**].
Through ogive based simulation(figure 5 & 6) the confidence interval(CI) is [**2.63, 5.01**].

6. Comparision

The Percentage Relative Efficiency (PRE) of estimators $\bar{d}_{cl}$, $\bar{d}_{sp}$ is defined as under:

$$\textbf{PRE} = \frac{\widehat{Var}(\bar{d}_{sp}) - \widehat{Var}(\bar{d}_{cl})}{\widehat{Var}(\bar{d}_{sp})} \times 100 = \frac{1.06 - 0.0228}{1.06} \times 100 = 97.84\%$$

The Percentage Relative Gain(PRG) over the length of confidence intervals is defined as:

$$\textbf{PRG} = \frac{(\text{length of CI})_{SPP} - (\text{length of CI})_{CBP}}{(\text{length of CI})_{SPP}} \times 100 = \frac{4.3805 - 2.9047}{4.3805} \times 100 = 33.69\%$$

Using ogive based simulation, the Percentage Relative Gain is:

$$(\textbf{PRG})_{ogive} = \frac{[(\text{length of CI})_{SPP}]_{ogive} - [(\text{length of CI})_{CBP}]_{ogive}}{[(\text{length of CI})_{SPP}]_{ogive}} \times 100 = \frac{3.04 - 2.38}{3.04} \times 100 = 21.71\%$$

7. Reliability of Social Networks as an Application

As considered, the average degree estimation of a social network leads to monitoring the reliability of the network. People join the social network at any point of time and leave it at any other instant. Addition and deletion in a social network is a common continuous process. A network is said to be reliable if the average degree of social network remains controlled over the time framework. The upper limit and lower limit of confidence intervals are useful measures to make a benchmark for checking of growth or decay of social networks over the time domain.

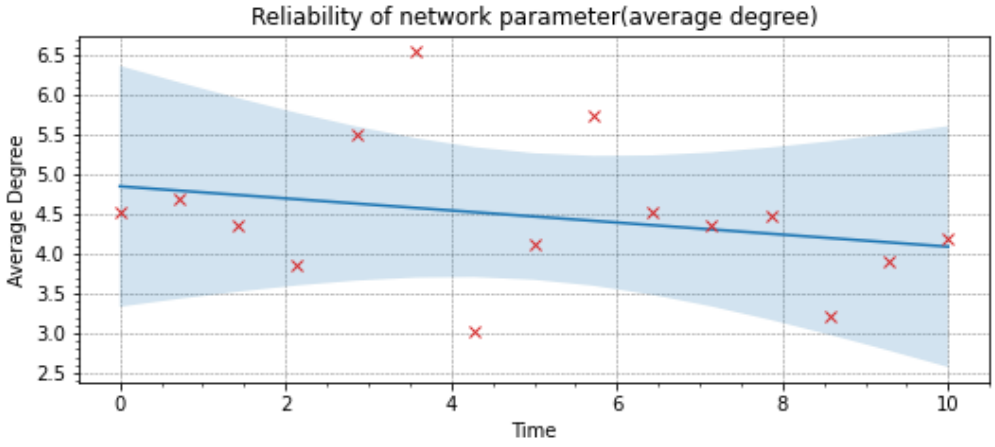

Figure 7: *Network reliabiity based on vertex degree estimate.*

8. Discussion

The two methods SPP and CBP are compared in a common setup of a social network for the objective of computation of average degree. A social network in general can be represented as a graph of vertices and edges. Clusters of vertices are formed by using both methods SPP and CBP. After comparison of percentage relative efficiency, the CPP found efficient by 97.8% over SPP. The simulated confidence interval for clique procedure (CBP) is [2.8-5.7] which is catching the true value of average degree 4.58 of vertices, which is also supported by figures 5 and 6. The same calculation for simulated confidence interval using shortest path procedure(SPP) is [2.4-6.8] which is longer than earlier (see figure 3 & 4). Ogive based simulation procedure also supports

for better efficiency of [(2.38, 5.42) for SPP, (2.63, 5.01) for CBP] the proposed for network degree evaluation and the proposed is useful for network reliability (figure-7).

