

In the Name of God

Journal of Information Systems & Telecommunication

Vol. 7, No. 4, October-December 2019, Serial Number 28

Research Institute for Information and Communication Technology
Iranian Association of Information and Communication Technology
Affiliated to: Academic Center for Education, Culture and Research (ACECR)

Manager-in-Charge: Habibollah Asghari, ACECR, Iran

Editor-in-Chief: Masoud Shafiee, Amir Kabir University of Technology, Iran

Editorial Board

Dr. Abdolali Abdipour, Professor, Amirkabir University of Technology, Iran
Dr. Mahmoud Naghibzadeh, Professor, Ferdowsi University, Iran
Dr. Zabih Ghasemlooy, Professor, Northumbria University, UK
Dr. Mahmoud Moghavvemi, Professor, University of Malaya (UM), Malaysia
Dr. Ali Akbar Jalali, Professor, Iran University of Science and Technology, Iran
Dr. Alireza Montazemi, Professor, McMaster University, Canada
Dr. Ramezan Ali Sadeghzadeh, Professor, Khajeh Nasireddin Toosi University of Technology, Iran
Dr. Hamid Reza Sadegh Mohammadi, Associate Professor, ACECR, Iran
Dr. Sha'ban Elahi, Associate Professor, Tarbiat Modares University, Iran
Dr. Shohreh Kasaei, Professor, Sharif University of Technology, Iran
Dr. Mehrnoush Shamsfard, Associate Professor, Shahid Beheshti University, Iran
Dr. Ali Mohammad-Djafari, Associate Professor, Le Centre National de la Recherche Scientifique (CNRS), France
Dr. Saeed Ghazi Maghrebi, Assistant Professor, ACECR, Iran
Dr. Rahim Saeidi, Assistant Professor, Aalto University, Finland

Executive Manager: Shirin Gilaki

Executive Assistants: Ali Mokhtarani, Mahdokht Ghahari

Print ISSN: 2322-1437

Online ISSN: 2345-2773

Publication License: 91/13216

Editorial Office Address: No.5, Saeedi Alley, Kalej Intersection., Enghelab Ave., Tehran, Iran,

P.O.Box: 13145-799

Tel: (+9821) 88930150 Fax: (+9821) 88930157

E-mail: info@jist.ir , infojist@gmail.com

URL: www.jist.ir

Indexed by:

- | | |
|---|-------------------------|
| - SCOPUS | www.Scopus.com |
| - Index Copernicus International | www.indexcopernicus.com |
| - Islamic World Science Citation Center (ISC) | www.isc.gov.ir |
| - Directory of open Access Journals | www.Doaj.org |
| - Scientific Information Database (SID) | www.sid.ir |
| - Regional Information Center for Science and Technology (RICEST) | www.ricest.ac.ir |
| - Iranian Magazines Databases | www.magiran.com |

Publisher:

Regional Information Center for Science and Technology (RICEST)
Islamic World Science Citation Center (ISC)

This Journal is published under scientific support of
Advanced Information Systems (AIS) Research Group and
Digital & Signal Processing Research Group, ICTRC

Acknowledgement

JIST Editorial-Board would like to gratefully appreciate the following distinguished referees for spending their valuable time and expertise in reviewing the manuscripts and their constructive suggestions, which had a great impact on the enhancement of this issue of the JIST Journal.

(A-Z)

- Alizadeh Noughabi, Havva, Islamic Azad University of Gonabad, Iran
- Ahmadizad, Arman, University of Kurdistan, Kurdistan, Iran
- Alavi, Seyed Enayatallah, Shahid Chamran University, Ahvaz, Iran
- Azizi, Sadoon, University of Kurdistan, Iran
- Asgari Tabatabaee, Mohammad Javad, University Of Torbat Heydarieh, Razavi Khorasan, Iran
- Ashrafi Payaman, Nosratali, University of Science and Technology, Tehran, Iran
- Amintoosi, Haleh, Ferdowsi University of Mashhad, Iran
- Behshid Behkamal, Ferdowsi University of Mashhad, Iran
- Darmani, Yousef, K. N. Toosi University of Technology, Tehran, Iran
- Ebadati, Omid Mahdi, Kharazmi University, Tehran, Iran
- Ebrahimpour, Nader, Islamic Azad University of Mahabad, West Azerbaijan, Iran
- Feshari, Majid, Kharazmi University, Tehran, Iran
- Farbeh, Hamed, Amirkabir University, Tehran, Iran
- Ghazvini, Mahdieh, Shahid Bahonar University, Kerman, Iran
- Hasanpour, Hamid, Shahrood university of Technology, Iran
- Haghbin, Afroz, Islamic Azad University Science and Research Branch, Tehran, Iran
- Jamshidnejad, Amir, Academy of Gondishapur, Ahvaz, Iran
- Jazayeriy, Hamid, Babol Noshivani University of Technology, Mazandaran, Iran
- Mavaddati, Samira, University of Mazandaran, Iran
- Mirzaei, Abbas, Islamic Azad University, Ardabil, Iran
- Mohammadzadeh, Sajad, University of Birjand, South Khorasan, Iran
- Mohammadpour, Davoud, Zanzan University, Zanzan, Iran
- Mirroshandel, Seyed Abolghasem, University of Guilan, Rasht, Iran
- Masdari, Mohammad, Islamic Azad University Urmia, Iran
- Mahdieh, Omid, University of Zanzan, Zanzan, Iran
- Moradi, Gholamreza, Amirkabir University, Tehran, Iran
- Pirgazi, Jamshid, Zanzan University, Zanzan, Iran
- Rezvanian, Alireza, Amirkabir University, Tehran, Iran
- Rahmani, Amir Masoud, Islamic Azad University Science and Research Branch, Tehran, Iran
- Rasi, Habib, University of Technology, Shiraz, Iran
- Soleimani Gharehchopogh, Farhad, Islamic Azad University Urmia, Iran
- Shirvani Moghaddam, Shahriar, Shahid Rajaei Teacher Training University, Tehran, Iran
- Tanhaei, Mohammad, Ilam University, Ilam, Iran
- Yadollahzadeh tabari, meisam, Babol Islamic Azad University, Iran
- Yaghoobi, Kaebeh, ManavRachna International University, India

Table of Contents

• Energy Efficient Clustering Algorithm for Wireless Sensor Networks	238
Maryam Bavaghar, Amin Mohajer and Sara Taghavi Motlagh	
• A Study of Fraud Types, Challenges and Detection Approaches in Telecommunication	248
Kasra Babaei, ZhiYuan Chen and Tomas Maul	
• A Fast Machine Learning for 5G Beam Selection for Unmanned Aerial Vehicle Applications	262
Wassa Shafik, S Mojtaba Matinkhah and Mohammad Ghasemzadeh	
• Investigate Network Simulation Tools in designing and managing intelligent Systems	278
Fatemeh Fakhari	
• A New Capacity Theorem for the Gaussian Channel with Two-sided Input and Noise Dependent State Information	294
Nima S Anzabi Nezhad, Ghosheh Abed Hodtani	
• Embedding Virtual Machines in Cloud Computing based on Big Bang–Big Crunch Algorithm.....	305
Afshin Mahdavi, Ali Ghaffari	
• Reallocation of Virtual Machines to Cloud Data Centers to Reduce Service Level Agreement Violation and Energy Consumption Using the FMT Method	316
Hojjat Farrahi Farimani, Seyed Reza Kamel Tabbakh, Davoud Bahrepour and Reza Ghaemi	

Energy Efficient Clustering Algorithm for Wireless Sensor Networks

Maryam Bavaghar*

Information Technology Institute, Iran Telecommunication Research Center (ITRC)
maryam.bavaghar@gmail.com

Amin Mohajer

Communications Technology Institute, Iran Telecommunication Research Center (ITRC)
scholar.mohajer@gmail.com

Sara Taghavi Motlagh

Information Technology Institute, Iran Telecommunication Research Center (ITRC)
sarataghavimotlagh@gmail.com

Received: 04/Jan/2020

Revised: 24/Apr/2020

Accepted: 08/May/2020

Abstract

In Wireless Sensor Networks (WSNs), sensor nodes are usually deployed with limited energy reserves in remote environments for a long period of time with less or no human intervention. It makes energy efficiency as a challenging issue both for the design and deployment of sensor networks. This paper presents a novel approach named Energy Efficient Clustering Algorithm (EECA) for Wireless Sensor Networks which is based on two phases clustering model and provides maximum network coverage in an energy efficient way. In this framework, an effective resource-aware load balancing approach applied for autonomous methods of configuring the parameters in accordance with the signaling patterns in which approximately the same bit rate data is provided for each sensor. This resource-efficient clustering model can also form energy balanced clusters which results in increasing network life time and ensuring better network coverage. Simulation results prove that EECA is better than LEACH, LEA2C and EECS with respect to network lifetime and at the same time achieving more network coverage. In addition to obtained an optimal cluster size with minimum energy loss, the proposed approach also suggests new and better way for selecting cluster heads to reduce energy consumption of the distributed nodes resulting in increased operational reliability of sensor networks.

Keywords: Wireless Sensor Networks; Energy-Efficient Clustering; Cluster Head Selection; Network Coverage; Network Life Time.

1- Introduction

With recent advancement in micro electro mechanical system (MEMS) technologies, low cost and low power micro electro nodes have become popular. WSN consists of tiny sensor nodes forming an ad hoc distributed data sensing and propagation network which collects the information from the surrounding environment. These networks combine wireless communication (i.e. transceiver) and minimal on-board computational facilities (i.e. processor or microcontroller) with sensing and monitoring. All these components together in a single device constitute a so called 'sensor node' or simply a 'sensor'. Wireless sensor networks lifetime mainly depends on battery, which are small and generally irreplaceable. Among all the task performed by Wireless Sensor Networks radio communication consumes most of the energy. It is therefore important to design proper

clustering algorithm, so that the inter cluster and intra cluster communication cost is minimum.

These networks are widely used in both the military and civilian applications such as target tracking, surveillance, and security management [1], [2]. With their capabilities for monitoring and control the network can provide a fine global picture of the target area through the integration of the data collected from many sensors each providing a coarse local view.

The main objective of sensor network is to collect data from monitoring environment and finally send it to base station via multi hop communication [3]. Based on network structure data routing protocols are divided into three structures [4]. Among them Hierarchical (cluster based) routing protocols are most energy efficient and widely used [5]. For e.g. say LEACH [6], LEACH-C [7], LEA2C [8], [9], EECS [10] and so on. In case of clustering those nodes which are geographically closer nodes form cluster. Each cluster we have a cluster head

* Corresponding Author

which acts as local base station. [6], [7] have proposed the LEACH protocol and a centralized version of this protocol, called LEACH-C. These protocols are based on clustering. Among various approaches available for energy efficient wireless sensor networks, the clustering approach in which data are gathered by one representative sensor of each group. It allows good scalability for the sensor network consisting of hundreds and thousands of nodes. Another advantage of clustering approach is balanced energy consumption among the nodes and thus increased network lifetime. The chain based approach tries to save the energy by forming a chain from the source to the sink, and so only one node will be transmitting data to the base station in any given transmission time frame. Here data fusion occurs at every node in the sensor network which allows all relevant information to permeate across the network.

In case of LEACH the job of cluster head is to collect data from their surrounding node and pass it to the base station. It is dynamic algorithm because the job of cluster head is rotated among the nodes but, this rotation is based on some probability for becoming a cluster head in each round [4]. LEACH-C, LEACH-N are modified version of LEACH but all of them failed to prevent the nodes from early energy dissipation of energy and hence leading to early end of network lifetime. The authors of [11] propose a protocol called Tree based clustering (TBC) here nodes in a cluster form a tree with cluster head as root. It effectively reduces and balances the energy consumption among the nodes. Compared with last dead node, improvement of TBC over LEACH is by 70%. However, the tree structure becomes complicated with number of rounds passing a more and more number of nodes get dying. The author of [12] has proposed a zone based hierarchical framework (ZBHF). Key feature in this scheme is to minimize the energy consumption during the self-organizing clustering scheme for energy efficient WSNs. Though the result obtained in this scheme is better than that of LEACH and LEACH-C but their network topology constraint them from being applied in a large scale network.

However, in case of LEA2C which based on two phase clustering, the lifetime of network improved remarkably around 50 percent [8, 9]. In case of LEA2C the cluster head was selected on the basis of maximum energy node. Our study on increasing network lifetime has seeded the idea for EECA from LEA2C.

In this paper we've done initial regrouping of clusters using K-means concept over multi criteria; energy and distance. This regrouping provides uniform energy distribution in all clusters. The difference of our proposed protocol with the previous clustering protocol lies in using multi criteria. We are able to adaptively cluster the nodes not only based on their topological closeness but also based on their energy levels by using K-means concept. We are able to reduce the computation time. Simulation results show that our new protocol can extend the network lifetime by 144% over

EECS and 62% over LEA2C, when the cluster head retained takes maximum energy criteria. Also random dying of the sensor nodes ensures more network coverage. Energy-performance trade-off for sensors processor operations is undergoing intense research considering the challenges with the evolving technology of wireless sensor computing. However, to guarantee energy-efficient processor operation, layout and architecture, it is necessary to identify and integrate optimization techniques and parameters influencing energy-performance trade-off in various energy efficiency domain. Existing literature on energy optimization in sensors focuses primarily on individual sub-domains such as offloading methods.

Our paper is organized as follows: first we present an energy consumption model in the WSNs. We have then given our proposed protocol followed by its explanation, where we have also described our proposed method K-Means_Initial, which by using multi criterion for clustering is helping us in building an energy efficient clustering model. Finally, we show through a series of experiments, some validation of our new algorithm and we present the future prospects.

2- Energy Consumption Model

Theoretically, our study is focused on power utilization of the wireless sensors nodes. But segregating the energy drained by it from the total energy drain is quite a difficult task given the association of network, and operating structure for sensor operations along with several other external factors influencing the network performance. Apparently, these factors have made this study very challenging task.

- Categorizing subcomponents of wireless sensor domain and defining their behavior w.r.t. the power consumption.
- Detecting uneven energy drains observed for a wide range of operations in WSN and identifying their root causes is quite a difficult task given the substantial amount of operations performed concurrently by the device.
- The energy minimization techniques applied varies with the varying functional and operation domains of the device processor. Assembling these techniques and finding application of common approach over varying domains is quite a challenging task.

Apparently, most of the current schemes focus on the subcomponents as a discrete entity. Due to these challenges, modeling a perfect optimal-energy system for wireless sensor networks turns out to be a challenging task. Energy optimization has become a crucial factor in wireless sensors as evolution of battery technology has failed to keep pace with evolution of computational sensor network technology. Moreover, the limitations imposed on

battery size intended to keep the device lightweight has made energy consumption by various software and hardware components a critical factor. Numerous energy models have been proposed so far depicting various factors playing crucial roles in WSN energy consumption. Eventually, the power consumed by sensor has been broadly described by as:

$$P_{Sensor} = P_{Display} + P_{Processor} + P_{Network}$$

where, P_{Sensor} is the overall node's power consumption, which is the sum of power drain by processor, network, and display components individually. These major consumers have been further classified into their specific functional areas consuming the device energy. The gravity of power issue for sensors can be seen through the exponential growth of the device processing capabilities along with the unwanted energy drains which go undetected for most of the times.

The communication consumes maximum energy than other tasks, in emission as well as in reception. Fig 1 shows an antenna model and the energy consumption rules associated [1].

