International Journal of Mathematics, Statistics and Operations Research
Vol. 5; Number 1; 2025, pp. 41-56
https://doi.org/10.47509/IJMSOR.2025.v05i01.03

ARF INDIA
Academic Open Access Publishing
www.arfjournals.com

Cluster Creation and Comparison in Social Network in View to Risk of Covid-19 Pandemic using PPSWR Cluster Sampling

Vivek kumar Gupta[1], Diwakar Shukla[2] and Astha Jain[3]*
[1,2,3] Department of Mathematics and Statistics,
Dr. Harisingh Gour Vishwavidyalaya, Sagar, M.P., India

Abstract

In the recent past, the Covid-19 pandemic has caused trouble due to lockdown in their day-to-day life of the people around the world. The risk of exposure of such major source of spread of disease is caused by human interaction with others or touching objects around the infected person, which ultimately creates a chain or clusters of infected people. Graphs are used to model the infected person's network. In recent past, graph sampling is used to draw sample subgraphs from network(graph) in order to study different network parameters. This paper presents a comparison of clique based procedure (CBP) and shortest path based procedure (SPP) to estimate the average degree of Covid-19 patient network using an overlapping cluster sampling. A probability proportional to size based cluster sampling 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 the clique based sampling algorithm (CBP) for risk evaluation is more efficient (relative gain of 13.26%) than the shortest path based sampling algorithm (SPP). The estimate of average degree of patient network provides a value about the nature of spread of infection that a patient has generated in society. This estimate also indicates the intensity of risk of spread of disease of a particular type of Covid-19 infection variant. Results of this study showed that this model can be used as a valuable tool for design based sampling on network like structure.

Covid-19, Clique, Cluster Sampling, Confidence Interval (CI), Graph, Sampling, Pandemic, Probability proportional to size with replacement(PPSWR), Shortest Path, Social Network

*CONTACT Author[3]. Email:asthajain2597@gmail.com

Article History
Received : 04 January 2025; Revised : 06 February 2025; Accepted : 05 June 2025; Published : 26 June 2025

To cite this paper

Vivek kumar Gupta, Diwakar Shukla and Astha Jain (2025). Cluster Creation and Comparison in Social Network in View to Risk of Covid-19 Pandemic using PPSWR Cluster Sampling. *International Journal of Mathematics, Statistics and Operations Research.* 5(1), 41-56.

1 Introduction

In couple of years due to spread of the Covid-19 pandemic, a specific situation has appeared in the world. This disease spreads through human interaction with infected people [14]. Interaction for various reasons creates a network assumed as a social network of contacts of persons with infected [11], covid transmission network, infection transport network [10], etc. To prevent such various steps are recommended by medical councils, one of them is to identify and isolate the infected person by which the chain of transmission can be broken. Covid-19 transmits through the touch of surfaces, objects, patients, and inhalation of droplets of an infected person. After identifying and isolating the patient next step is to find close contacts of infected who shared close air space usually for a prolonged period [5]. Close contact forms a cluster of persons which may be part of a chain of spread. Once cases of Covid are identified and received appropriate treatment, the priority of Covid control program is to classify contacts into active or passive cases, some cases may be without symptoms but infected. Active cases are at high risk to develop new Covid patients and areas where treatment of patients with active Covid is assured. In an outbreak, there may be several active cases, with each having many contacts that may or may not overlap. To perform a quick test one can select a clique or shortest path [15][19] between Covid contacts (nodes) as a sample instead of studying all the contacts. The usual sampling procedure in such a situation shall be the two-stage cluster sampling to find out the unknown as it is easy, fast, convenient, and economical.

For dealing with sampling scheme contacts of different active cases spread in groups can be treated as clusters, which are selected at the first stage. In the second stage, a few contacts are selected from each selected cluster. The sampling strategy uses as a tool for studying Covid-19 transmission, but this strategy is equally good and applicable to other contagious diseases also. Sometimes to avoid preparation of an aggregate list of units, cluster sampling is used. Clusters are formed either before selecting the sample (CBS) or after selecting the sample. In most practices, the CBS system of cluster formation is used and clusters may be either overlapping or non-overlapping [7],[13]. For non-overlapping many methodologies are in the literature on survey sampling, but, for overlapping clusters, the literature is limited [6].

2 Definition and Related Work

A undirected network(graph) [11] G = (V, L) contains two component, set of nodes V = $v_1, v_2, \ldots, v_N$and set of links (edges) L = $l_1, l_2, \ldots, l_M$. A link (edge) li is connection between a pair of nodes (v_i, v_j). For directed graph (v_i, v_j) is not same as (v_j, v_i). The network G is simple, undirected, and unweighted which contains N number of nodes and M number of links. In an undirected graph, degree of a node is defined as the number of links incident on a node. A subgraph is a subset of the nodes and links between the nodes. A path is a subgraph of the network defined as a continuous sequence of links and nodes.

The shortest path between pair of nodes in a graph is a path between nodes such that weights of its constituent edges are minimized. An undirected graph or subgraph is called complete if every pair of distinct nodes is connected and has unique links. A maximal complete subgraph of a network is called a clique [7][11]. The word "maximal" means that no other nodes can be added to the clique without making it less connected.