9. Conclusion

This paper contains an overlapping cluster sampling based comparative approach using the shortest path and cliques method over created clusters. A graphical structure has been taken under consideration representing the social network. In order to estimate the unknown parameter(like average degree), the proposed sampling method takes into account the cliques and compares with shortest paths between several pairs of vertices in a setup of the overlapping cluster of degree sequence. The proposed method is examined by conducting an experiment on a well-known real network keeping in view that the average degree is an important property of network. To evaluate the comparative statistical significance of proposed procedure CBP, the 95% confidence intervals were computed for both methods. It has been found as an outcome of the study that 95% confidence intervals contain the true value. The Ogive based simulation procedure has been implemented which shows cluster based method using clique (CBP) provides a better estimate of the parameter(average degree) than the cluster based method using shortest path (SPP). The network reliability could be monitored over the long time domain by the bench-mark values of confidence intervals. This contribution opens up new avenues and opportunities for network degree parameter estimation. One can think of the inclusion of the additional network measures for future studies that will help to bring up new insights to the development of graph sampling cluster methods. In order to have more a comprehensive evaluation of the existing social networking, the sampling methods could be considered by involving the other kinds of parametric network measures and properties.

References

[1] Cochran, W. G. (2005). Sampling Techniques, *John Willey and Sons, New York*.
[2] Singh S. (1988). Estimation in overlapping clusters, *Communications in Statistics: Theory and Methods*, 17:613-621.
[3] Rezvanian, A. and Meybodi, M. R. (2015). Sampling social networks using shortest paths, *Physica A: Statistical Mechanics and its Applications*, 424: 254-268.
[4] Zhang, LC. and Patone, M. (2017). Graph sampling. *Metron*, 75: 277-299.
[5] Alim, A. and Shukla, D. (2021). Double sampling based parameter estimation in big data and application in control charts. *Reliability: Theory & Applications,* 16(2 (62)): 72-114.
[6] Katzir, L., Liberty, E., Somekh, O. & Cosma, I. A. (2014). Estimating sizes of social networks via biased sampling, *Internet Mathematics*, 10:3-4, 335-359.
[7] Kurant, M.,Butts, C. T. and Markopoulou, A. (2012). Graph size estimation. *CoRR, abs/1210.0460*.
[8] McCormick, T. H., Moussa, A., Ruf, J., DiPrete, T. A., Gelman, A., Teitler J., and Zheng, T. (2013). A practical guide to measuring social structure using indirectly observed network data. *Journal of Statistical Theory and Practice*, 7(1):120–132.
[9] McCormick, H., Salganik, M.J. and Zheng, T. (2010). How many people do you know?: Efficiently estimating personal network size. *JASA*, 105(489):59–70.
[10] Dijkstra, E. W. (1959). A note on two problems in connexion with graphs, *Numerische Mathematik*, 1(1):269–271.
[11] Newman, M.E. (2001). Scientific collaboration networks. II. Shortest paths, weighted networks and centrality, *Physical Review E Stat Nonlin Soft Matter Phys.*, 64(1):016132.
[12] Chen, W., Wang, C. and Wang, Y. (2010). Scalable influence maximization for prevalent viral marketing in large-scale social networks, *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovxery and Data Mining*, pp. 1029–1038.
[13] Milgram, S. (1967). The small world problem, *Psychology Today*, 2(1): 60–67

[14] Travxers, J. and Milgram, S. (1969). An experimental study of the small world problem, *Sociometry* ,32(4):425–443.

[15] Zachary, W. W. (1977). An information flow model for conflict and fission in small groups, *Journal of Anthropological Research* 33(4): 452–473.

[16] Milan, R. and Shukla, D. (2021). Kernel sampling based parameter estimation in detected community in weighted graph in big data. *Reliability: Theory & Applications*, 16(4 (65)), 105-120.

[17] Pandey, K. K. and Shukla, D. (2021). Stratified linear systematic sampling based clustering approach for detection of financial risk group by mining of big data. *International Journal of System Assurance Engineering and Management*, 1-15.

[18] Lee, C., Reid, F., McDaid, A., Hurley, N. (2010). Detecting highly overlapping community structure by greedy clique expansion. *SNA-KDD: Social Network Mining and Analysis* pp. 33–42.

[19] Shang, R., Luo, S., Li, Y., Jiao, L. and Stolkin, R. (2015). Large-scale community detection based on node membership grade and sub-communities integration. *Physica A: Statistical Mechanics and its Applications*, 428: 279-294.

[20] Shen, H., Cheng, X., Cai, K. and Hu, M. B. (2009). Overlapping and hierarchical community structure in networks. *Physica A: Statistical Mechanics and its Applications*, 388(8): 1706-1712.