To transmit a k bits message over a distance of d meters, the transmitter consumes:

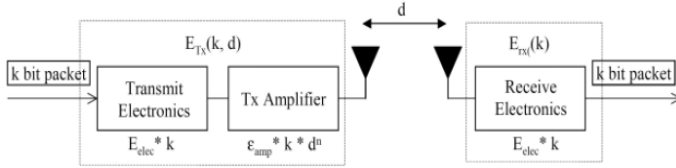


Fig. 1. Communication model in WSNs

$$E_{Tx}(k, d) = E_{Tx}(l) + E_{Tx_amp}(k, d) \tag{1}$$

And

$$E_{Tx}(k, d) = \begin{cases} k.E_{elec}(k, d) + k.\epsilon.d^2 & \text{if } :d < d_{crossover} \\ E_{Tx}(k, d) = k.E_{elec}(k, d) + k.\epsilon.d^4 & \end{cases} \tag{2}$$

The energy consumption for receiving k bits of data is computed as:

$$E_{Rx}(k) = k.E_{elec} \tag{3}$$

where:

- **E elec** : energy of electronic transmission/reception;
- **k** : size of a message;
- **d** : distance between the transmitter and the receiver;
- **E_{Tx}** : transmission energy,
- **E_{Tx amp}** : amplification energy;

- **ε** : amplification factor;
- **d crossover** : limit distance over which the transmission factors change of value.
- **E_{Rx}**: receiving energy

3- Proposed Protocol (EECA)

The EECA approach proposed here reduces the energy consumption of the network resulting in increased network life time. The residual energy of the node, distance, and the data overhead are taken into account for selection of cluster head in this proposed Energy Efficient Clustering Scheme (EECS). The waiting time of the mobile sink is estimated. The node having high residual energy and capable of transmitting a maximum number of packets is being chosen as CH. The transition from one state to other state is estimated using Markov model. The operation of the node and transition in Markov model are mainly based on the present state and not on the past history.

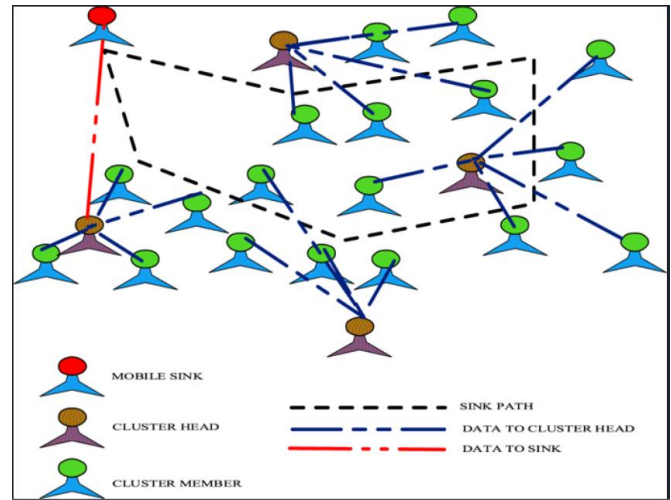


Fig. 2. The Architecture of Wireless Sensor Network

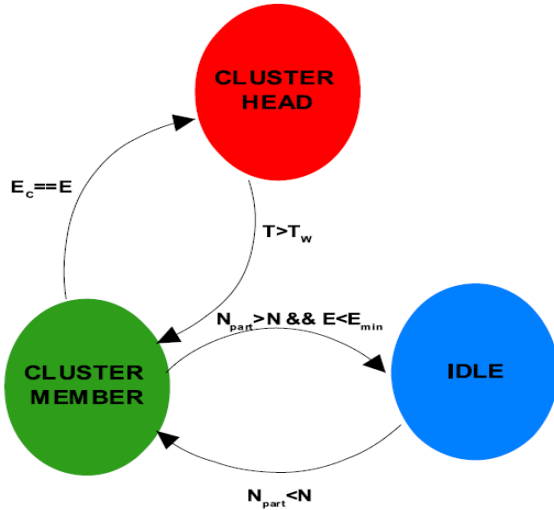


Fig. 3. Routing procedure in the energy efficient clustering algorithm

We first discuss here the system model starting from the assumptions, then after giving pseudo code for the proposed EECA algorithm, we have discussed each phase of the proposed routing scheme in detail:

3-1- Algorithm Assumptions

Our clustering algorithm is strongly related with LEA2C. The operations are divided into rounds as in LEA2C. Each round starts with a cluster setup phase, in which cluster organization takes place, which is followed by a data transmission phase. In data transmission phase data from ordinary nodes are transferred to the cluster head. Cluster head aggregate data and transmit it to the base station. In every cluster setup phase base station has to set appropriate role for each node in the cluster; we have here 0 for inactive normal node, 1 for active normal node and 2 for cluster head. We assume that there is no constraint about the energy for base station. Also base station has total knowledge about energy level and position of all the nodes of the network (most probably by the use of GPS receiver in each node). The sensor nodes are assumed to be homogenous i.e. they have same energy and communication as well as computation capabilities at algorithm start. List 1 below contains some of the term definitions used in our algorithm.

Definition:

1. aliveNodes: Number of nodes having energy more than threshold energy
2. element: Total number of sensor nodes initially
3. Olattice: It is an array of structure for cluster heads.
4. Centroid: Mean value of a cluster.
5. inputTemp: Array of structure of Input Nodes to hold normalized Input Nodes.
6. centroidTemp: Array of structure of Centroids to hold normalized value of Centroids.

7. m: Value for m is checked by Davies-Bouldin index, for m number of cluster heads.
8. I[N]: It is an array of structure of Input Nodes.
9. Number_of_centroids: Total number of centroids.
10. Number_of_ith_clusterElement: Total number of sensor nodes in i^{th} cluster.
11. Sumx: Variable to hold sum of x-coordinates of a cluster.
12. Sumy: Variable to hold sum of y-coordinates of a cluster.
13. x_of_newCentroid_of_ith_cluster: x coordinate value for newCentroid of ith cluster.
14. y_of_newCentroid_of_ith_cluster: y coordinate value for newCentroid of ith cluster.
15. x_of_oldCentroid_of_ith_cluster: y coordinate value for old Centroid of ith cluster.
16. newCentroid: Updated value of Centroid. oldCentroid: Old Centroid value.
17. I[N] – It represent array of structure of Input Nodes.
18. Role –Role is used to identify, normal node, cluster head node and inactive node.
19. I[i].c- Cluster Head Node or Normal Input Node.
I[i].c=1 for cluster head node. I[i].c=0 for normal input nodes.

EECA Algorithm 1:

Set of Input: {Number of nodes, Deployment area, Round} Output : {Alive nodes after each round}
Function Call: Algorithms 2,3 ,4, 5 and 6.

BEGIN:

1. Repeat steps 1.2 to 1.4 For round=1 to aliveNodes!=0 do
- 1.2 If round=1 then

CALL Algorithm 2; [Initialize node energy and distance coordinate]

CALL Algorithm 3; [Initial cluster formation]

CALL Algorithm 4; [Further optimization of the initial clusters]

Else

If aliveNodes! = element then:
Put the nodeID value for those nodes whose energy is below threshold as zero;

End If

CALL Algorithm 2; [Select m most energetic nodes as cluster heads.]

Save the cluster heads obtained above in Olattice;

CALL Algorithm 3; [Form initial cluster based on maximum energy.]

```

        CALL Algorithm 4; [Further
        optimization, with cluster head
        as maximum energy.]
    End If
    1.3 CALL Algorithm7; [Role to each node
    is provided.]
    1.4 CALL Algorithm 8; [Number of
    aliveNodes are updated. ]
    End For [Outer for Loop
    Ends.
    Signifying end of each round.]
END

```

Cluster Head Selection- Algorithm 2

Set of Input: {Array of structure of Input Nodes} Output: {m cluster heads}

BEGIN

```

If round== 1 then:
    Initialize x, y with random
    values; Initial energy to every
    node= 0.5J;
    Value for m is checked by Davies-Bouldin
    index, for m number of cluster heads;
    Select random cluster head values for m cluster
    heads;
    Normalize input set and cluster head set by using
    Min-Max Normalization;
    Copy Normalized Input Nodes into inputTemp;
    Copy Normalized Cluster Head into
    centroidTemp;
Else,
    Select m most energetic nodes as cluster heads.
Update centroidTemp with new set of normalized cluster
heads;
End If
END

```

K-means Initial (Initial cluster formation)- Algorithm 3:

```

Set of Input: {m cluster heads, Array of structure of
Input
Nodes, inputTemp, centroidTemp}
Output : {m energy balanced initial clusters}
BEGIN
    Repeat For i= 0 to i< element do
        If I[i]. nodeID!= 0 Then
            Find the Euclidian distance of this node of inputTemp
            with all nodes of centroidTemp taking both energy and x,
            y coordinates;

```

```

        Get the nearest node with this
        node;
        Form cluster with ith node belonging to its
        nearest node; If Ends
    For Ends
END

```

Cluster Optimization- Algorithm 4:

Set of Input: { m Initial energy balanced clusters}
Output : {Optimized n clusters}

BEGIN

```

Initialize old Centroid by Olattice;
Initialize newCentroid by zero;
CALL Algorithm 5; [Recompute
centroid] Return 0;

```

END

Recompute Centroid - Algorithm 5:

Set of Input: { Array of structure of Input Nodes, NewCentroid, Centroids}
Output : {Set of new Clusters}

BEGIN

```

Recompute centroid:
For i=0 to i< number_of_centroids do
    If Number_of_ith_clusterElement >1 then do
        For j=0 to j< number_of_ith cluster
        Element do Sumx =
            sumx+x_of_jth_element_of_ith
            Cluster;
            Sumy=sumy+y_of_jth_element_of_i
            th Cluster;
        End For
        x_of_newCentroid_of_ith_cluster=Su
        mx/ Number_of_ith_clusterElement;

```

```

        y_of_newCentroid_of_ith_cluster=Sumy
        / Number_of_ith_clusterElement;

```

```

Else
    x_of_newCentroid_of_ith_cluster
    =
    x_of_oldCentroid_of_ith_cluster;

```

```

        y_of_newCentroid_of_ith_cluster=
        y_of_oldCentroid_of_ith_cluster ;

```

End If

End For

```

If newCentroid is same as oldCentroid then:
    Return 0; [Returning with new cluster set]
Else
    CALL Algorithm 6; [Recompute Cluster]
End If
END

```

Recompute Cluster- Algorithm 6:

Set of Input: {Set of new centroid, Array of Structure of Input Nodes}
Output : {New cluster set of the Network}

```

BEGIN:
    Repeat For i=0 to i< element do
        If I[i]. nodeID! =0 then
            For j=0 to j< number_of Centroids
                do Compute euclidian distance
                between ith input node and jth
                centroid;
            End
            For
                Obtain the closest centroid to which
                this ith node belongs;
            End
            If
                End For
            CALL Algorithm 5; [Recompute centroid]
        End For
    End For
END

```

Role Allocation by base station -Algorithm 7:

Set of Input: {Optimized n clusters}
Output: { Each node with updated Role assigned by Base Station}

4- Cluster Setup and Installation

4-1- Cluster Setup Phase

The protocol uses a two phase clustering method. K-means initial followed by K-means algorithm. In K-means initial clustering we are doing initial regrouping considering two different data, energy and coordinates. For this we have used min max normalization method [13] in which min_a and max_a are minimum and maximum values for any given attribute a, for e.g. in our case we have taken x coordinate, y coordinate and energy as attributes and have fitted each of these attributes in the range of (0,1) with the help of min max normalization. Min max normalization maps a value v in the range of (0,1) by simply computing:

$$V' = (v - \min_v) / (\max_v - \min_v) \quad (4)$$

BEGIN

```

Repeat For i=0 to i< element
    do
        If ith node energy>threshold energy AND
        I[i].c==1 then:// For centroid
            ROLE
            =2; End If
        If ith node energy>threshold energy AND
        I[i].c==0 then://For cluster member
            ROLE
            =1;
            End If
        If ith node energy<threshold energy,
            then: ROLE=0;
        End If
    End For
END

```

AliveNodes Updating-Algorithm 8

Set of Input: {n clusters with updated Role} Output : {Update Number of aliveNodes}

BEGIN

```

Update data for each node with ROLE=1;
Aggregate data at the node with ROLE= 2;
Send data to base station; Update
energy of each node;
If node energy < threshold energy, then:
    aliveNodes--; End
If
    Return aliveNodes;

```

Where n is the number of clusters, c_i is the centroid of cluster i , δ_i is the average distance of all elements in cluster to centroid c_i , and $d(c_i, c_j)$ is the distance between centroids c_i and c_j . Since algorithms that produce clusters with low intra cluster distances and high inter cluster distances will have a low DB index, the clustering algorithm that produces a collection of clusters with the smallest DB index is considered the best algorithm based on this criterion

After every transmission phase we update energy of the network system by using equations 1, 2 and 3. Unlike previous algorithms in our proposed algorithm we reform cluster of the network and form new cluster heads with normalized data are then passed through K-means_Initial method for forming initial clusters. Value for k is checked by Davies-Bouldin index. Davies-Bouldin (DB) index actually compute the ratio of intra-clusters dispersion to inter-cluster distances by:

$$I_{db} = 1/n \sum_{i=1}^n \max((\delta_i + \delta_j)/d(C_i, C_{ij})) \quad (5)$$

4-2- Cluster Head Selection Phase

The cluster head is very important in cluster based WSN. They are responsible for data aggregation of each cluster member nodes and sending the same to the desired location. Cluster head is selected on different criterion. These are mainly based on maximum energy level or the nearest sensor to the base station. We have chosen the node as cluster head with maximum energy.

4-3- Data Transmission Phase

When clusters have been formed and cluster head has been selected, now it's time to send data packets sensed at normal nodes to their related cluster heads and after applying data aggregation function at cluster head, the same will be sent to base station. And after each round energy consumption is computed.

The criterion to be minimized in K-means is defined as:

$$\sum_{k=1}^k \sum_{x \in Q_k} \| (X - C_k) \|^2 \quad (6)$$

Where Q_k is K^{th} cluster, C_k centroid of cluster Q_k . The K-means_Initial algorithm introduced here produces initial clusters with energy and distance both criterion. The cluster obtained here is further optimized by K-means algorithm. In eq. 6, Q_k is K th cluster and C_k is the centroid of cluster Q_k .

Updated roles for each node. In the next section we have discussed the simulation and results of our proposed algorithm.

5- Simulation Results

The proposed algorithm is implemented using C programming language. We have also implemented LEA2C and EECS algorithm in order to compare our simulation results.

TABLE I. Parameters of simulation

Sensor Deployment Area	100x100
Base Station Location	50x200
Number of Nodes	400
Data Packet Size	800 bits
Initial Energy	0.5 J
Stand by State Energy Los	0.00006 J
Energy per bit spent by transmitter Circuits	50 nJ/bit
Amplifier Energy	10 pJ/bit/m ²

Table 1 shows the input sets provided to the algorithm EECA. The data provided here is containing same value as taken for LEA2C, for better comparison of results.