2.1 Motivation

During a pandemic, a single infected person may be a major source of spread of infection [5],[16] in many venues like homes, offices, community centers [12][14], etc. Therefore, clusters sampling may include many overlapping clusters [13]. In network, the Clique [11] is used to detect overlapping communities. It is an agglomerative algorithm that recursively merges similar vertices and communities. The best possible community structure [19][20] occurs when all vertices of a subgraph are connected, i.e. when they form a clique, the clique uses the k-cliques present in a graph to determine communities. Two k-cliques are said to be adjacent if they differ by only one vertex, and a collection of adjacent k-cliques then forms a community. This allows communities to overlap as a vertex may be part of two separate k-cliques. In network, the transmission of information from one node to another is passed through the shortest path [11]. Similarly, for tracing of infected person one can use shortest path between two infected persons in a community [10][15].

2.2 Confidence Interval (CI)

A confidence interval [3][16], is an interval used to obtain an estimated range that is likely to include an unknown population parameter, the estimated range of confidence intervals is obtained by a set of sample data. The 95% confidence interval is defined as P $[a < \theta < b] = 0.95$, where θ is an unknown parameter and a, b are real numbers.
where,

$$a = \text{Estimated average} - 1.96\sqrt{\text{Estimated variance}}, \tag{1}$$

$$b = \text{Estimated average} + 1.96\sqrt{\text{Estimated variance}}, \tag{2}$$

.

2.3 Probability proportional to size with replacement(PPSWR) cluster sampling [3][13]

Let M_0 be the total number of units in all clusters and M_i represents size of i^{th} clusters. The unbiased estimate of population total Y is

$$\hat{Y} = \frac{M_0}{N}\left(\sum_{i=1}^{n} \bar{y}_i\right) \tag{3}$$

where, $\bar{y}_i$ is the sample mean of i^{th} clusters of size M when n clusters are in sample.
The variance under PPSWR sampling is

$$Var\left(\bar{Y}\right) = \left(\frac{M_0}{N}\right) \sum_{i=1}^{N} M_i \left(\bar{y}_i - \bar{y}\right)^2 \tag{4}$$

where, $\bar{Y}$ represent the population aggregate mean. One can consider the size of hospital in terms of number of beds available or number of registration of patients in order to estimate unknown parameter like average discharge time of admitted patient among all Covid-19 hospitals in a city a Covid-19 hospital is a cluster. Let M_i be the measure of size of cluster (i.e. hospital of Covid-19 patients) and M_0' be $=\sum_{i=1}^{N} M_i'$.
Define,

$$y_i = \frac{M_i'}{M_0'} \tag{5}$$

then

$$\hat{Y}_{PPS} = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{y_i}{z_i}\right) \tag{6}$$

where $\hat{Y}_{PPS}$ is an unbiased estimate of population total with variance.

$$Var(\hat{Y}_{PPS}) = \frac{1}{n} \sum_{i=1}^{n} g_i \left(\frac{y_i}{g_i} - \hat{Y}_{PPS}\right)^2 \tag{7}$$

unbiased estimate of $Var(\hat{Y}_{PPS})$is

$$est[Var(\hat{Y}_{PPS})] = \frac{1}{n(n-1)} \sum_{i=1}^{n} g_i \left(\frac{y_i}{g_i} - \hat{Y}_{PPS}\right)^2 \tag{8}$$

where $\hat{Y}_{PPS}$ is the estimated sample mean used for $\hat{Y}$.

3 Overlapping Cluster Sampling [13]

Let P be a population network of an outbreak under investigation of N distinct and identifiable contacts of K active cases. These N contacts may be expressed in the form of K overlapping clusters [9][13] with $N_i (i = 1, 2, ..., K)$ contacts associated with the i^{th} active case and

$$\sum_{i=1}^{K} N_i = M \geq N \tag{9}$$

The equality holds for non-overlapping clusters, which is a situation where no contact is common to more than one active case. But in fact, a contact may

be associated with more than one active case and let F_{ij} be the frequency of j^{th} contact occurring in clusters of active cases. Let d be the characteristic of study and the parameter of interest be the population mean $\bar{A}$. Define,

$$A_{ij} = \frac{MD_{ij}}{NF_{ij}}; i = 1, 2, \ldots, K \text{and} j = 1, 2, \ldots, N. \tag{10}$$

where D_{ij} is the value of d for the j^{th} contact in the i^{th} cluster. The proposed strategy in two stage-sampling consists of the following steps :

- Let k clusters out of K be selected by probability proportional to size with replacement (PPSWR) with initial probabilities $P_i = \frac{N_i}{M}$, $i = 1, 2, \ldots, K.(k < K)$
- If i^{th} cluster is selected α_i times in the sample at the first stage, then at the second stage a random sample of size $\alpha_i n_i$ contacts is selected by simple random sampling without replacement (SRSWOR) from the i^{th} selected cluster of size N_i, assuming $\alpha_i n_i \leq N_i, i = 1, 2, \ldots, K$.

This strategy leads to a fully without replacement sample, i.e., there is no chance of repetition in the ultimate sample of contacts.