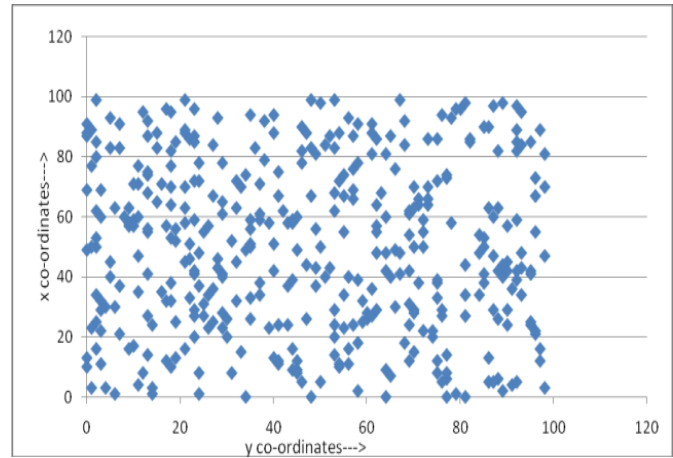


Fig 4. Initial random deployment of 400 nodes in 100x100 deployment area

Fig 4 shows initial random deployment of nodes in given area. We can see in the diagram that nodes are in the 100x100 area. Previous results were available under this area, for comparison purpose we have kept the area and number of nodes same.

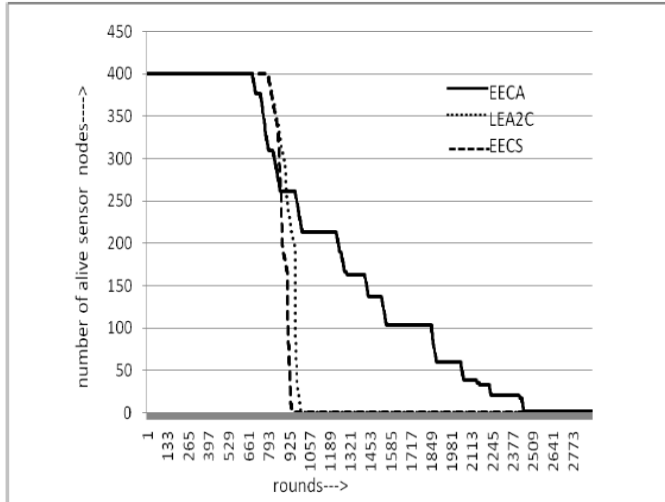


Fig. 5. Alive nodes VS rounds in EECA, LEA2C and EECS

We can observe in figure 5 the lifetime of the network in case of EECS, LEA2C and EECA. Initially all the algorithms are provided with 400 nodes as input. All these three algorithms are compared under same unit of cost for better comparison. We can see clearly that when cost function for EECS is applied, life span of network is up to

950 rounds and first node is dying near 780 rounds. In case of LEA2C, because of maximum energy criterion for cluster head selection, first node death time increases remarkably compared to other conditions. The result of LEA2C can be observed clearly which is giving better result over EECS in terms of both first node death time, as well as total network life time. In this case first node is dying at around 800 rounds. Total life time of the network is about 1010 rounds.

In case of EECA the first node dies at 700 rounds and after that it decreases with rounds. Lifespan of the network is up to 2893 rounds as can be seen from figure 5. We can see that network lifetime in case of our proposed algorithm EECA has a remarkable improve over the other two algorithms LEA2C and EECS. In figure 5 we can see that node is first dying nearly close with LEA2C and EECS but the network lifespan taking maximum energy as criterion is remarkably noticeable.

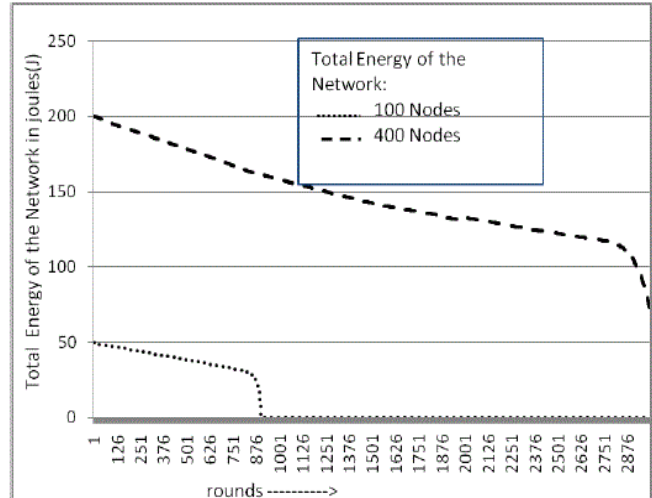


Fig 6. Total Energy of the Network available during each round in EECA

In figure 6 we can see total energy of the network has been shown during each round. Total energy available is represented here with bold blue curve for 400 sensor nodes and the one in bold red is for only 100 sensor nodes. We can see from the figure 6 that energy loss during each round is very less. As each time in our algorithm we are taking the node as cluster head which has maximum energy, thus it prevents early exhausting of energy of any particular node. Initially for 400 nodes total energy of the network available is 200 J and 50 J for the curve of 100 nodes in figure 6, as 0.5 J of energy is the initial assumed energy. It can be seen clearly that it is only during the near to last round that there is a severe fall in network energy level.

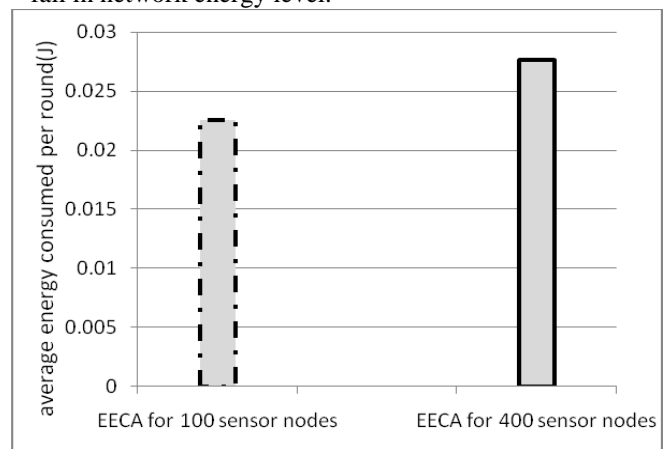


Fig. 7. Average energy consumed per round

In figure 7 average energy consumption of the network in each round has been shown for network containing 400 nodes and 100 nodes. In the column bar graph shown in figure 7 one in red color is representing result for 400 nodes and the other bar in blue is for network having 100

nodes only. Energy consumption is represented in terms of joules. We can observe clearly that average energy consumed during each round is very low because of energy efficient clustering. Unlike previous algorithms like LEACH, EECS, etc. We are having two phase clustering. The K-means_Initial algorithm proposed in our algorithm creates an initial energy efficient cluster which is then further optimized by K-means algorithm. Thus the average energy consumed is very low, which certainly impact in life enhancement of the network lifetime. We can also observe from figure 7 that though the number of nodes is increased by 4 times in case of network having 400 nodes as compared to network having only 100 nodes, the average energy consumed per round is increased by 0.005 only, which is certainly very small change.

In comparison with the proposed Cluster-based Dynamic Routing Approach (CDRA), we have compared the performance of the proposed scheme with [14] which applied a long short term memory (LSTM) to detect the channel characteristics automatically.

In Figure 8, the performance of the proposed CDRA approach and LSTM have been compared from the perspective of total consumption energy for different number of connected subscribers. As it is obvious in this figure, the total consumed energy of LSTM is completely close to CDRA.

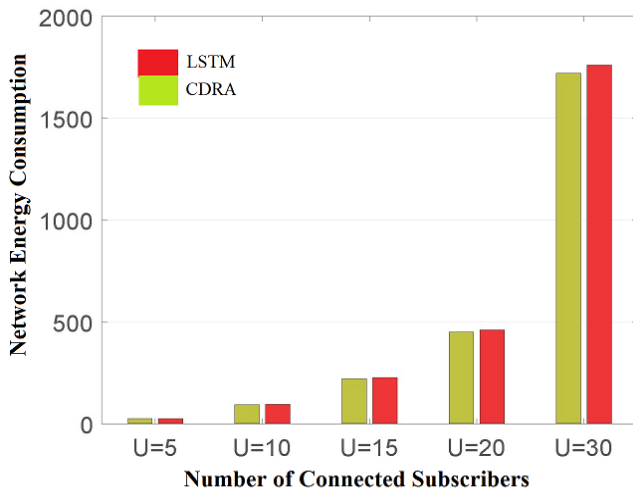


Fig. 8. Network energy consumption (mj) versus the Number of connected users (#)

Figure 9 demonstrates the effect of Number of resource block on the energy consumption. This result proves that we can decrease the consumed energy by increasing the number of resource blocks. It has been also exhibited that CDRA has better performance than other two cooperative distributed resource allocation approaches.

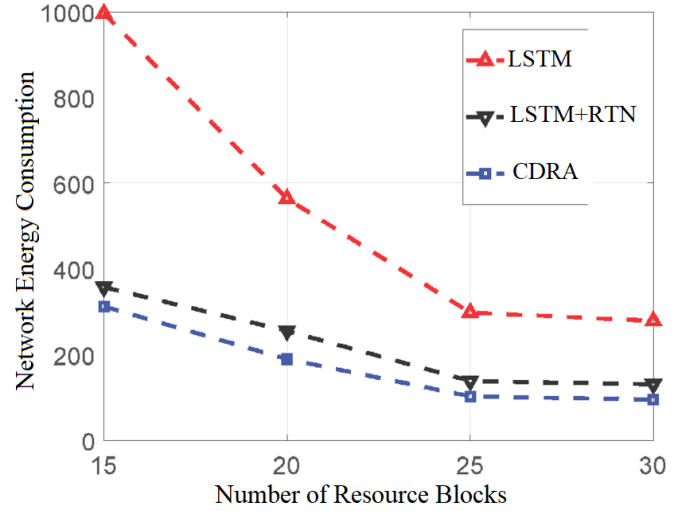


Fig. 9. Total energy consumption (mj) versus the Number of resource block (#)

The CDRA frameworks are able to dynamically choose the transmit power of all nodes according to their current channel conditions in every TS. Compared with other alternative approaches, our framework is able to provide better EE under different transmit power limitations, and are applicable in various moving speed conditions by adjusting the parameters of networks, which proves the effectiveness of our proposed frameworks as it is exhibited in the achieved results.

6- Conclusions

In this paper we tried to improve the total network lifetime using our proposed approach, Energy Efficient Clustering Algorithm (EECA), for wireless sensor networks. EECA ensures a positive benefit compared with EECS and LEA2C. The results obtained are very promising as compared with respect to network life time. We have obtained an optimal cluster size where the nodes send their data with minimum energy loss. Two phase clustering has been performed where initial clusters has been done using both criteria energy and distance. Initial energy balanced clusters thus obtained by K-means_Initial algorithm has been further optimized by K-means clustering algorithm. Our maximum energy as criterion for cluster head selection has established better result when simulate as compared to other popular clustering algorithm. As future work we can improve results by working on initial energy efficient clustering by using certain optimizing algorithms like topological self-organizing map, which will produce a complete energy-balanced cluster before passing it to second phase clustering. In our proposed algorithm we have done the initial energy efficient clustering by enhancing K-means

concept itself. Also the integration of other parameters in the clustering process, such as the moving speed of the sensors in case of mobiles sensors can also be considered for the future works.

References

- [1] El-mawla, Nesma Abd, Mahmoud Badawy, and Hesham Arafat. "SECURITY AND KEY MANAGEMENT CHALLENGES OVER WSN (ASurvey)." *International Journal of Computer Science & Engineering Survey (IJCSSES)* 10, no. 1 (2019): 15-34.
- [2] Mohajer, Amin, Maryam Bavaghar, Rashin Saboor, and Ali Payandeh. "Secure dominating set-based routing protocol in MANET: Using reputation." In *2013 10th International ISC Conference on Information Security and Cryptology (ISCISC)*, pp. 1-7. IEEE, 2013.
- [3] Singh, Omkar, and Vinay Rishiwal. "QoS Aware Multi-hop Multi-path Routing Approach in Wireless Sensor Networks." *International Journal of Sensors Wireless Communications and Control* 9, no. 1 (2019): 43-52.
- [4] Habib, Md Ahsan, and Mohammad S. Hasan. "A performance analysis of backbone structures for static sink based starfish routing in wsn." In *2017 4th International Conference on Networking, Systems and Security (NSysS)*, pp. 1-6. IEEE, 2017.
- [5] Ghosh, Indranil. "Study on hierarchical cluster-based energy-efficient routing in wireless sensor networks." *International Research Journal of Engineering and Technology (IRJET)* 5 (2018): 688-691.
- [6] Mohajer, Amin, Morteza Barari, and Houman Zarrabi. "An Efficient Resource Allocation Mechanism including Load-aware Handover Decision."
- [7] Nam, Deok. "Comparison Studies of Hierarchical Cluster-Based Routing Protocols in Wireless Sensor Networks." *Proceedings of 35th International Confer* 69 (2020): 334-344.
- [8] Kaur, Khushpreet, and Er Lovepreet Kaur. "Energy Efficient Protocol On Wireless Sensor Network: A Review." *International Journal of Scientific Research in Science, Engineering and Technology* 3, no. 3 (2017): 40-46.
- [9] Shukla, Anurag, and Sarsij Tripathi. "An Effective Relay Node Selection Technique for Energy Efficient WSN-Assisted IoT." *Wireless Personal Communications* (2020): 1-31.
- [10] Ramesh, Dr V. "Energy-Efficient Clustering Scheme (EECS) With Secure Data Aggregation for Mobile Wireless Sensor Networks." *International Journal of Electrical Electronics & Computer Science Engineering* 4, no. 5 (2017).
- [11] Sohan, Radhika, Nitin Mittal, Urvinder Singh, and Balwinder Singh Sohi. "An Optimal Tree-Based Routing Protocol Using Particle Swarm Optimization." In *Nature Inspired Computing*, pp. 117-124. Springer, Singapore, 2018.
- [12] Swathi, N., S. Santosh Kumar, KN Sunil Kumar, and P. Rajendra Prasad. "Zone based hierarchical energy efficient clustering scheme for WSN." In *2016 IEEE International Conference on Recent Trends in Electronics, Information &*

Communication Technology (RTEICT), pp. 136-140. IEEE, 2016.

- [13] Aksu, Gökhan, Cem Oktay Güzeller, and Mehmet Taha Eser. "The Effect of the Normalization Method Used in Different Sample Sizes on the Success of Artificial Neural Network Model." *International Journal of Assessment Tools in Education* 6, no. 2 (2019): 170-192.
- [14] Gui, Guan, Hongji Huang, Yiwei Song, and Hikmet Sari. "Deep learning for an effective nonorthogonal multiple access scheme." *IEEE Transactions on Vehicular Technology* 67, no. 9 (2018): 8440-8450.

Maryam Bavaghar received the B.S. degree in Information Technology Engineering from Birjand University, Birjand, Iran in 2009, and M.S. degree in Information Security from Malek Ashtar University of Technology, in 2013. Her research Interests is including Network Security and Robust Network Optimization in wireless environments, Intrusion Detection, Penetration Test, Wireless Sensor Networks and Cellular Resource Management.

Amin Mohajer received his PhD from Malek Ashtar University of Technology, Tehran, Iran. His research area includes intelligent resource management in wireless communication systems, application of robust optimization theory and recommender systems in self-organized wireless networks. His current works are more related to designing of a Self-Optimization Networking model in Next Generation Mobile Networks using data analysis and artificial intelligence approaches. His technical activity includes 8+ Years of experience in the field of RF Planning & Optimization; roll out project experience of operator, sound knowledge of LTE Advanced/5G RF Planning & Optimization in addition to more than 30 courses for GSM/GPRS/3G/LTE RF Planning & Optimization and academic training.

Sara Taghavi Motlagh received the B.S. degree in Information Technology Engineering from Islamic Azad University, in the field of wireless & mobile ADHOC network Performance in earthquake disasters. in 2009, she finished the M.S. in Information Technology (IT), Information Technology Management from Islamic Azad University, Science and Research branch of Tehran, in 2013. Her research Interests is including USSD Gateway, Cloud Computing, Mobile Wallet, SMS Gateway, GSM Networks Application Software.

A Study of Fraud Types, Challenges and Detection Approaches in Telecommunication

Kasra Babaei

Faculty of Science and Engineering, University of Nottingham Malaysia
Khyx6kbb@nottingham.edu.my

ZhiYuan Chen *

Faculty of Science and Engineering, University of Nottingham Malaysia
zhiyuan.chen@nottingham.edu.my

Tomas Maul

Faculty of Science and Engineering, University of Nottingham Malaysia
tomas.maul@nottingham.edu.my

Received: 13/Nov/2019

Revised: 02/May/2020

Accepted: 26/May/2020

Abstract

Fraudulent activities have been rising globally resulting companies losing billions of dollars that can cause severe financial damages. Various approaches have been proposed by researchers in different applications. Studying these approaches can help us obtain a better understanding of the problem. The aim of this paper is to investigate different aspects of fraud prevention and detection in telecommunication. This study presents a review of different fraud categories in telecommunication, the challenges that hinder the detection process, and some proposed solutions to overcome them. Also, the performance of some of the state-of-the-art approaches is reported followed by our guideline and recommendation in choosing the best metrics.

Keywords: Fraud Detection; Machine Learning, Telecommunication

1- Introduction

Telecommunications companies have long been suffering from fraudulent. In addition to the financial losses caused by fraudulent activities, companies that are incapable of foiling these activities will lose their customers as well. However, by developing adaptive and automatic systems it is possible to hinder fraud.

Telecommunication fraud is eventuating a tremendous financial loss for companies annually. Hardly possible to calculate and state the financial loss caused by fraudulent activities in the telecommunications industry, because some companies prefer not to reveal to protect their reputation. In addition, not all frauds are detected by telecommunication companies and the efficiency of their detection systems is not clear. Nonetheless, based on some analyses, it was concluded that telecommunication fraud caused 46.3 billion USD in 2013 globally, which was about 2 percent of the worldwide telecom revenues [1]. Also, according to [2], telecommunication companies lose around 7% of their revenue due to fraudulent activities. This financial loss can produce pernicious effects on companies' revenues [3]. It is worth to note that even though wireless communication has become more predominant, telecommunication companies are still suffering, particularly in developing countries such as China [4]. As listed in Table 1, companies all over the world lose a considerable amount of their revenue due to fraudulent activities [2].

In this paper, the aim is to provide a thorough overview of different fraud related systems, namely fraud detection systems and fraud prevention systems, followed by the techniques and challenges that cause problems to these systems. As depicted in Figure 5, the paper tries to keep the focus on research works that were published during the past decade but also covers some of the earlier works that we find relevant. Also, an extensive review of different evaluation metrics used for performance measurement is carried out to understand the most employed and appropriate metrics in telecommunication fraud.

Table 1 Revenue loss in 2015 [2]

Fraud type		Fraud loss in B. of dollars		
		Glob ally	Western Europe	North America
Fraud type	International Revenue Share Fraud (IRSF)	10,75	2,07	3,21
	Interconnect Bypass Fraud	5,97	1,15	1,78
	Premium Rate Service Frau	3,74	0,72	1,12
Fraud methods	Subscription Frau	8,05	2,4	1,55
	PBX Hacking IP PBX Hacking	7,47	2,22	1,44
	Wangiri Fraud	1,77	0,53	0,34
	Phishing	1,57	0,47	0,3
	Abuse of Service Terms and Condition	1,17	0,53	0,34
	SMS Faking or Spoofing	0,79	0,23	0,15

* Corresponding Author

2- Related Works

Fraud detection is very important for companies to thwart fraudsters from causing financial loss, reputational damage, and invading their customers' private information. Various surveys have reviewed electronic fraud (also known as e-fraud), which is any sort of illegal action committed by using electronic technology and equipment such as computers. Some of the main categories of fraud that have been covered include credit card fraud, money laundering, insurance fraud, financial statement fraud, and mortgage fraud [5]–[7].

Financial fraud includes a vast area and researchers have reviewed and categorised them differently, which is important to be studied. Recent research in financial fraud such as [6] investigated methods and approaches in various areas including telecommunications fraud, credit card fraud, and insurance fraud. In [8], the authors reviewed four other types of financial fraud, namely computer intrusion, money laundry, telecommunications fraud, and credit card fraud. Four different types of fraud in telecommunications were defined by [8] (i.e. superimposed or surfing fraud, subscription fraud, ghosting fraud, and insider fraud), and they also investigated some major issues and challenges as well as tools that were used to detect them.

Another review was conducted by [9] in which the authors looked into three fraud areas, namely credit card fraud, computer intrusion, and telecommunications fraud. In telecom fraud, fraud attacks were categorised into superimposed fraud and subscription fraud. Each category includes some subcategories as well, such as phone cloning and ghosting that are under the superimposed fraud category. The review emphasised on three major approaches to detect telecom fraud (i.e. rule-based, neural networks, and visualisation).

Another comprehensive survey was conducted by [10] in which data-mining methods used in fraud detection within a 10 year period (i.e. from 2000 to 2010) were reviewed. The review was more focused on the data-mining methods used, including semi-supervised and also one-class classification methods in which the model is trained with only one class, which is often the non-fraudulent class.

A recent and comprehensive review was done in [6] that covered five different areas of fraud (i.e. telecommunications, health insurance, credit card, and online auction) and investigated four major challenges along with the efforts made to overcome them in each area. Issues and challenges in fraud detection regardless of the area were likewise studied in [5].

Fraud systems are very important in providing secure and reliable services and eliminating financial losses incurred by fraudsters. Studying different approaches proposed in the literature can provide useful insights related to the problems and challenges in this area, which can lead to

identifying the gaps for further investigation. Also, such research works embody a pool of ideas that can be refined, extended, and combined, in order to further improve current performance levels in this area.

3- Fraud

This section presents a comprehensive definition of fraud in general and also in the specific context of telecommunication. It also reviews the different motivations that push people to commit fraudulent activities in this area.

There are numerous definitions of fraud. The Cambridge Advanced Learner's Dictionary defines fraud as "the crime of getting money by deceiving people", and the Merriam-Webster Dictionary defines it as "the crime of using dishonest methods to take something valuable from another person". In other words, any deliberate action with the purpose of making unfair or unlawful gain is known as fraud [1]. In telecommunications, fraud refers to the misuse of services provided by telecom companies, including voice or data, without gaining permission and without the intention of paying [11]–[13]. Fraud detection refers to the efforts made to spot and catch undesirable behaviours relating to this misuse [14]. These undesirable behaviours include delinquency, intrusion, and account defaulting [9].

It is important to understand the motivation behind fraudulent activities. One main motivation is to use services with no intent to pay for them (self-usage), where another is based on financial gain obtained from reselling premium services to customers for a lower price [15]. In [8], the motives for fraudulent activities were categorised into two groups based on revenue, namely revenue fraud and non-revenue fraud, where in the former the fraudster tries to earn money and in the latter the purpose is only to gain free services. Furthermore, fraudsters can make untraceable communications and hide their identity [16], which is very useful for criminals and terrorists who want to stay hidden to perpetrate their vicious plans. The opacity of communication is partly due to the complex topology and massive size of networks that make it extremely difficult, time consuming, and costly to identify and find the location of the fraudsters [17].

3-1- Fraud Types

There are several forms of telecommunication fraud and previous works have categorised them differently; however, almost all of them have categorised fraud in telecommunications based on the methods used by fraudsters to gain unauthorised access [14]. A very broad categorisation was made by [18] that divided fraud into subscription fraud and superimposed fraud. In subscription fraud, fraudsters possess an account whose services they

do not intend to pay for (high debt fraud is also under this category). The account is completely genuine, and fraud happens when it is active. Another fraud case in subscription fraud is registering with a false identity. It is worth noting that there are two types of users, namely domestic and commercial, where in the latter case, the cost is at a higher rate because the usage of commercial users is at a higher level [19]. A very common subscription fraud happens when a commercial user registers with a false identity as a domestic user to reduce the cost of communication. In [20] the authors divided subscription fraud into two subcategories based on intention: (a) for making profit, and (b) for personal usage. Detecting subscription fraud is, arguably, the most challenging kind of fraud in telecommunication, and this type can cause a huge revenue loss for companies [21]. Superimposed fraud happens when fraudsters take control of an account, which in fact belongs to a legitimate customer. Scrutinising calling records on the bill is a very common method for detecting superimposed fraud [22], [23]. Also, [9] used the same approach and classified fraud into superimposed fraud and subscription fraud. The authors further subcategorised superimposed fraud into other types such as phone cloning, ghosting, insider, and tumbling, while insolvent cases were considered a subcategory of subscription fraud. Another classification was proposed by [18] in which fraud types were categorised based on their source and nature, into internal fraud and external fraud. In external fraud, fraudsters' identities are hidden due to the nature of the source, which is from outside of the organisation, often with no geographical limitation. In contrast, the source of an internal attack is from within the organisation, which makes the investigation process easier. Some common examples of internal fraud are [18]:

- **Ghosting:** using technical means to get a cheap or free rate.
- **Sensitive Information Disclosure:** selling important and sensitive information to external entities.
- **Secret Commissions:** Secret profits (e.g. vouchers) are traded for obtaining goods or services.

On the other hand, common examples of external fraud are [18]:

- **Surfing:** obtaining another customer's service without their authorisation, for instance, by cloning SIM cards, or manipulation of Private Branch Exchange (PBX).
- **Premium Rate Fraud (PRS):** fraudsters inflate the revenue payable to a provider by sending traffic to a PRS line [24].
- **Roaming Fraud:** a fraudulent subscriber uses the long delay of transferring Call Detail Records (CDR) between the visiting network and the home network to refuse payment.

As illustrated in Figure 1 and explained in [15], fraud types can be divided into three main categories known as the 3M's classification, namely motive, means, and methods. The motive includes non-revenue fraud and revenue fraud (refer to Section 3 for detail), and the means are the nature or form of the fraud which satisfy the motive, where some examples are:

- **Call Selling:** selling high rate calls, often international calls, below the real price.
- **Sensitive Information Disclosure:** an internal fraud in which the fraudster sells important information such as access codes.
- **Content Selling:** obtaining content such as games and ringtones for free by exploiting the payment system.

Referring to generic methods to perpetrate fraud, the authors in [15] defined four main methods:

- **Subscription Fraud:** obtaining an account with true or false credentials with no intent to pay for it.
- **Technical Fraud:** exploiting vulnerabilities in the network for financial benefits.
- **Internal Fraud:** committing fraud from inside the organisation.
- **Point of Sale:** fabricating sale documents to increase the compensations which should be paid by the telecommunications company.

Another fraud classification was made by [25] where fraud was divided into four groups:

- **Contractual Fraud:** obtaining a service with no intention of paying. An example of this type is subscription fraud and premium rate fraud.
- **Hacking Fraud:** misusing system vulnerabilities to make revenue by exploiting or selling functionalities. Network attack and Private Automatic Branch Exchange (PABX) fraud are two examples of hacking fraud.
- **Technical Fraud:** exploiting the technical vulnerabilities of the network to perpetrate fraud. Detecting the vulnerabilities often requires technical knowledge, however, once discovered, non-technical fraudsters can also utilise it to their benefit. Cloning and technical internal fraud are examples of this type.
- **Procedural Fraud:** where a fraudster tries to attack the implemented procedure, normally business procedures, whose goal is to minimise exposure to fraud. Often, the purpose of procedural fraud is to grant access to the system. Examples of this type are roaming fraud and voucher ID duplication.

Another fraud classification was given in [14]. Telecommunication fraud was classified into two main groups based on transmission medium, namely traditional networks and Voice over Internet Protocol (VoIP). In traditional networks, fraud can be subcategorised into further types such as subscription fraud, SIM cloning, Premium Rate Service (PRS), dealer fraud, roaming fraud, calling card fraud, and internal fraud. In VoIP, fraud is committed by employing VoIP techniques, and some examples of this category are Arbitrage fraud, call transfer fraud, location route number, and bypass fraud. Other types of attacking methods in VoIP are Man in the middle Attack, Replay Attack, Teardown Attacks, Flooding Attacks and SPIT (Spam over IP Telephony) [26].

With the growth of smartphones and broadband Internet, users prefer to use VoIP to make their calls or send their messages to reduce cost. Consequently, new types of fraudulent activities have also emerged such as registration hijacking, spam, and message tampering [27]. Currently, smartphone advertising is used in many mobile applications, which has attracted fraudsters who use

computer bots to generate abundant click events on advertisements thus earning money from them [28]. Arguably, there is no single perfect classification framework for fraud, and scholars have come up with various categories such as in [29] and [30] often dividing fraud into similar groups in which superimposed fraud and subscription fraud are the two dominant types [11].

Table 2 shows a summary of fraud detection systems since 2011, based on the type of fraud that the reported systems were designed to deal with. As depicted in **Error! Reference source not found.** and

Table 2, subscription fraud, targeted by almost half of the research papers published since 2011, was the most studied type of fraud, followed by SIM box fraud. It is worth mentioning that there are only a scant number of papers targeting a specific type of telecommunication fraud while in some research works such as [31], [32] and [33] the authors have tried to detect any type of telecommunication fraud instead of identifying merely a specific type.

4- Fraud Systems

In this section, different fraud management systems are reviewed, followed by methods commonly used in each system. The increase of fraudulent activities in the telecommunications industry and the financial losses incurred by this lucrative crime have compelled companies to look for automatic and intelligent systems that can foil fraud. These systems generally fall into two main categories which are prevention systems and detection systems. The following subsections will explain the differences between the aforementioned systems.

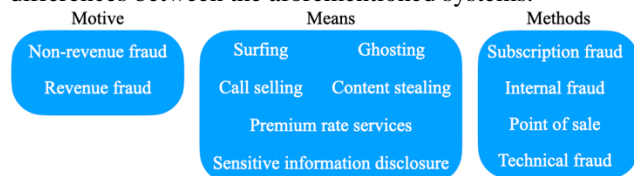


Figure 1 3M's fraud classification proposed by [15]

4-1- Fraud Prevention Systems

The idea behind fraud prevention systems (FPSs) is to block or prevent any fraudulent activity from occurring [8], [34]. Fraud prevention systems are the first barrier in controlling and confronting fraudulent activities. There are various mechanisms for this purpose such as using a firewall, encryption, or other forms of procedures such as Personal Identification Number (PIN) or Subscriber Identity Module (SIM) used in Private Branch Exchange (PBX) [34], and analysing applications and identifying potential customers before providing any service [20]. The problem with these kinds of systems is that they are not

infallible, their performance is usually questionable, and perpetrators can usually adapt and change their methods to overcome the prevention mechanisms [18]. Besides the low effectiveness of these systems, FPSs are usually intrusive from the users' perspective [8]. For example, assigning a security code is a typical approach for protecting SIM card users [16], however, users often forget the code as it is rarely used, and repeatedly re-entering incorrect codes can result in SIM lock.