Theorem 3.1. *An unbiased estimator for population mean $\bar{A}$ is given by*

$$\bar{a}_{PPSW} = \frac{1}{K} \sum_{i=1}^{K} \frac{1}{n_i} \sum_{j}^{\alpha_i n_i} a_{ij} \tag{11}$$

where α_i's are Binomial random variables with parameters k and $P_i = \frac{N_i}{M}$, $\sum_{i=1}^{K} \alpha_i = k$ and a_{ij} corresponds to the sample value for the j^{th} contact in the i^{th} cluster. where,

$$\bar{A} = \frac{1}{K} \sum_{i=1}^{K} \sum_{1}^{N_i} A_{ij}$$

Proof. Let E_1 denotes the conditional expectation for a given sample of clusters and E_2 the expectation over all such samples, then we have

$$E(\bar{a}_{PPSW}) = E_1 \frac{1}{K} \sum_{i=1}^{K} \alpha_i E_2(\bar{a}_i) = \frac{1}{k} \sum_{i=1}^{K} E_1(\alpha_i) \bar{A}_i \tag{12}$$

where $\bar{a}_i$represents the ith cluster mean over $\alpha_i m_i$ contacts. Since $E_1(\alpha_i) = \frac{KN_i}{M}$ and transforming A_{ij}, one can get

$$E(\bar{a}_{PPSW}) = \frac{1}{K} \sum_{i=1}^{K} \frac{1}{n_i} \sum_{j}^{\alpha_i n_i} a_{ij}$$

□

Theorem 3.2. *The variance of the estimator* $\overline{a}_{PPSW}$ *is given as*

$$Var\left(\overline{a}_{PPSW}\right) = \frac{\sigma^2_{ba'}}{k} + \frac{1}{k}\sum_{i=1}^{K}\frac{N_i}{M}\left(\frac{1}{n_i} - \frac{1}{N_i}\right)S^2_{ia} - \frac{k-1}{k}\sum_{i=1}^{K}\frac{N_i}{M^2}S^2_{ia}, \quad (13)$$

where

$$\sigma^2_{ba'} = \sum_{i=1}^{K}\frac{N_i}{M}\left(\overline{A}_i - \overline{A}\right)^2,\ S^2_{ia} = \sum_{j=1}^{N_i}\frac{\left(\overline{A_i} - \overline{A}\right)^2}{(N_i - 1)} \quad (14)$$

and

$$\overline{A} = \sum_{i=1}^{K}\frac{N_i}{M}\overline{A_i} = \overline{Y}.$$

Theorem 3.3. *An unbiased estimator of* $Var\left(\overline{a}_{PPSW}\right)$ *is given by*

$$\widehat{Var}\left(\overline{a}_{PPSW}\right) = \frac{s^2_{ba'}}{k} + \frac{1}{k\,(k-1)}\sum_{i=1}^{k}\left(\frac{1}{n_i} - \frac{1}{n_i\alpha_i}\right)s^2_{ia'} - \frac{1}{k}\sum_{i=1}^{k}\frac{s^2_{ia'}}{M} \quad (15)$$

where,

$$s^2_{ba'} = \frac{1}{k-1}\sum_{i=1}^{K}\alpha_i(\overline{a}_{i.} - \overline{a}_{PPSW})^2, s^2_{ia'} = \frac{1}{\alpha_i n_i - 1}\sum_{j=1}^{\alpha_i n_i}(a_{ij} - \overline{a}_{i.})^2 \quad (16)$$

and $\overline{a}_{i.}$ is the i^{th} cluster mean over $\alpha_i n_i$ contacts.

4 Proposed Sampling Algorithm and Dataset

Graph sampling techniques [6][7][11] are of mainly two types: random selection and graph exploration techniques. In random selection, nodes or edges are selected as samples uniformly at random or proportional to some particular characteristic like the degree of a node or its Page Rank score. In graph exploration, a sample is obtained by taking a neighborhood of a randomly selected seed node using random walks, shortest path [8][13], or another strategy. In this paper, samples are taken by exploration.

4.1 General Computational Algorithm

The computational algorithm for estimation of average degree of a given network G(V, E), where V = set of nodes $(v_1, v_2, \ldots, v_N)$, E = set of edges $e_1, e_2, \ldots, e_M$ is expressed under:

- Randomly select a K pair or set of vertices from G using simple random sampling [11][16].
- Form overlapping clusters using different procedures.

- Create a degree sequence of nodes (clusters).
- Estimate network parameters [8] using overlapping clusters [9] sampling mean estimation method [3].

To estimate network parameters, samples of nodes or edges are required. For drawing a sample, one can use different methodologies like random node, random edge, etc. In this paper, to collect network samples, shortest path method and method using cliques between random pair of vertices method are used to compute by which nodes of network are collected and clustered.

4.2 Shortest Path Procedure(SPP)

Shortest path between pairs can be obtained by various algorithms such as Dijkstra's algorihm [4], Prim's algogrithm etc. In the proposed, Dijkastra's algorithm [4] based Shortest path procedure (SPP) is used and described as under :
The Computaional procedure SPP proposed is as under:

Step 1: Select K random pair of nodes from G.

Step 2: Using Dijkstra's algorithm [4] find shortest path between K pairs of nodes.

Step 3: Find degree of selected nodes of shortest paths which form degree sequence.

Step 4: Degree sequence are taken as overlapping cluster units.

Step 5: By SRSWR rule, select a sample of k cluster units from K overlapping cluster.

Step 6: By SRSWR rule, sample n_i units from N_i nodes within k cluster.

4.3 Clique Based Procedure (CBP)

Cliques are complete subgraph used to form community [11][16], component and clusters. In this, K nodes are randomly chosen as seed nodes to find cliques of different lengths who ultimately form clusters [18]. Such clusters create clique node degree sequence.
The computational procedure CBP is proposed as under:

Step 1: Select K nodes randomly.

Step 2: Obtain cliques using v_i as seed node.

Step 3: Find degree of selected nodes of cliques which constitute overlapping cluster.

Step 4: By SRSWR, choose k clusters from K overlapping clusters.

Step 5: By SRSWR, select sample of ni size from $N_i(n_i < N_i)$ nodes from k clusters.

To evaluate and compare the efficiency of proposed methodologies, a real-world datasets has been considered which represents of Covid-19 trasmission network. The recognition that growth and preferential attachment [1] work simaltaneously in real networks has inspired a minimal model called the Barabási-Albert model [1][2], which are used to generate scale-free networks. It is defined as follows: Start with m_0 nodes, the links between nodes are chosen arbitrarily, as long as each node has at least one link. The network results into following:

- At each time step uset adds a new node with $m(\leq m_0)$ links that connect the new node to m nodes already in the network.
- The probability Π(d) that a link of new node connects to node i depends on the degree d_i as $\prod(d_i) = \frac{d_i}{\sum_{i=1} d_i}$ $(i = 1, 2, 3, \ldots, m)$

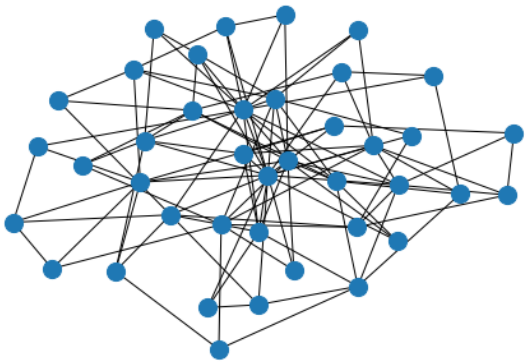

Figure 1: Albert Barabasi Graph

Table 1: Description of Dataset(network)

Network Description	Node	Edges	m
Albert-Barabasi Network [2]	40	111	3

By Using SPP and CBP, degree sequence are created using table-2.

To estimate average degree by overlapping cluster sampling mean methodologies, aggregate unit value must be normalized. Suppose a network exists of size N nodes(vertices) each with degrees $d_i, i = 1, 2, \ldots, N$ representing the connections (Number of infected person by node i) between node i and all other nodes. The $A_{ij} = \frac{MD_{ij}}{F_{ij}}$ $i = 1, 2, \ldots, N$are used to normalized degrees by equation (3.1) where, N = number of nodes.

Table 2: Number of Contacts in Covid-19 Trasmission Network Model (fig 1)

Person (node)	Infected Person (degree)	Person (node)	Infected Person (degree)	Person (node)	Infected Person (degree)
V_0	5	V_{14}	4	V_{28}	4
V_1	7	V_{15}	8	V_{29}	3
V_2	11	V_{16}	5	V_{30}	3
V_3	14	V_{17}	7	V_{31}	3
V_4	16	V_{18}	3	V_{32}	3
V_5	11	V_{19}	4	V_{33}	4
V_6	20	V_{20}	5	V_{34}	3
V_7	9	V_{21}	3	V_{35}	4
V_8	5	V_{22}	3	V_{36}	3
V_9	6	V_{23}	3	V_{37}	3
V_{10}	5	V_{24}	4	V_{38}	3
V_{11}	3	V_{25}	4	V_{39}	3
V_{12}	8	V_{26}	3		
V_{13}	4	V_{27}	3		

M = number of nodes in overlapping clusters.
M_s = number of nodes in overlapping clusters formed by SPP.
M_c = number of nodes in overlapping clusters formed by CBP.
F_s= Frequency of nodes in clusters formed by SPP.
F_c= Frequency of nodes in clusters formed by CBP.
A_{ij}^S= Normalised degree of nodes in clusters created by SPP.
A_{ij}^C= Normalised degree of nodes in clusters created by CBP.

By using SPP and CBP, multiple samples are obtained, as displayed in Table-3.

5 Ogive Based Simulation Procedure

To find estimated confidence interval (CI) of estimated unknown parameters, ogive simulation procedure is whose steps are as under :

Step 1: Take k cluster units by SRSWR from K clusters (k < K).

Step 2: Sample ni nodes from N_i within k clusters.