4-2- Fraud Detection Systems

Fraud detection systems (FDSs) are the next defensive system where it is assumed fraudsters have managed to bypass the FPS, or in other words, fraud has already occurred. An optimum detection system should be capable of identifying and reporting fraud activities at the time of their occurrence (also known as real-time detection). Fraud detection systems can help system managers to overcome the limitations of prevention systems by continuous monitoring. As depicted in Figure 2 and according to [33], the mechanisms used by FDSs can be divided into three main categories that are rule-based systems, visualisation systems, and user-profiling systems. Authors in [35] categorised detection systems into statistical and probabilistic, or machine learning and rule-based.

It is also worth mentioning that a data mining process can be categorised into offline and online modes [36]. In the offline mode, relevant data has been already been collected and stored, and models are trained to be used later for predicting the outcome of unseen data. In fraud detection, data usually comes in the form of streams that require an online mode of data mining [37]. The focus of this survey is on online fraud detection systems. This is due to the potential and flexibility of these systems (i.e. they can automatically adapt to new types of fraud). Besides, with an online system it is possible to take actions such as terminating the call while the call is still in progress, but with an offline system, detection is generally not possible until the user terminates the call [38]. These characteristics have made online systems more attractive to both researchers and the industry. The following subsections provide an overview of the various detection techniques used in FDSs. The techniques are categorised into the following methods: 1) rule-based systems; 2) visualisation systems; and 3) user-profiling.

In rule-based systems, a set of rules are defined by a field expert, and an alarm is triggered when a certain criterion is met. Although these systems are straightforward, effective and efficient, they come with some deficiencies which are [39], [40]:

- Vulnerability to unknown fraud attacks.
- Rules should be programmed precisely for every possible fraud.
- A field expert and prior knowledge is needed for setting new rules.

- Setting new rules is not immune to human error.
- Defining new rules is time consuming and often complicated.

As mentioned above, a big disadvantage of rule-based systems is that adversaries can adapt and change their attacking methods to avoid triggering an alarm, which makes rule-based approaches ineffective against new attacking patterns. In [41], a rule-based expert system for detecting superimposed fraud was proposed to evaluate a user's account upon the user's request.

Another technique used by FDSs is visualisation in which human visual pattern recognition is required to identify any sudden changes in the patterns of subscribers' activities such as a location change or a dramatic increase in usage [42], [43]. After an initial anomaly detection, further investigation of the visualised data is still required to detect fraud. A disadvantage of this technique is that it is not a fully automated technique and relies on a human field expert to scrutinise and pick out suspicious cases for further investigation.

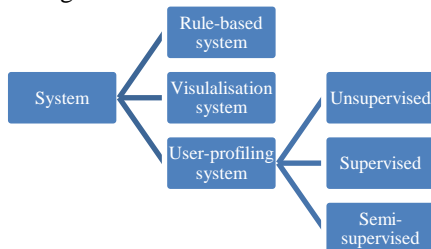


Figure 2 Classification of methods used in fraud detection systems [33]

The basic idea behind user-profiling consists of accumulating user characteristics to build a profile (also known as 'user dictionary') that represents the user's behaviour [34]. This profile that shows the user's behaviour in the past is then used to compare with the recent activities to determine significant changes, which are often signs of fraud. The data used by the system to describe a user's behaviour is usually derived from call detail records (CDRs) that contain information such as call duration, call location, time of the call, the destination and call cost [18]. Recently, user-profiling has attracted a lot of attention because of its effectiveness in automatic fraud detection and learning new fraud patterns [33]. User-profiles are constructed using data mining methods. Previous works have used various statistical methods based on the availability of labelled data, which can be mainly divided into supervised and unsupervised learning approaches [40]:

In supervised learning, a portion of the dataset, which is labelled as "fraudulent" or "non-fraudulent", is used as a training set. There are two main types of supervised learning models: classification and regression. In a classification model, the outcomes are discrete, and the model tries to map unseen instances into defined classes. Alternatively, when the outcomes are continuous, regression models are used, where the aim is to predict

continuous values. Supervised learning requires a labelled training set which is known as a limitation, given the cost (temporal and financial) of generating labels. Without a labelled dataset, it is not possible to train the model. Moreover, and as a result of the above mentioned cost of labelling, training sets are often not large enough to effectively train the model [44]. Common supervised learning methods consist of support vector machines (SVMs), artificial neural networks (ANNs), decision trees, naïve bayes, and *k*-nearest neighbours (KNNs). The authors of [45] tried to accumulate characteristics of a user based on weekly activities and then detected fraudulent accounts using feed-forward neural networks (FF-NNs).

Table 2 Fraud Type & Reference & Description

Fraud type	Refence	Description
Superimposed fraud	[22], [23], [46]	Taking control of a legitimate account and making unauthorised calls
Subscription fraud	[1], [19], [30], [47]	Obtain a genuine account with no intention to pay for its services
Toll fraud	[39]	Make costly long-distance calls without authorisation that will be paid by subscribers
SIM box fraud	[11], [14], [35]	Channel national and international calls away from mobile operators and deliver them as local calls

Unlike supervised learning, unsupervised learning does not require labelled training data. This represents a significant cost saving, which in turn avoids the insufficient training data problem mentioned in the previous point [33]. It is usually a better approach when the majority of the dataset is negative for fraud; however, it can also produce a high false alarm rate if this assumption is not met [48]. Some common unsupervised learning algorithms consist of hierarchical clustering, self-organising maps (SOMs), and gaussian mixture models (GMMs). Generally, supervised learning algorithms can achieve higher detection rates and lower false positive rates while in unsupervised learning it is possible to detect unseen attacks [49]. The authors in [50] tried two different clustering methods to detect fraud based on weekly accumulated characteristics of users. Another unsupervised approach was conducted by [51] in which they used expectation maximisation for tuning a hierarchical regime-switch model for call-based detection.

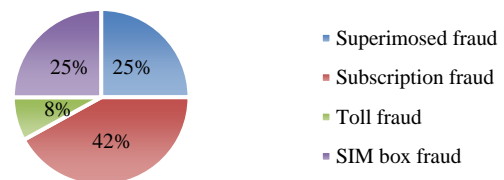


Figure 3 Types of Telecommunication Fraud Investigated Since 2011

There is also another method known as semi-supervised learning that lies between supervised learning and

unsupervised learning. Semi-supervised learning is used under various circumstances such as paucity of training data or lack of certainty about all instances' labels [30]. This method uses both labelled data and unlabelled data for the learning process [52], which makes the method very suitable for fraud detection where the number of positive instances in the dataset is very small [53]. Table 3 presents a summary of various approaches and techniques used for telecommunication fraud detection in the literature since 2011. We believe these recent works that we studied are more influential and can represent various learning methods.

As Table 3 illustrates, recently, there has been an upward trend in unsupervised learning methods in this area. A recent unsupervised learning approach was reported in [39] to overcome the limitations of rule-based systems in which Local Outlier Factor (LOF) was utilised on real call data to detect toll fraud attacks, and prevent VoIP fraud. Another user-profiling approach using unsupervised methods was used in [33] that tried to use Latent Dirichlet Allocation (LDA) and a straightforward threshold-type classifier with automatic threshold setting. They also used three different approximation methods to calculate the Kullback Leibler divergence (KL-divergence) between two layers of LDA, ultimately finding the most effective

method. In an earlier work, they introduced four different approximation methods to compute the KL-divergence between two LDAs, and unlike other similar work their approach aimed to detect the whole fraudulent accounts instead of merely one single fraudulent call [54].

Another extensive research was conducted by [30] in which a semi-supervised approach was applied to the dataset to detect subscription fraud in telecommunications. The authors proposed a framework consisting of 3 phases: preprocessing, clustering, and classification. After data cleaning, transformation, and dimensionality reduction in the preprocessing phase, SOM and *k*-means techniques were employed for clustering the data. In the classification phase, three different classifiers including Decision Trees (DTs), SVMs, and Neural Networks (NNs) were utilised to label the accounts into fraudulent and non-fraudulent. An ensemble of the three aforementioned classifiers was also built and compared to the original classifier results, with the ensemble showing superior performance.

Recently, deep learning techniques have also been employed. After their success in other fields such as image processing, authors in [55] proposed an approach based on a deep learning architecture. In particular they employed Deep Convolutional Neural Networks (DCNN), to separate normal behaviours from fraudulent ones.

Table 3 Telecommunication Detection Techniques Used Since 2011

Strategy	Learning Method	Reference	Detailed Description
Knowledge-based	Rule-based	[47], [58]–[60]	Triggering an alarm based on pre-defined rules
Supervised	SVM & ANN	[35]	Comparing the performances of SVMs and ANNs
	ANN	[11]	Applied a supervised learning method using multilayer perceptron (MLP)
	One-Class SVM	[22]	Applying Quarter-Sphere SVM which is a formulation of One-Class SVM
	Naïve-Bayesian	[19]	Used Naïve-Bayesian classification to calculate the probability and KL-divergence to detect subscription fraud
	Fuzzy logic	[14]	Used the Min and Max values for 5 predefined patterns to design the fuzzy logic membership function
Unsupervised	Local Outlier Factor (LOF)	[39], [61]	An outlier detection approach based on local density
	Self-Organising Map (SOM)	[32]	A framework based on SOM clustering with a threshold classifier
	Gaussian Mixture Model (GMM)	[23]	Applied a probabilistic model for superimposed fraud
	Latent Dirichlet Allocation (LDA)	[33], [62]	A probabilistic approach that used LDA and a secondary phase for separating fraudulent profiles
	ROCK algorithm and Subspace	[31]	Constructed a bi-level clustering methodology using ROCK clustering algorithm and subspace clustering
	Graph-based	[1]	Used a graph-based approach and a threshold classifier
Semi-supervised	SOM and <i>k</i> -means with an ensemble	[30]	Used bagging and boosting ensembles to create classifiers from Decision Trees, SVMs, and ANNs

5- Challenges

The fraud detection process is hindered by various challenges that are explained briefly in this section.

5-1- Concept Drift

Concept drift refers to the condition of an online supervised learning system where the distribution of the input and output changes, which will affect the prediction model, and can be defined as [36]:

$$\exists X: p_{t_0}(X, y) \neq p_{t_1}(X, y)$$

Equation 1

where p_{t_0} is the joint distribution at time t_0 , X refers to the input features, and y refers to the output. In supervised learning, the model is trained with the input features X and the respective output y . In the prediction phase, a new set of (previously unseen) input features X is given and the aim is to predict the output y . Concept drift can happen when normal behaviours keep evolving or altering, for example when the purchasing behaviour of customers changes on especial occasions such as the new year, or when fraudsters change their attacking methods. Hence, the model cannot perform accurate predictions since, under a more general perspective of drift, the relationship between the input features and the output has changed. Concept drift thus requires either updating the model incrementally or re-training it with recent batches of data [36], [56]. Adaptive learning is a solution to the concept drift problem where classical learning is not suitable. It is an advanced method of incremental learning in a non-stationary environment where the system has the capability of adapting to the stream of data [36], [57].

In certain cases, the occurrence of drift is cyclic and expected (e.g. changes in the buying preferences of customers during holidays) [56] while typically it is unanticipated and it may happen erratically. An optimal fraud detection system is expected to be able to adapt to concept drift quickly whether it is cyclic or unexpected, and also to distinguish it from noise (some learning algorithms interpret noise as concept drift) [56].

According to [56], there are three types of approaches that can handle concept drift, namely instance selection, instance weighting, and ensemble learning. In the instance selection approach, instances that are relevant to the concept are selected from the recent batches of data using a window. The pertinence of the instances is determined by how well the current model can classify them. In instance weighting, algorithms that can handle concept drift by themselves using weighted instances are used (e.g. support vector machines), however, instance weighting is prone to overfitting and [63] showed that it is inferior to instance selection. Instances are weighted by two factors, namely their age, and their appropriateness to the current concept. The ensemble learning approach tries to regularly replace the old batches of data with the most relevant and recent batches of data [64]. It hoards a series of concept descriptions, predictions that are merged by voting, weighted voting, or merely the most pertinent description is picked.

5-2- Imbalanced Data Distribution

A common problem in real-world datasets is that distributions are often imbalanced (also known as skewed data distributions). In an imbalanced binary dataset, the

instances are not equally distributed amongst classes as one class, usually known as the majority class, includes more instances than the other class, which is called the minority class [65]. For instance, in a data set that is related to medical diagnosis, there might be only a few cases that have cancer with many cases being normal. This is a serious problem for supervised learning algorithms where often there are only scarce abnormal instances for training, which makes training hard due to the resulting skewed distribution [66]. In a typical imbalanced dataset, the ratio between the minority and majority classes can be, for example, 1 to 100, 1 to 1,000, 1 to 10,000 or even more [67]. The proposed methods for dealing with the imbalanced data distribution problem can be categorised into algorithmic methods and data level methods [6], [68]. At the data level, some instances are replicated or removed to balance the dataset. Under-sampling is the notion used when a portion of the majority class is removed in order to re-balance the distribution of the dataset. In contrast, over-sampling is the process by which some instances of the minority class are replicated to obtain the balance. Both approaches come with some disadvantages. Under-sampling can remove useful data while over-sampling often causes over-fitting and also increases the training time as it enlarges the dataset [69]. It is also possible to apply both under-sampling and over-sampling especially when the dataset is profoundly imbalanced or when the minority class is extremely small [67].

At the algorithmic level, there are various approaches including: (i) cost-sensitive learning that tries to offset the misclassification by putting a cost-variable, (ii) adjusting the decision threshold when using one-class classification where the model is trained merely with the target class, and (iii) adjusting the probability of the estimate when using decision trees [67]. Another solution is to apply various algorithms that are capable of dealing with skewed distributions (meta-learning) [68], [70]. In meta-learning, various classifiers are utilised to carry out the classification task, and then, their performances are integrated, via an ensemble, to outperform classification with a single classifier.

5-3- Curse of Dimensionality

Telecommunication companies produce a large amount of data every day [34]. One aspect of this consists of a significant number of attributes, which together form a high-dimensional space that can cause several problems, often encapsulated by the term 'curse of dimensionality'. In high-dimensional space, data instances become more spread out, leading to decreased density, which in turn causes the convex hull to become stretched and difficult to distinguish [71]. High-dimensional datasets are very complicated, require larger amounts of memory and cause

longer computing time that make the detection process extremely difficult and time consuming [30], [72]. Therefore, dimensionality reduction is a crucial preprocessing step especially in telecommunication fraud detection. Its goal is to reduce the dimensions and complexity of a high-dimensional dataset without losing valuable information [73]. There are two main approaches for dimensionality reduction, namely feature selection and feature extraction. The aim of feature selection methods is to extract a smaller portion of the features that contains useful information and excludes noisy, redundant and irrelevant features [74]. In feature extraction, the goal is to embed the high-dimensional dataset into a lower dimensional space thus reducing the number of effective attributes [75].

Feature selection includes three methods, which are filter, wrapper, and embedded methods [75][74]. In filter methods, which act as a pre-processing step, the features are ranked using different criteria and scoring functions, then top ranked features are selected. Wrapping methods use the classifier itself to evaluate the features and have three categories, namely forward wrapping, backward wrapping, and forward-backward wrapping. Forward wrapping adds features gradually to the classification until the optimum feasible improvement is achieved. In contrast, backward wrapping tries to remove features gradually until no further improvement is feasible, and in forward-backward wrapping, features are added and also removed until maximum improvement is achieved. In embedded methods, the optimal features are selected during the model construction process using classifiers that have embedded feature selection methods.