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

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

Table 3: Degree Sequence by SPP and CBP using table-2

S. No.	Seed Pair of Node	Shortest Path	Degree Sequence	Seed Node	Clique	Degree Sequence
S_1	(V_3, V_{36})	$[V_3, V_0, V_{12}, V_{36}]$	[14, 5, 8, 3]	V_4	$[V_3, V_{18}, V_4, V_5]$	[14, 3, 16, 11]
S_2	(V_{11}, V_{22})	$[V_{11}, V_1, V_3, V_{22}]$	[3, 7, 14, 3]	V_{32}	$[V_{13}, V_{32}]$	[4, 3]
S_3	(V_{15}, V_{34})	$[V_{15}, V_{12}, V_{24}, V_{34}]$	[8, 8, 4, 3]	V_{29}	$[V_{12}, V_{29}]$	[8, 3]
S_4	(V_{38}, V_{29})	$[V_{38}, V_{16}, V_3, V_{29}]$	[3, 5, 14, 3]	V_{19}	$[V_3, V_{19}, V_5]$	[14, 4, 11]
S_5	(V_{25}, V_{35})	$[V_{25}, V_4, V_{39}, V_{35}]$	[4, 16, 3, 4]	V_{17}	$[V_{17}, V_4, V_{21}]$	[7, 16, 3]
S_6	(V_6, V_{37})	$[V_6, V_1, V_3, V_{37}]$	[20, 7, 14, 3]	V_{14}	$[V_3, V_{14}, V_5]$	[14, 4, 11]
S_7	(V_2, V_{32})	$[V_2, V_3, V_1, V_{32}]$	[11, 14, 7, 3]	V_{11}	$[V_6, V_4, V_{11}]$	[20, 16, 3]
S_8	(V_5, V_{36})	$[V_5, V_4, V_{17}, V_{36}]$	[11, 16, 7, 3]	V_9	$[V_0, V_9, V_{25}]$	[5, 6, 4]
S_9	(V_{26}, V_{27})	$[V_{26}, V_6, V_{27}]$	[3, 20, 3]	V_1	$[V_{20}, V_1, V_{35}]$	[5, 7, 4]
S_{10}	(V_8, V_{39})	$[V_8, V_2, V_4, V_{39}]$	[5, 11, 16, 3]	V_8	$[V_6, V_8, V_{31}]$	[20, 5, 3]
S_{11}	(V_{13}, V_{31})	$[V_{13}, V_4, V_6, V_{31}]$	[4, 16, 20, 3]	V_{23}	$[V_{23}, V_7]$	[3, 9]
S_{12}	(V_{17}, V_{30})	$[V_{17}, V_{16}, V_3, V_{30}]$	[7, 5, 14, 3]	V_{15}	$[V_6, V_{15}, V_{26}]$	[20, 8, 3]
S_{13}	(V_{12}, V_{18})	$[V_{12}, V_0, V_3, V_{18}]$	[8, 5, 14, 3]	V_{39}	$[V_{39}, V_4, V_5]$	[3, 16, 11]
S_{14}	(V_{21}, V_{28})	$[V_{21}, V_4, V_{28}]$	[3, 16, 4]	V_{24}	$[V_{24}, V_{34}]$	[4, 3]
S_{15}	(V_{16}, V_{33})	$[V_{16}, V_3, V_{19}, V_{33}]$	[5, 14, 4, 4]	V_{36}	$[V_{36}, V_{28}]$	[3, 4]
S_{16}	(V_9, V_{17})	$[V_9, V_0, V_4, V_{17}]$	[6, 5, 16, 7]	V_{33}	$[V_{33}, V_{37}]$	[4, 3]
S_{17}	(V_4, V_{23})	$[V_4, V_2, V_7, V_{23}]$	[16, 11, 9, 3]	V_{30}	$[V_{30}, V_{15}]$	[3, 8]
S_{18}	(V_1, V_{24})	$[V_1, V_6, V_7, V_{24}]$	[14, 5, 8, 3]	V_4	$[V_3, V_{18}, V_4, V_5]$	[14, 3, 16, 11]
S_{18}	(V_1, V_{24})	$[V_1, V_6, V_7, V_{24}]$	[7, 20, 9, 4]	V_{22}	$[V_2, V_3, V_{22}]$	[11, 14, 3]
S_{19}	(V_{14}, V_{20})	$[V_{14}, V_3, V_1, V_{20}]$	[4, 14, 7, 5]	V_6	$[V_6, V_{16}, V_{38}]$	[20, 5, 3]
S_{20}	(V_0, V_{27})	$[V_0, V_3, V_{10}, V_{27}]$	[5, 14, 5, 3]	V_{10}	$[V_6, V_{10}, V_{27}]$	[20, 5, 3]

Step 5: Draw two Ogive curves separately for lower limit and upper limit of Confidence Intervals (CI).

Step 6: Using intersecting point of Ogive curve, find estimate of CI.

5.1 Numerical Illustration

Albert-Barabasi network [2] dataset (figure-1) has taken for representation of Covid-19 spread model which has N = 40 distinct nodes. In Table-3, degree sequences were obtained using CBP and SPP, which form overlapping clusters of nodes(infected person) of a network. The objective is to estimate average degree [10] of network and the relative efficiency of estimate using confidence interval size. For numerical evaluation, one can take samples at two stages. Sample of size k = 15 clusters at the first stage are taken from K = 20 clusters. In the second stage sample of vertices from each of these clusters are taken randomly. Further, one can use ogive simulation procedure as in Section 5 for P times.

Table 4: Normalized Degree Sequence of Aggregate Units using table-3.

Node	Degree	F_S	F_C	A_{ij}^S	A_{ij}^C	Node	Degree	F_S	F_C	A_{ij}^S	A_{ij}^C
V_0	5	4	1	2.44	6.75	V_{20}	5	1	1	9.75	6.75
V_1	7	5	1	2.73	9.45	V_{21}	3	1	1	5.85	4.05
V_2	11	3	1	7.15	14.85	V_{22}	3	1	1	5.85	4.05
V_3	14	10	4	2.48	4.73	V_{23}	3	1	1	5.85	4.05
V_4	16	7	4	4.46	5.4	V_{24}	4	2	1	3.9	5.4
V_5	11	1	4	21.45	3.71	V_{25}	4	1	1	7.8	5.4
V_6	20	4	5	9.75	6.75	V_{26}	3	1	1	5.85	4.05
V_7	9	2	1	8.78	12.15	V_{27}	3	2	1	2.93	4.05
V_8	5	1	1	9.75	6.75	V_{28}	4	1	1	7.8	5.4
V_9	6	1	1	11.7	8.1	V_{29}	3	1	1	5.85	4.05
V_{10}	5	1	1	9.75	6.75	V_{30}	3	1	1	5.85	4.05
V_{11}	3	1	1	5.85	4.05	V_{31}	3	1	1	5.85	4.05
V_{12}	8	3	1	5.2	10.8	V_{32}	3	1	1	5.85	4.05
V_{13}	4	1	1	7.8	5.4	V_{33}	4	1	1	7.8	5.4
V_{14}	4	1	1	7.8	5.4	V_{34}	3	1	1	5.85	4.05
V_{15}	8	1	2	15.6	5.4	V_{35}	4	1	1	7.8	5.4
V_{16}	5	1	1	9.75	6.75	V_{36}	3	2	1	2.92	4.05
V_{17}	7	3	1	4.55	9.45	V_{37}	3	1	1	5.85	4.05
V_{18}	3	3	1	1.95	4.05	V_{38}	3	1	1	5.85	4.05
V_{19}	4	1	1	7.8	5.4	V_{39}	3	2	1	2.93	4.05

5.2 Shortest Path Based Procedure for Parameter Estimation

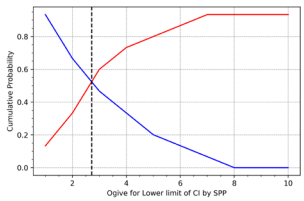

Figure 2: Ogive for lower limit of CI for SPP.

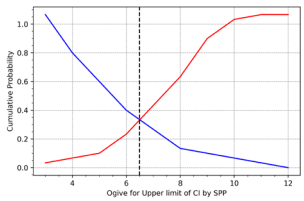

Figure 3: Ogive for upper limit of CI for SPP.

Followings are the output of analysis: using SPP:

- Estimated average degree = 5.05 (using eqn(11))
- Estimated variance for average degree = 0.024 (using eqn (15))
- Average confidence interval(CI) size = [7.62 - 2.48] = 5.14

Table 5: Cluster Sample Units(using table-3 and 4)

S. No.	Seed Pair of Node	Normalized Degree Sequence	Sample Shortest Path	Seed Node	Clique	Normalized Degree Sequence
S_1	(V_{11}, V_{22})	[5.85, 2.73, 5.85]	[V_{11}, V_1, V_{22}]	V_4	[V_3, V_{18}, V_4]	[4.73, 4.05, 5.4]
S_2	(V_{15}, V_{34})	[5.2, 3.9, 5.85]	[V_{12}, V_{24}, V_{34}]	V_{32}	[V_{13}, V_{32}]	[5.4, 4.05]
S_3	(V_{38}, V_{29})	[5.85, 9.75, 5.85]	[V_{38}, V_{16}, V_{29}]	V_{29}	[V_{12}, V_{29}]	[10.8, 4.05]
S_4	(V_{25}, V_{35})	[7.8, 4.46, 2.93]	[V_{25}, V_4, V_{39}]	V_{19}	[V_3, V_{19}]	[4.73, 5.4]
S_5	(V_2, V_{32})	[7.15, 2.48, 2.73]	[V_2, V_3, V_1]	V_{17}	[V_4, V_{21}]	[5.4, 4.05]
S_6	(V_3, V_{36})	[2.48, 2.44, 2.92]	[V_3, V_0, V_{12}]	V_{11}	[V_6, V_4]	[5.4, 5.4]
S_7	(V_8, V_{39})	[7.15, 4.46, 2.93]	[V_2, V_4 , V_{39}]	V_9	[V_0, V_9]	[6.75, 8.1]
S_8	(V_{17}, V_{30})	[4.55, 9.75, 2.48]	[V_{17}, V_{16}, V_3]	V_1	[V_{20}, V_{35}]	[6.75, 5.4]
S_9	(V_{12}, V_{18})	[5.2, 2.44, 2.48]	[V_{12}, V_0, V_3]	V_8	[V_6, V_8]	[5.4, 6.75]
S_{10}	(V_{21}, V_{28})	[5.85, 7.8]	[V_{21}, V_{28}]	V_{15}	[V_{15}, V_{26}]	[5.4, 4.05]
S_{11}	(V_{16}, V_{33})	[9.75, 2.48, 7.8]	[V_{16}, V_3, V_{19}]	V_{39}	[V_{39}, V_4]	[4.05, 5.4]
S_{12}	(V_9, V_{17})	[4.55, 2.44, 4.46]	[V_0, V_4, V_{17}]	V_{24}	[V_{24}, V_{34}]	[5.4, 4.05]
S_{13}	(V_4, V_{23})	[4.46, 7.15, 5.85]	[V_4, V_2, V_{23}]	V_{30}	[V_{30}, V_{15}]	[4.05, 5.4]
S_{14}	(V_{14}, V_{20})	[2.48, 2.73, 9.75]	[V_3, V_1, V_{20}]	V_{22}	[V_{22}, V_3]	[4.05,4.73]
S_{15}	(V_0, V_{27})	[2.48, 9.75, 2.93]	[V_3, V_{10}, V_{27}]	V_6	[V_6, V_{16}]	[5.4, 6.75]

Table 6: Sample Based Estimation using SPP(using table 5).