5-4- Real Time Detection

As mentioned earlier, there are two different modes in fraud detection, namely online and offline modes. In a fraud detection system that is working in online mode, it is crucial to minimise the gap between the time when the fraud happened and the time when it was detected (known as the median duration) [76]. In fact, minimising the median duration can profoundly decrease the financial loss caused by the fraudster. Reducing the amount of data needed is considered as an effective method to achieve a system that is capable of real time detection as it can cause less memory usage and shorter computing time [30], [72].

5-5- Availability of Data

The paucity of publicly accessible data to perform research on is one of the issues that hinders doing research in this area [11], [35]. Companies are usually not keen on providing their data to researchers due to the confidential information that the data contain. Also, sometimes there are laws that prevent companies from furnishing researchers with data for experimental purposes.

Companies also avoid exposing details of their FDSs, because they believe this can help fraudsters understand the underlying mechanisms and create new techniques to avoid detection [34].

5-6- Noisy Data

Most real-world datasets are incomplete, noisy, and contain redundant, or obsolete records [30]. Therefore, many researchers tend to apply a preprocessing step before designing their model to clean the dataset and transform the dataset to a suitable form. As [30] explains, data preprocessing consists of three steps, namely data cleaning, data integration, and data dimensionality reduction. In data integration, the purpose is to deal with missing values, outliers, and erratic data. Data integration tries to deal with data (usually disparate) that are derived from various sources and maintaining them in one set. Data dimensionality reduction, as explained earlier, tries to transform high dimensional data into a lower dimension space. Noisy data can cause severe effects on the fraud detection process, especially with regards to accuracy. Noise is known as meaningless data that can cause variations in observations [5]. The difference between noise and outliers is that the former is not in the interest of the system and could have been caused by human error for instance. On the other hand, an outlier is a meaningful anomaly and is generally of interest to the system. It is worth noting that sometimes algorithms consider noise as outliers (in this case the outlier is the fraud instance) [77], which proves the importance of a preprocessing step prior to training the model. This is basically because the root cause of noise can be random or intentional (i.e. generated by a fraudster) [5]. Thus, FDSs should be capable of distinguishing between noise and actual outliers.

5-7- Misclassification Costs

Misclassification happens when a non-fraudulent instance is incorrectly classified as a fraudulent instance (also known as a false positive), or **when** a fraudulent instance is classified as a non-fraudulent instance (also known as a false negative). In fraud detection, the cost of a false positive misclassification is unequal to the cost of a false negative misclassification [78]. To explain further, the cost of a false negative is more expensive than a false positive because a false positive can be classified correctly after further investigation, but a false negative means that the fraudster has managed to stay undetected and can continue committing fraud. Therefore, in an FDS, a lower false negative error rate is much more important than a false positive error rate. However, it should not be deemed that the false positive rate is trivial. Further investigation requires human resources and is expensive, thus, a system with a high false positive rate can be a problem especially

for companies and organisations with limited budget and human resources [79].

Table 4 Accuracy and AUC results of papers in Table 3

Ref.	Method Investigated	Accuracy	AUC
[30]	SOM, k-means with an ensemble of Decision Trees, SVM, and ANN	83.4% - 89.8%	0.796 - 0.948
[33]	A probability approach that used LDA and an automatic threshold classifier		0.967 - 0.998
[11]	Applied a supervised learning method using a multilayer perceptron (MLP)	56.1% - 98.71%	0.997
[32]	A framework based on SOM clustering with a threshold classifier	60% - 87.75%	0.717 - 0.936
[22]	Applying Quarter-Sphere SVM which is a formulation of One-Class SVM	90.0%	
[35]	A comparison between the performances of SVM and ANN	98.67% - 98.87%	0.997 - 0.985

6- Evaluation Metrics

Model evaluation is very important, because it allows one to conduct performance comparisons between different proposed systems. Besides that, evaluation makes it possible to compare different approaches, to find the best algorithm, and optimise it further for a specific problem. This section reviews the evaluation metrics used by the papers that are tabulated in Table 3.

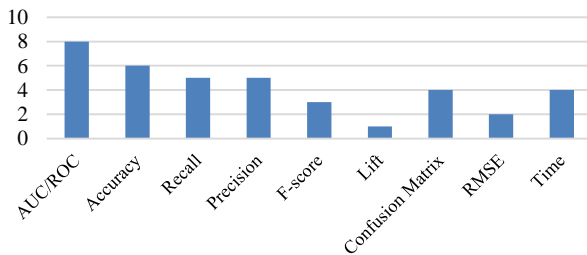


Figure 4 The number of papers in Table 3 that used evaluation metrics

As Figure 4 shows, the most widely used evaluation metrics in telecommunication fraud detection systems are accuracy (Equation 2) and AUC/ROC. However, according to [80], accuracy can lead to incorrect conclusions when it is used under certain conditions including skewed data, whereby the metric becomes biased towards the class with the majority of instances (please refer to Section 5-2 for more details).

Another very common performance metric used in this area for evaluation is the Receiver Operating Characteristic (ROC) curve, which basically visualises the probability of fraud detection versus the probability of false alarm, and the Area under ROC, also named AUC, is

the area under the curve in which 1 is the perfect value [32], [81]. In cases where the model does not depend on a threshold classifier, the area under ROC metric can be superior to accuracy [30]. Although, imbalanced distributions have no effect on ROC, which makes it very attractive for datasets with skewed distributions [82], ROC curves can generate an optimistic performance evaluation in the case that the data is significantly skewed [83]. Besides the aforementioned challenges, there are other issues that should be noted. An integral requirement of a supervised learning approach is labelled data, however, its availability is often an issue, moreover, labelling can be costly, time consuming, and requires an expert [71]. Also, an anomaly can have various meanings in different application domains as some have a more generic form while others have a specific form [48], and often it is very hard and expensive in some areas to provide labels for anomalous cases such as failures in aircraft engines [84]. Moreover, the performance of fraud detection systems depends heavily on the sources of data, given that data often originates from different sources with different formats and standards [76]. For instance, attributes can be binary, categorical, continuous, or a mixture of these.

Table 5 Summary of recommended evaluation metrics

Metric	Advantages	Disadvantages
Accuracy	Frequently used, very traditional, general and intuitive	Can be biased towards the majority class
AUC/ROC	Conceptually simple, visual performance evaluation and immune to skewed data	Can generate optimistic performance evaluations under large skewed distributions
FNR	Simple and an important financial factor	Not a thorough measurement

It is also good to introduce some metrics that are used less frequently. While some of the authors paid more attention to the time factor by revealing the detection time and also the computation time of their approaches, Root Mean Square Error (RMSE), F-score (Equation 5), and lift (Equation 6) evaluation metrics were rarely used. F-score is a measurement that shows the harmonic mean of precision and recall at a certain threshold, and lift (LFT) evaluates the true positive rate in the fraction of instances that are higher than the threshold [85]. In Table 4 the performance of the research papers that are tabulated in Table 3 are presented based on accuracy and area under ROC.

Choosing the right evaluation metric is often a problem dependent process. While a metric might fit perfectly to some problems, it may be unsuitable for other problems. Based on previous works in this area, it can be concluded that accuracy is a useful metric for performance evaluation, although it should not be concluded that it is sufficient for determining whether a proposed approach is suitable or not. In telecommunication fraud detection, ROC and AUC are vital metrics that can present important

information. The advantages of ROC consist of being conceptually simple and useful for experiments in which the data is skewed, and giving a more extensive measure of classification performance [82]. Also, the false negative (FN) rate should be considered when evaluating an FDS. A high FN rate means a large number of fraudulent cases are determined as non-fraudulent cases by the system, which can cause huge financial losses as there will be no further investigation on them. However, it should be noted that solely FN rate cannot represent a comprehensive evaluation of the system as it basically concentrates on merely one specific factor. Table 5 shows a summary of the pros and cons of three evaluation metrics that are recommended by this research work.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad \text{Equation 2}$$

$$Recal = \frac{TP}{TP + FN} \quad \text{Equation 3}$$

$$Precision = \frac{TP}{TP + FP} \quad \text{Equation 4}$$

$$F - score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad \text{Equation 5}$$

$$Lift = \frac{\%of\ positive\ above\ the\ threshold}{\%of\ dataset\ above\ the\ threshold} \quad \text{Equation 6}$$

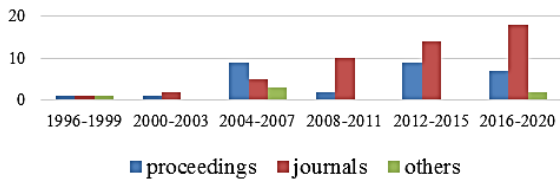


Figure 5 Source and published year of references in this paper

7- Discussion and Analysis

In previous sections, fraud types such as superimposed fraud, and subscription fraud were explained briefly. Also, various major challenges and issues in telecommunication fraud that hinder the detection process such as the curse of dimensionality and data imbalance data were reviewed.

The challenge that has received the least attention in telecommunications is arguably the problem of imbalanced class distributions. The two major works covering this issue are [20] and [30]. The authors in [20] preferred to use oversampling (they tripled the fraudulent cases), and they claimed this would avoid biasing the neural network toward classes with more instances. In contrast, in [30] the authors argued that oversampling has no advantage as it adds no new information. Therefore,

they preferred to employ under-sampling to balance the dataset and tackle the issue of a skewed class distribution. In an attempt to design a detection system that is capable of performing real-time subscription fraud detection, [20] designed a prediction model that used a multilayer perceptron neural network to evaluate customers and detect potential fraudsters at the time of subscribing, and a classification module that employed fuzzy rules to classify subscribers into four categories based on their previous behaviour. However, the performance of this approach becomes questionable when there is no previous record of new customers. In [30] the authors argued that by reducing the dimensionality of the data it is possible to reduce the time and memory needed for the algorithm, however, they believed that the problem of classifying residential subscribers (i.e. subscription fraud) is not under the category of real-time applications as there is time to perform the detection. Nonetheless, the proposed model was claimed to be capable of working in real-time as well. In another approach [33], by using merely three variables instead of a large range of variables, the authors managed to build a model that was capable of performing close to real-time detection, and did not require waiting for additional data to perform detection.

Recall that one of the major problems of fraud detection in the area of telecommunications was concept drift. Within Table 3, [31] tried to use every profile (also known as signature) only for a short period of time, and updated the profile gradually as behaviours of users evolved. They also discarded the old records or decreased their weights.

To avoid taking into consideration the noise and redundancy caused by combining all features of the high-dimensional space of the original data, [31] used a subspace clustering method, which is capable of disregarding unimportant attributes in each cluster. In [30] and [22] the authors took a different approach and preferred to use principle component analysis (PCA) to select the best features. To reduce the inner-dimensionality (fields of records) of variables, [32] employed SOMs to plot the variables to a SOM grid, thus projecting a multidimensional space into a 2-dimensional space, reflecting pattern similarities. Pruning was used to reduce the size of data in [1]. The authors of [39] selected half of the variables at their disposal to decrease complexity, and generated two additional variables for the purpose of better detection.

Cleaning the dataset from outliers or noise is also important to improve the performance (refer to Section 5 for more details). The authors of [35] utilised descriptive statistics, graphical methods, and Z-score standardisation to identify outlier values and remove them. In another work, [11] developed a model by employing neural networks that are capable of producing good performance even when the dataset contains noise. In [30] a thorough

preprocessing step to clean data and deal with missing values, outliers and inconsistent data, was used.

Some of the research works in Table 3 such as [30] and [19] are based on real-world data sets while others such as [33] are based on simulated data sets. Also it is worth mentioning that some research work such as [4] is based on data sets received from third parties like banks. Another type of data set used in telecommunication fraud detection is internal audit data, which is used to detect employee fraud and misconduct. The authors of [86] tried to detect fraudulent activities in a telecommunication company based on internal audit data.

8- Conclusions

In this paper, the aim was to provide a review of different fraud systems in telecommunications. Two different fraud systems, namely fraud detection systems and fraud prevention systems, were investigated with more focus on the former system as it has arguably more potential for improvement. The mechanisms used in fraud detection systems were studied and divided into three categories, namely rule-based, visualisation, and user-profiling systems. There is no standard or uniform way to categorise fraud types, therefore, researchers have divided fraud into several groups based on different factors. This research paper attempted to review the most prevalent and thorough approaches of categorising fraud types. Four fraud types that recently were investigated in the literature are superimposed fraud, subscription fraud, toll fraud, and SIM box fraud, where superimposed fraud witnessed the most attention from researchers. Likewise, various major challenges and problems that hinder the fraud detection process were studied in this research work. Some major challenges that hinder performance are real-time constraints, skewed data, concept drift, and high-dimensionality. In addition, we tried to explore evaluation metrics that were frequently used in the literature for measuring the performance and efficiency of the proposed systems and recommended three metrics that are profoundly vital in measuring system performance, namely accuracy, AUC/ROC, and FN rate. Moreover, the paper presented the performance of several recently designed fraud detection systems since 2011 in terms of accuracy and area under ROC.

9- Future Work

The performance of a fraud detection system is very dependent upon the data that it was designed for. Therefore, it would be impactful to investigate the performance of the same system on different datasets. Moreover, FDSs in the literature have often focused on building a system that is exclusively designed for either

offline or online detection. However, an optimal system should be able to effectively integrate both these types of fraud detection.

References

- [1] W. Henecka and M. Roughan, "Privacy-Preserving Fraud Detection Across Multiple Phone Record Databases," *Dependable Secur. Comput. IEEE Trans.*, vol. 12, no. 6, pp. 640–651, Nov. 2015.
- [2] E. I. Tarmazakov and D. S. Silnov, "Modern approaches to prevent fraud in mobile communications networks," in *Proceedings of the 2018 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering, ElConRus 2018*, 2018, vol. 2018-Janua, pp. 379–381.
- [3] M. Pejic-Bach, "Invited Paper: Profiling Intelligent Systems Applications in Fraud Detection and Prevention: Survey of Research Articles," in *2010 International Conference on Intelligent Systems, Modelling and Simulation*, 2010, pp. 80–85.
- [4] Y.-J. Zheng, X.-H. Zhou, W.-G. Sheng, Y. Xue, and S.-Y. Chen, "Generative adversarial network based telecom fraud detection at the receiving bank," *Neural Networks*, vol. 102, pp. 78–86, 2018.
- [5] M. Behdad, L. Barone, M. Bennamoun, and T. French, "Nature-Inspired Techniques in the Context of Fraud Detection," *IEEE Trans. Syst. Man, Cybern. Part C (Applications Rev.)*, vol. 42, no. 6, pp. 1273–1290, Nov. 2012.
- [6] A. Abdallah, M. A. Maarof, and A. Zainal, "Fraud detection system: A survey," *J. Netw. Comput. Appl.*, vol. 68, pp. 90–113, 2016.
- [7] J. West and M. Bhattacharya, "Intelligent financial fraud detection: A comprehensive review," *Comput. Secur.*, vol. 57, pp. 47–66, 2016.
- [8] R. J. Bolton and D. J. Hand, "Statistical Fraud Detection: A Review," *Stat. Sci.*, vol. 17, no. 3, pp. 235–255, 2002.
- [9] Y. Kou, C.-T. Lu, S. Sirwongwattana, and Y.-P. Huang, "Survey of fraud detection techniques," in *Networking, Sensing and Control, 2004 IEEE International Conference on*, 2004, vol. 2, pp. 749–754 Vol.2.
- [10] S. Wang, "A comprehensive survey of data mining-based accounting-fraud detection research," in *2010 International Conference on Intelligent Computation Technology and Automation, ICICTA 2010*, 2010, vol. 1, pp. 50–53.
- [11] A. H. Elmi, S. Ibrahim, and R. Sallehuddin, "Detecting SIM Box Fraud Using Neural Network," in *IT Convergence and Security 2012*, J. K. Kim and K.-Y. Chung, Eds. Dordrecht: Springer Netherlands, 2013, pp. 575–582.
- [12] P. Gosset and M. Hyland, "Classification, detection and prosecution of fraud in mobile networks," *Proc. ACTS Mob. summit, Sorrento, Italy*, 1999.
- [13] V. Jain, "Perspective analysis of telecommunication fraud detection using data stream analytics and neural network classification based data mining," *Int. J. Inf. Technol.*, vol. 9, no. 3, pp. 303–310, 2017.