S. No.	Seed Pair of Node	Normalized Degree Sequence	Cluster mean	s_{ia}^2	$s_{ba'}^2$	95% C.I.	C.I. length
S_1	(V_{11}, V_{22})	[5.85, 2.73, 5.85]	4.81	3.24	0.058	[2.8, 6.85]	4.05
S_2	(V_{15}, V_{34})	[5.2, 3.9, 5.85]	4.98	0.99	0.005	[3.86, 6.11]	2.25
S_3	(V_{38}, V_{29})	[5.85, 9.75, 5.85]	7.15	5.07	4.41	[4.6, 9.7]	5.1
S_4	(V_{25}, V_{35})	[7.8, 4.46, 2.93]	5.06	6.2	0.0001	[2.25, 7.9]	5.65
S_5	(V_2, V_{32})	[7.15, 2.48, 2.73]	4.12	6.9	0.86	[1.15, 7.1]	5.95
S_6	(V_3, V_{36})	[2.48, 2.44, 2.92]	2.61	0.07	5.95	[2.31, 2.91]	0.6
S_7	(V_8, V_{39})	[7.15, 4.46, 2.93]	4.85	4.56	0.04	[2.43, 7.26]	4.83
S_8	(V_{17}, V_{30})	[4.55, 9.75, 2.48]	5.59	14.03	0.29	[1.35, 9.83]	8.48
S_9	(V_{12}, V_{18})	[5.2, 2.44, 2.48]	3.37	2.5	2.82	[1.58, 5.16]	3.58
S_{10}	(V_{21}, V_{28})	[5.85, 7.8]	6.83	1.9	3.17	[4.91, 8.74]	3.83
S_{11}	(V_{16}, V_{33})	[9.75, 2.48, 7.8]	6.68	14.16	2.66	[2.42, 10.93]	8.51
S_{12}	(V_9, V_{17})	[4.55, 2.44, 4.46]	3.82	1.42	1.51	[2.47, 5.17]	2.7
S_{13}	(V_4, V_{23})	[4.46, 7.15, 5.85]	5.82	1.81	0.59	[4.3, 7.34]	3.04
S_{14}	(V_{14}, V_{20})	[2.48, 2.73, 9.75]	4.99	17.03	0.0036	[0.32, 9.66]	9.34
S_{15}	(V_0, V_{27})	[2.48, 9.75, 2.93]	5.05	16.59	0	[0.44, 9.66]	9.22
	Avg Value		5.05		1.59	[2.48, 7.62]	5.14

The 95% confidence interval estimate using SPP is [2.48, 7.62](see table-6). Through ogive based simulation(figure 2 & 3) the confidence interval(CI) is [2.54, 6.01] .

5.3 Clique Based Procedure (CBP) for Parameter Estimation

Table 7: Sample Based Estimation using CBP(using table5).

S. No.	Seed Node	Normalized Degree Sequence	Cluster mean	s^2_{ia}	$s^2_{ba'}$	95% C.I.	CI length
S_1	V_4	[4.73, 4.05, 5.4]	4.73	0.46	0.45	[3.96, 5.49]	1.53
S_2	V_{32}	[5.4, 4.05]	4.73	0.91	0.45	[3.4, 6.05]	2.65
S_3	V_{29}	[10.8, 4.05]	7.43	22.78	4.12	[0.81, 14.04]	13.23
S_4	V_{19}	[4.73, 5.4]	5.07	0.22	0.11	[4.41, 5.72]	1.31
S_5	V_{17}	[5.4, 4.05]	4.73	0.91	0.45	[3.4, 6.05]	2.65
S_6	V_{11}	[5.4, 5.4]	5.4	0	0	[5.4, 5.4]	0
S_7	V_9	[6.75, 8.1]	7.43	0.91	4.12	[6.1, 8.75]	2.6
S_8	V_1	[6.75, 5.4]	6.08	0.91	0.46	[4.75, 7.4]	2.65
S_9	V_5	[5.4, 6.75]	6.08	0.91	0.46	[4.75, 7.4]	2.65
S_{10}	V_{15}	[5.4, 4.05]	4.73	0.91	0.45	[3.4, 6.05]	2.65
S_{11}	V_{39}	[4.05, 5.4]	4.73	0.91	0.45	[3.4, 6.05]	2.65
S_{12}	V_{24}	[5.4, 4.05]	4.73	0.91	0.45	[3.4, 6.05]	2.65
S_{13}	V_{30}	[4.05, 5.4]	4.73	0.91	0.45	[3.4, 6.05]	2.65
S_{14}	V_{22}	[4.05, 4.73]	4.39	0.23	1.02	[3.72, 5.06]	1.34
S_{15}	V_6	[5.4, 6.75]	6.07	0.91	0.45	[4.75, 7.4]	2.65
	Avg Value		5.4		1.59	[2.48, 7.62]	5.14

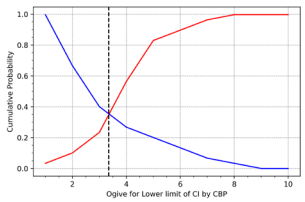

Figure 4: Ogive for lower limit of CI by CBP.

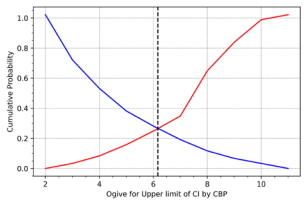

Figure 5: Ogive for upper limit of CI by CBP.

Following are the output of analysis using CBP:

- Estimated average degree = 5.4 (using eqn(11))

- Estimated variance for average degree = 0.028 (using eqn(15))
- Average confidence interval(CI) size = [6.86 - 3.94] = 2.92

The 95% confidence interval estimate using CBP is [3.94, 6.86].(See table 7) Through ogive based simulation(figure 4 & 5) the confidence interval(CI) is [3.27, 6.28].