- [14] H. M. Marah, O. M. Elrajubi, and A. A. Abouda, "Fraud detection in international calls using fuzzy logic," in *Proceedings - International Conference on Computer Vision and Image Analysis Applications, ICCVIA 2015*, 2015.
- [15] L. Cortesão, F. Martins, A. Rosa, and P. Carvalho, "Fraud management systems in telecommunications: a practical approach," in *12th International Conference on Telecommunications*, 2005, pp. 167–182.
- [16] T. Fawcett and F. Provost, "Adaptive Fraud Detection," *Data Min. Knowl. Discov.*, vol. 1, no. 3, pp. 291–316, 1997.
- [17] C. S. Hilas and J. N. Sahalos, "An Application of Decision Trees for Rule Extraction Towards Telecommunications Fraud Detection," in *Knowledge-Based Intelligent Information and Engineering Systems: 11th International Conference, KES 2007, XVII Italian Workshop on Neural Networks, Vietri sul Mare, Italy, September 12-14, 2007. Proceedings, Part II*, B. Apolloni, R. J. Howlett, and L. Jain, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 1112–1121.
- [18] J. Lopes, O. Belo, and C. Vieira, "Applying User Signatures on Fraud Detection in Telecommunications Networks," in *Advances in Data Mining. Applications and Theoretical Aspects: 11th Industrial Conference, ICDM 2011, New York, NY, USA, August 30 -- September 3, 2011. Proceedings*, P. Perner, Ed. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 286–299.
- [19] P. Saravanan, V. Subramaniaswamy, N. Sivaramakrishnan, M. Arun Prakash, and T. Arunkumar, "Data mining approach for subscription-fraud detection in telecommunication sector," *Contemp. Eng. Sci.*, vol. 7, no. 9–12, pp. 515–522, 2014.
- [20] P. A. Estévez, C. M. Held, and C. A. Perez, "Subscription fraud prevention in telecommunications using fuzzy rules and neural networks," *Expert Syst. Appl.*, vol. 31, no. 2, pp. 337–344, 2006.
- [21] F. M. Kau and O. P. Kogeda, "Impact of Subscription Fraud in Mobile Telecommunication Companies," in *2019 Open Innovations (OI)*, 2019, pp. 42–47.
- [22] S. Patnaik, S. Subudhi, and S. Panigrahi, "Quarter-Sphere Support Vector Machine for Fraud Detection in Mobile Telecommunication Networks," *Procedia Comput. Sci.*, vol. 48, pp. 353–359, 2015.
- [23] M. I. M. Yusoff, I. Mohamed, and M. R. A. Bakar, "Fraud detection in telecommunication industry using Gaussian mixed model," in *2013 International Conference on Research and Innovation in Information Systems (ICRIIS)*, 2013, pp. 27–32.
- [24] M. Arafat, A. Qusef, and G. Sammour, "Detection of Wangiri Telecommunication Fraud Using Ensemble Learning," in *2019 IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT)*, 2019, pp. 330–335.
- [25] G. Phil and H. Mark, "Classification Detection and Prosecution of Fraud on Mobile Networks," in *Proceedings of ACTS Mobile Summit, Sorrento Italy, 1999*.
- [26] S. Kamas and M. A. Aydin, "SPIT detection and prevention," *Istanbul Univ. - J. Electr. Electron. Eng.*, vol. 17, pp. 3213–3218, 2017.
- [27] L. Carvajal, L. Chen, C. Varol, and D. Rawat, "Detecting unprotected SIP-based Voice over IP traffic," in *4th International Symposium on Digital Forensics and Security, ISDFS 2016 - Proceeding*, 2016, pp. 44–48.
- [28] G. Cho, J. Cho, Y. Song, D. Choi, and H. Kim, "Combating online fraud attacks in mobile-based advertising," *Eurasip J. Inf. Secur.*, vol. 2016, no. 1, pp. 1–9, 2016.
- [29] M. Yelland, "Fraud in mobile networks," *Comput. Fraud Secur.*, vol. 2013, no. 3, pp. 5–9, 2013.
- [30] H. Farvareh and M. M. Sepehri, "A data mining framework for detecting subscription fraud in telecommunication," *Eng. Appl. Artif. Intell.*, vol. 24, no. 1, pp. 182–194, 2011.
- [31] L. P. Mendes, J. Dias, and P. Godinho, "Bi-level clustering in telecommunication fraud," in *ICORES 2012 - Proceedings of the 1st International Conference on Operations Research and Enterprise Systems*, 2012, pp. 126–131.
- [32] D. Olszewski, "Fraud detection using self-organizing map visualizing the user profiles," *Knowledge-Based Syst.*, vol. 70, pp. 324–334, 2014.
- [33] D. Olszewski, "A probabilistic approach to fraud detection in telecommunications," *Knowledge-Based Syst.*, vol. 26, pp. 246–258, 2012.
- [34] C. S. Hilas and J. N. Sahalos, "User profiling for fraud detection in telecommunication networks," in *In: 5th Int. Conf. technology and automation*, 2005, pp. 382–387.
- [35] R. Sallehuddin, S. Ibrahim, A. M. Zain, and A. H. Elmi, "Detecting SIM box fraud by using support vector machine and artificial neural network," *J. Teknol.*, vol. 74, no. 1, pp. 137–149, 2015.
- [36] J. Gama, I. Zliobaite, A. Bifet, M. Pechenizkiy, and A. Bouchachia, "A survey on concept drift adaptation," *ACM Comput. Surv.*, vol. 46, no. 4, pp. 44:1–44:37, Mar. 2014.
- [37] Z. Shaeiri, J. Kazemitabar, S. Bijani, and M. Talebi, "Behavior-Based Online Anomaly Detection for a Nationwide Short Message Service," *J. AI Data Min.*, vol. 7, no. 2, pp. 239–247, 2019.
- [38] L. Manunza, S. Marseglia, and S. P. Romano, "Kerberos: A real-time fraud detection system for IMS-enabled VoIP networks," *J. Netw. Comput. Appl.*, vol. 80, pp. 22–34, 2017.
- [39] K.-I. Kim, T. Kim, N.-W. Cho, and M. Kim, "Toll Fraud Detection of VoIP Service Networks in Ubiquitous Computing Environments," *Int. J. Distrib. Sens. Networks*, vol. 2015, 2015.
- [40] J. Li, K.-Y. Huang, J. Jin, and J. Shi, "A survey on statistical methods for health care fraud detection," *Health Care Manag. Sci.*, vol. 11, no. 3, pp. 275–287, 2008.
- [41] C. S. Hilas, "Designing an expert system for fraud detection in private telecommunications networks," *Expert Syst. Appl.*, vol. 36, no. 9, pp. 11559–11569, 2009.
- [42] K. C. Cox, S. G. Eick, G. J. Wills, and R. J. Brachman,

- “Brief Application Description; Visual Data Mining: Recognizing Telephone Calling Fraud,” *Data Min. Knowl. Discov.*, vol. 1, no. 2, pp. 225–231, 1997.
- [43] W. N. Dilla and R. L. Raschke, “Data visualization for fraud detection: Practice implications and a call for future research,” *Int. J. Account. Inf. Syst.*, vol. 16, pp. 1–22, 2015.
- [44] J. A. Lasserre, C. M. Bishop, and T. P. Minka, “Principled Hybrids of Generative and Discriminative Models,” in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’06)*, 2006, vol. 1, pp. 87–94.
- [45] C. S. Hilas and J. N. Sahalos, “Testing the fraud detection ability of different user profiles by means of FF-NN classifiers,” *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 4132 LNCS, pp. 872–883, 2006.
- [46] C. S. Hilas and P. A. Mastorocostas, “An application of supervised and unsupervised learning approaches to telecommunications fraud detection,” *Knowledge-Based Syst.*, vol. 21, no. 7, pp. 721–726, 2008.
- [47] S. S. Rajani and M. Padmavathamma, “A Model for Rule Based Fraud Detection in Telecommunications,” in *International Journal of Engineering Research and Technology*, 2012, vol. 1, no. 5 (July-2012).
- [48] V. Chandola, A. Banerjee, and V. Kumar, “Anomaly Detection: A Survey,” *ACM Comput. Surv.*, vol. 41, no. 3, pp. 15:1–15:58, Jul. 2009.
- [49] J. Tamboli and M. Shukla, “A survey of outlier detection algorithms for data streams,” in *2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)*, 2016, pp. 3535–3540.
- [50] C. Hilas, P. Mastorocostas, and I. Rekanos, “Clustering of telecommunications user profiles for fraud detection and security enhancement in large corporate networks: a case study,” *Appl. Math. Inf. Sci.*, vol. 9, pp. 1709–1718, 2015.
- [51] J. Hollmén and V. Tresp, “Call-based Fraud Detection in Mobile Communication Networks Using a Hierarchical Regime-switching Model,” in *Proceedings of the 1998 Conference on Advances in Neural Information Processing Systems II*, 1999, pp. 889–895.
- [52] X. Zhu and A. B. Goldberg, “Introduction to semi-supervised learning,” *Synth. Lect. Artif. Intell. Mach. Learn.*, vol. 3, no. 1, pp. 1–130, 2009.
- [53] A. Daneshpazhouh and A. Sami, “Semi-Supervised Outlier Detection with Only Positive and Unlabeled Data Based on Fuzzy Clustering,” *Int. J. Artif. Intell. Tools*, vol. 24, no. 03, p. 1550003, 2015.
- [54] D. Olszewski, “Fraud Detection in Telecommunications Using Kullback-Leibler Divergence and Latent Dirichlet Allocation,” in *Adaptive and Natural Computing Algorithms*, 2011, pp. 71–80.
- [55] A. Chouiekh and E. L. H. I. E. L. Haj, “ConvNets for Fraud Detection analysis,” *Procedia Comput. Sci.*, vol. 127, pp. 133–138, 2018.
- [56] A. Tsymbal, “The Problem of Concept Drift: Definitions and Related Work,” 2004.
- [57] J. L. Lobo, J. Del Ser, M. N. Bilbao, I. Laña, and S. Salcedo-Sanz, “A probabilistic sample matchmaking strategy for imbalanced data streams with concept drift,” *Stud. Comput. Intell.*, vol. 678, pp. 237–246, 2017.
- [58] S. Augustin *et al.*, “Telephony fraud detection in next generation networks,” in *AICT 2012 - 8th Advanced International Conference on Telecommunications*, 2012, pp. 203–207.
- [59] Y. Alraouji and A. Bramantoro, “International call fraud detection systems and techniques,” in *MEDES 2014 - 6th International Conference on Management of Emergent Digital EcoSystems, Proceedings*, 2014, pp. 159–166.
- [60] X. Liu and X. Wang, “A Network Embedding Based Approach for Telecommunications Fraud Detection,” in *Cooperative Design, Visualization, and Engineering*, 2018, pp. 229–236.
- [61] G. Kaiafas, C. Hammerschmidt, R. State, C. D. Nguyen, T. Ries, and M. Ourdane, “An Experimental Analysis of Fraud Detection Methods in Enterprise Telecommunication Data using Unsupervised Outlier Ensembles,” in *2019 IFIP/IEEE Symposium on Integrated Network and Service Management (IM)*, 2019, pp. 37–42.
- [62] N. Ruan, Z. Wei, and J. Liu, “Cooperative Fraud Detection Model With Privacy-Preserving in Real CDR Datasets,” *IEEE Access*, vol. 7, pp. 115261–115272, 2019.
- [63] R. Klinkenberg, “Learning drifting concepts: Example selection vs. example weighting,” *Intell. Data Anal.*, vol. 8, no. 3, pp. 281–300, 2004.
- [64] A. D. Pozzolo, G. Boracchi, O. Caelen, C. Alippi, and G. Bontempi, “Credit card fraud detection and concept-drift adaptation with delayed supervised information,” in *2015 International Joint Conference on Neural Networks (IJCNN)*, 2015, pp. 1–8.
- [65] Y. Yan, M. Chen, M.-L. Shyu, and S.-C. Chen, “Deep Learning for Imbalanced Multimedia Data Classification,” in *Proceedings - 2015 IEEE International Symposium on Multimedia, ISM 2015*, 2015, pp. 483–488.
- [66] S. Al-Stouhi and C. K. Reddy, “Transfer learning for class imbalance problems with inadequate data,” *Knowl. Inf. Syst.*, vol. 48, no. 1, pp. 201–228, 2016.
- [67] N. V. Chawla, N. Japkowicz, and A. Kotcz, “Editorial: Special Issue on Learning from Imbalanced Data Sets,” *SIGKDD Explor. Newsl.*, vol. 6, no. 1, pp. 1–6, Jun. 2004.
- [68] C. Phua, D. Alahakoon, and V. Lee, “Minority Report in Fraud Detection: Classification of Skewed Data,” *SIGKDD Explor. Newsl.*, vol. 6, no. 1, pp. 50–59, Jun. 2004.
- [69] R.-C. CHEN, T.-S. CHEN, and C.-C. LIN, “A new binary support vector system for increasing detection rate of credit card fraud,” *Int. J. Pattern Recognit. Artif. Intell.*, vol. 20, no. 02, pp. 227–239, 2006.
- [70] S.-C. Lin, Y. I. Chang, and W.-N. Yang, “Meta-learning for Imbalanced Data and Classification Ensemble in Binary Classification,” *Neurocomput.*, vol. 73, no. 1–3, pp. 484–494, Dec. 2009.
- [71] V. J. Hodge and J. Austin, “A Survey of Outlier Detection Methodologies,” *Artif. Intell. Rev.*, vol. 22, no. 2, pp. 85–126, 2004.