6 Comparision

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

$$\text{PRG} = \frac{\text{length of CI}_{SPP} - \text{length of CI}_{CBP}}{\text{length of CI}_{SPP}} = \frac{5.14 - 2.92}{5.14} \times 100 = 43.19\% \tag{17}$$

Using ogive based simulation, the Percentage Relative Gain is:

$$\begin{aligned} \text{PRG}_{ogive} &= \frac{(\text{length of CI}_{SPP})_{ogive} - (\text{length of CI}_{CBP})_{ogive}}{(\text{length of CI}_{SPP})_{ogive}} \\ &= \frac{5.14 - 2.92}{5.14} \times 100 = 13.26\% \end{aligned} \tag{18}$$

7 Discussion

For a Covid-19 risk transmission network, using the Albert-Barabasi network model a network of infected people is created. Two procedures SPP and CBP are compared for a dataset and also used to evaluate average degree of network in a common setup. The Covid-19 transmission network can be represented as a graph of nodes and edges with weight. Overlapping clusters of nodes are formed by using two methods SPP and CBP. After comparison of percentage relative efficiency, the CBP is efficient by 43.2% over SPP (see sec6 eqn(17)). The simulated confidence interval for clique procedure(CPP) is [2.48-7.62] which includes the true value of average degree which is 5.55 of nodes, also supported by figures 3 and 4. The same calculation for simulated confidence interval using shortest path procedure(SPP) is [3.94-6.86] which is longer than earlier (See figure 2 & 3). Ogive based simulation procedure also supports better efficiency of CBP [(2.54, 6.01) for SPP, (3.27, 6.28) for CBP] (see figures 4 & 5). Compared to the length of CI as an efficiency measure, the CBP is 13.26% [see sec-6 eqn(18)] better than SPP. The ogive based simulation procedure seems a strong tool for confirmation of the findings. ■

8 Conclusion

The objective of this paper is to investigate average rate of risk of infection (average risk degree) of Covid-19 transmission network model by overlapping

cluster sampling based methods and evaluate efficiency by a comparative approach. The shortest path and cliques based procedure is used as procedural tools for generating clusters. Albert-Barabasi network model is used to create a small Covid-19 transmission network to provide numerical support to provide average degree which is an unknown parameter of Covid-19 network. To evaluate, the proposed procedure for sampling takes the cliques and compares them with shortest paths between several pairs of vertices in a setup of the overlapping cluster of degree sequence. The 95% confidence intervals are computed for both methods which contain the true value. The ogive based simulation procedure is also used for comparative statistical significance showing cluster based method using clique (CBP) provides a better estimate of the parameter (average degree) than cluster based method on shortest path (SPP). This contribution could be extended to test large network datasets and opens up new avenues and opportunities for various network parameter estimations. Higher the average degree of network tends to the high risk rate of transmission of Covid-19 infection in a community. ■

References

[1] R. Albert, A.L. Barabasi, Statistical mechanics of complex networks. Rev Mod Phys 74, 47–97, 2002.

[2] R. Albert, A.L. Barabasi, Emergence of scaling in random networks. Science 286, 509-512, 1999.

[3] W.G. Cochran, *Sampling Techniques*, Wiley, New York, (1977), 3^{rd}edn.

[4] E.W. Dijkstra, A note on two problems in connexion with graphs. Numerische Mathematik, 1(1), 269–271, 1959.

[5] E. Edwards, Family clusters. A common pattern for how the coronavirus spreads.; https://www.nbcnews.com/health/health-news/family-clusters-common-pattern-how-coronavirus-spreads-n1150646 (accessed June 17, 2020), 2020.

[6] O. Frank, Survey sampling in networks. The SAGE Handbook of Social Network Analysis, 389–403, 2011.

[7] O. Frank, Sampling and estimation in large social networks. Soc. Netw., 1(1), 91–101, 1979.

[8] N. Gupta, A. Singh, H. Cherifi, Centrality measures for networks with community structure. Physica A: Statistical Mechanics and its Applications, 452, 46-59, 2016.

[9] Gergely Palla, Uncovering the overlapping community structure of complex networks in nature and society, Nature, 435.7043, 814–818, 2005.

[10] Y. Moreno, R. Pastor-Satorras, A. Vespignani, Epidemic outbreaks in complex heterogeneous networks. Eur Phy J B, 26(4), 521–529, 2002.

[11] M.E.J. Newman, *Networks: An Introduction*. University Press. Oxford, 2010.

[12] M. Newman, Spread of epidemic disease on networks, Phys. Rev. E, 66(1),016128, 2002.

[13] S.S. Osahan, Overlapping clusters of tuberculosis contacts: An improved sampling strategy. Biometrical journal, 39(6), 689-697, 1997.

[14] C. Phucharoen, N. Sangkaew, and K. Stosic, The characteristics of COVID-19 transmission from case to high-risk contact, a statistical analysis from contact tracing data. EClinicalMedicine, 27, 100543, 2020.

[15] A. Rezvanian, M.R. Meybodi, Sampling social networks using shortest paths, Physica A: Statistical Mechanics and its Applications, 424, 254-268, 2015.

[16] D. Shukla, Y.S. Rajput, N.S. Thakur, Estimation of spanning tree mean-edge using node sampling, Model Assisted Statistics and Applications. 4(1), 23-37, 2009.

[17] D. Shukla, S. Pathak and N.S. Thakur, Estimation of population mean using two auxiliary sources in sample surveys, Statistics in Transition new series. 1(13), 21-36, 2012.

[18] M. Stehlík, J. Kiseľák, A. Dinamarca, E. Alvarado E, F. Plaza, F.A. Medina, ... &Y. Lu, Regional emergency-driven adaptive cluster sampling for effective COVID-19 management. Stochastic Analysis and Applications, 1-35, 2022.

[19] Y. Yang, P.S. Sun, X. Hu, Z.J. Li, Closed walks for community detection, Physica A: Statistical Mechanics and its Applications, 397, 129-143, 2014.

[20] J.G. Young, A. Kirkley and M.E.J. Newman. Clustering of heterogeneous populations of networks. Physical Review E, 105(1), 014312, 2022.