- [72] L.-A. Gottlieb, A. Kontorovich, and R. Krauthgamer, "Adaptive metric dimensionality reduction," *Theor. Comput. Sci.*, vol. 620, pp. 105–118, 2016.
- [73] M. Sugiyama, "Local Fisher Discriminant Analysis for Supervised Dimensionality Reduction," in *Proceedings of the 23rd International Conference on Machine Learning*, 2006, pp. 905–912.
- [74] J. Miao and L. Niu, "A Survey on Feature Selection," *Procedia Comput. Sci.*, vol. 91, pp. 919–926, 2016.
- [75] S. Agarwal, P. Ranjan, and R. Rajesh, "Dimensionality reduction methods classical and recent trends: A survey," *Int. J. Control Theory Appl.*, vol. 9, no. 10, pp. 4801–4808, 2016.
- [76] A. Bănărescu, "Detecting and Preventing Fraud with Data Analytics," *Procedia Econ. Financ.*, vol. 32, pp. 1827–1836, 2015.
- [77] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, "LOF: Identifying Density-based Local Outliers," *SIGMOD Rec.*, vol. 29, no. 2, pp. 93–104, May 2000.
- [78] C. Phua, V. C. S. Lee, K. Smith-Miles, and R. W. Gayler, "A Comprehensive Survey of Data Mining-based Fraud Detection Research," *CoRR*, vol. abs/1009.6, 2010.
- [79] B. Barbarioli and R. M. Assuncao, "Anomaly Detection under Cost Constraint," in *Proceedings - 2016 5th Brazilian Conference on Intelligent Systems, BRACIS 2016*, 2016, pp. 247–252.
- [80] G. Haixiang, L. Yijing, J. Shang, G. Mingyun, H. Yuanyue, and G. Bing, "Learning from class-imbalanced data: Review of methods and applications," *Expert Syst. Appl.*, vol. 73, pp. 220–239, 2017.
- [81] Z. Chen, L. D. Van Khoa, E. N. Teoh, A. Nazir, E. K. Karuppiah, and K. S. Lam, "Machine learning techniques for anti-money laundering (AML) solutions in suspicious transaction detection: a review," *Knowl. Inf. Syst.*, vol. 57, no. 2, pp. 245–285, Nov. 2018.
- [82] T. Fawcett, "An introduction to {ROC} analysis," *Pattern Recognit. Lett.*, vol. 27, no. 8, pp. 861–874, 2006.
- [83] J. Davis and M. Goadrich, "The Relationship Between Precision-Recall and ROC Curves," in *Proceedings of the 23rd International Conference on Machine Learning*, 2006, pp. 233–240.
- [84] P. Gogoi, D. K. Bhattacharyya, B. Borah, and J. K. Kalita, "A Survey of Outlier Detection Methods in Network Anomaly Identification," *Comput. J.*, 2011.
- [85] R. Caruana and A. Niculescu-Mizil, "Data Mining in Metric Space: An Empirical Analysis of Supervised Learning Performance Criteria," in *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2004, pp. 69–78.
- [86] A. Nawawi and A. S. A. P. Salin, "Employee fraud and misconduct: empirical evidence from a telecommunication company," *Inf. Comput. Secur.*, vol. 26, no. 1, pp. 129–144, 2018.

Kasra Babaei is currently a PhD candidate at University of Nottingham Malaysia, received the MSc in Information Technology (Merit) from University of Nottingham in 2014, and his BSc in Business Administration from University of Payam Noor, Anzali Branch, Iran in 2011

Dr. Chen ZhiYuan currently is an Assistant Professor with the University of Nottingham Malaysia (UNM) and a Principal Consultant with MIMOS at the Accelerative Technology Lab. She received the MPhil and a PhD in Computer Science from the University of Nottingham. Before joining UNM, she has been a research associate in the UK Horizon Digital Economy Research Institute. Her research interests are in the area of computer science, machine learning, data mining, user modelling and artificial intelligence.

Tomas Maul received a BSc (Hons.) degree in Psychology from the University of St. Andrews, St. Andrews, UK, two MSc degrees in Computer Science from Imperial College, London, UK, and a PhD degree in Computer Science (Computational Neuroscience) from the University of Malaya, Kuala Lumpur, Malaysia. He is currently Associate Professor at the School of Computer Science, University of Nottingham Malaysia.

A Fast Machine Learning for 5G Beam Selection for Unmanned Aerial Vehicle Applications

Wasswa Shafik

Computer Engineering Department, Yazd University, Yazd, Iran
wasswashafik@stu.yazd.ac.ir

S. Mojtaba Matinkhah*

Computer Engineering Department, Yazd University, Yazd, Iran
matinkhah@yazd.ac.ir

Mohammad Ghasemzadeh

Computer Engineering Department, Yazd University, Yazd, Iran
m.ghasemzadeh@yazd.ac.ir

Received: 04/Mar/2019

Revised: 24/Aug/2019

Accepted: 08/Nov/2019

Abstract

Unmanned Aerial vehicles (UAVs) emerged into a promising research trend applied in several disciplines based on the benefits, including efficient communication, on-time search, and rescue operations, appreciate customer deliveries among more. The current technologies are using fixed base stations (BS) to operate onsite and off-site in the fixed position with its associated problems like poor connectivity. These open gates for the UAVs technology to be used as a mobile alternative to increase accessibility in beam selection with a fifth-generation (5G) connectivity that focuses on increased availability and connectivity. This paper presents a first fast semi-online 3-Dimensional machine learning algorithm suitable for proper beam selection as is emitted from UAVs. Secondly, it presents a detailed step by step approach that is involved in the multi-armed bandit approach in solving UAV solving selection exploration to exploitation dilemmas. The obtained results depicted that a multi-armed bandit problem approach can be applied in optimizing the performance of any mobile networked devices issue based on bandit samples like Thompson sampling, Bayesian algorithm, and ϵ -Greedy Algorithm. The results further illustrated that the 3-Dimensional algorithm optimizes utilization of technological resources compared to the existing single and the 2-Dimensional algorithms thus close optimal performance on the average period through machine learning of realistic UAV communication situations.

Keywords: Unmanned Ariel Vehicle; Multi-Armed Bandit; Reinforcement Learning Algorithms; Beam selection.

1- Introduction

The dynamics and advances in technology in respect to scientific studies have emerged and influenced a number of fields for instance distributed compressed sensing [1], text mining applications [2], besides its increased manifestation in wireless, semi-online and online application, in particular, Unmanned Aerial Vehicles (UAV).

Unmanned Ariel vehicles (UAVs) are aircraft deprived of an anthropological pilot either onboard or off board and a type of unmanned vehicle and are components of an unmanned aircraft system, to mention a UAV, a ground-based controller, and a system of communications. UAVs have got increased usage in a number of different industries like agri-businesses for precision agriculture operations, thermal imaging in road patrolling, path planning, rescue departments for post-natural disaster

missions, targets monitoring, crack assessments among other notable benefits as well [3]- [7].

Machine learning entails quite numerous categories for example supervised learning, unsupervised learning, semi-Supervised learning and reinforcement learning. This reinforcement learning appears in two categories that is to say negative and positive with unties like Markov Decision Process-learning among others. Reinforcement learning is specified under simple reinforcement learning and the deep reinforcement learning having deterministic policy gradient algorithms simultaneously learn a Q-function and a policy through the use of Bellman equation given that Deep Q-Networks modernizes the Q-value function of a state for a specific action simply; this proposed model uses a simple reinforcement learning and multi armed bandit not deep reinforcement learning.

Fast machine learning techniques confirm an enhanced potential in learning patterns and extracting attributes from a complex dataset includes the use of other learnings like deep learning for pattern recognition and medication innovation, although others do surface explicit encounters

* Corresponding Author

that ought to be worked on embrace anthropologies in sensor synthesis, associate medicinal analyses, and experimental decision-makings [9].

In appreciation of the models availed in problem analysis, we are proposing an alternative way to problem attack to network issues analysis to the identified medical network scenarios basing on the existing studies, models, and architectures, enlighten on the medical application illustrated above. A simplified mathematical expression of the armed bandit scenarios has been used, validated, and proofed calculations together with the illustration presented throughout this paper in comparison to the state-of-the-art models.

Some interesting questions are that machine learning tackles: Are there any correlations between crowdsourcing annotations with expert measurements to feed the fast machine learning training algorithms with quit satisfactory reliability? Can we use a hierarchical feature selection method for cancer detection? How we can make the multi-label biomedical compound more efficient? Is there any way to produce useful results by the use of computer games in embedding human intelligence-based tasks to train fast machine learning algorithms? Learning the lower-staged configuration of inventive data to achieve a more intangible portrayal is the central awareness of fast machine learning. How we design algorithms to implicitly capture the intricate associations and topographies of the larger-scale feedbacks?

The non-orthogonal schemes with multiple access capabilities for the 5G UAV social network that encompasses medical as well where they referred to the social network as communication. The author depicts that the systems of multiple access with no orthogonal ranges overtake extra numerous access of the same scheme capacities in the relation of summation capability, dynamism efficiency, and social phantom effectiveness. This accomplishes superior sum-rate on a subordinate loftiness that diminishes the inclusive dynamism outflow of the fifth generation UAV network.

The paper in the rest parts are structured as the following: Section 1.1 presents the motivation for this paper depicting early approaches and clearly showing the requisite for this study. In section 2, the study avails research literature focusing mainly on the UAVs' technological usage. In section 3, the detailed multi-armed bandit problem approach in comparison to other approaches in beam selection, in particular, we considered Thompson sampling, ϵ -greedy algorithm, the Bayesian Upper Confidence bounds bandits. Section 4 presents simulated results and discussion of the provided results provided. In section 5, we explicitly our simulation parameters and detailed

discussion of beam selection and section 6 hold the conclusion of this article.

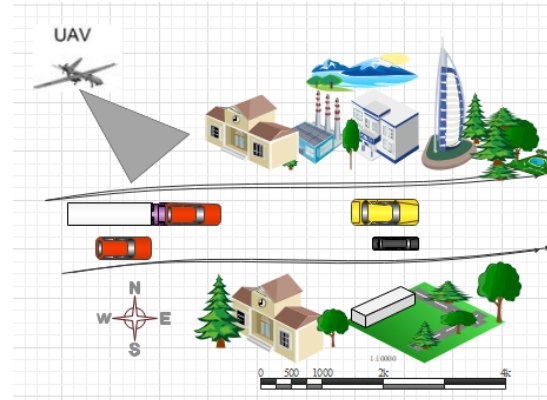


Fig. 1 The Simulation Atmosphere. The trolley and bus momentarily block the red car and black car correspondingly.

1-1- Motivation

UAVs preserves on advancing in capabilities and efficiency to increase its rapid evolution in different areas like medical, engineering, profitable, and entertaining applications, that is to say, aptitude to gather real-time data at cost-effectiveness, on-time delivery, on-time payloads deliveries. Regardless it's the sluggish extension in the multi- fields, some of the state-of-the-art studies do include a context-aware communication link [8], Improved communication security of UAVs [9], reduced communication delays through formation control [10], Multi-objective Placements [11], improved UAV wireless communication [12], mention but a few.

Based on the identified presented studies, it has been further identified that the beam that is emitted has not been given clear technological attention since current studies some of which focus on synthesis methods of the synchronized assemblage differential pulse-code modulation codec for UAV communication schemes [13], more, the application of UAV in communal protection communicating and optimizing of reportage [14]. Having discovered that beams emitted from the UAV stochastically and technologically behave like bandits, this paper, therefore, presents a multi-armed bandit' approach is selecting the best armed per the desire of the operator that has the approach to beam selection of UAVs.

In appreciation to the application of machine learning, the number of study analysis have been presented given a dynamic technological advancement, they led to the presence of the noise-resistant surface defect recognition tactics [15], color texture classification and identifications to solve difficulties involved in the texturing in computing to obtain better accuracy [16], according to the detailed study on learnings in remote sensing that categorizes the UAVs too included the concepts, apparatuses, and

encounters for the like a perceptive assignment investigation for emerging unmanned aerial vehicle boonies rifle sustenance [17]. Technologies involving expertise development monitoring, piloting tasks are seen in medical as well with optical brain imaging in conjunction with UAVs [18].

2- Research Literature

This section presents a general analysis of the proposed UAV models and frameworks that have been claimed in data gathering tasks, secure remote connection of social media platforms among others.

2-1- UAV Technological Usage

The 5G technologies are the subsequent cohort of wireless communication, submission earlier swiftness, and more consistent acquaintances on smart medical devices and supplementary devices than interminably before. The convergence of multiple networking functions to achieve charge, power, and complexity reduction is one of the promising advantages of 5G. The 5G expected to assist power an enormous intensification in the IoT technologies, in case of the arrangement compulsory to transfer mammoth quantities of data to have a smart world. The 5G involves multiple entities with different ideas like the extraordinary record, truncated expectancy, effectiveness enlargement, and yield extension.

2-2- Search and Rescue Operations

Murphy et al. [19] proposed autonomous UAV can find missing people using the broadcasting motions fashioned by the comprehensive schemes for itinerant communications booth. The UAVs act as a universal mobile communications entity BS and persuade the disappeared individual's maneuver to effort to variety communication. They recycled a constraint-centered topographic-based path arrangement tactic to harvest a course for the unmanned aerial vehicles to navigate in the inflight transient finished predictable signals since a huge quantity of conceivable spring settings for abundant resolutions; this assisted the telephone enactment accordingly prominent the request of augmented complexity, and efficiency.

2-3- Smart City Scenarios

The need to automate this data gathering practice, a network of UAV is a suitable option for a vehicle are the inspiration for this paper through performance analysis of 5G network and beyond. Authors to mention, El-Sayed et al. [20] deployed UAVs performing like moveable edges following traffic proceedings and overcrowding situations

prominent to the self-motivated attitudes, amplified mobility. The proposed traffic-awareness technique empowered the distribution of unmanned aerial vehicles in vehicular situations demonstrated the proposed technique to accomplish jam-packed social net analysis under diverse situations deprived of extra communication delay. This approach opened some technical questions that were not elaborated for instance what could the proposed model do in times of complex traffic from a big smart perspective [30].

2-4- A System with Pre-Computed Pre-coders

Meticulously within this subsection, M. Meng et al. [21] collective through the compensations of a fifth-generation mm-Wave detector, machine learning approaches with more methodological developments, for example, the multiple inputs multiple out to identify the "black flying" UAV. An operational resolution to disentangle the delinquent of distinguishing and categorizing a UAV by the fifth-generation mm-Wave detector of the Internet of Things, which has extraordinary applied solicitation assessment.

The existing technology improved the accessibility of social networks, rapid rate, nonetheless the next fifth-generation (5G) is anticipated to increase connectivity and increased data rate that multiplies the existing one thousand times regardless of the location and data size. The advance in technology still has a deficiency of proper infrastructure as services that can accommodate all possible scenarios of mobility, this contributed mainly to its supposition in this study.

After the close discovery that numerous artificial Intelligence applications are emerging entailing UAVs with artificial capabilities including knowledge engineering, edge mobility performance assessments, a number of reinforcement learning issues agent Q-learning, Learning Automata and many more [22]–[24], numerous studies have focused on low complexities leaving a group of the multi-cells. Scrutinizing the intricate, vibrant relations of the number of dynamic drifts in the numerous booths are acquaint with the resilient control of interfering amongst immediate BS [25].

Furthermore, in case the non-orthogonal learning classifications are recycled of uplink learning, it is revealed that the precoding environment recycled by the BS in single-cell converting to besmirched through the station amongst that base station and the users in extra cells [26]. Nevertheless, in multiple cells do support the customer tools necessities to assess the conduit public data besides nourish them spinal to the BS schedulers for

adaptabilities reserve supervision. This leads to a substantial rise of gesticulating overheads and feedbacks expectancy addicted to the complaisant networks of UAVs.

3- Multi-Armed Bandit and Beam Selection

This section presents the multi-armed bandit approach and most related arenas like Thompson sampling, ϵ -greedy algorithm, and Bayesian Upper confidence bounds (UCB) algorithm in solving medical devices. The MAB problematics are decisive problems in nature that well reveals the survey against the utilization quandary of medical computation. There exist numerous techniques to adopted using MAB containing no exploration connotation that the maximum unexperienced approach in 2-dimensional; vigorously this study follows but in 3-dimensional approach optimization per the model explained, according to 22 and 23 of the pseudo [28].

This concept proves that j-armed problems associated with the bandit might be elucidated by disentangling j-armed glitches. Contemplate a chronological verdict delinquent in that at correspondingly stage there are j potential actions, optimal action of j results in an estimation being occupied from the j^{th} experimentation and the received the mathematical assessment of this observation as a reward. The annotations taken may provide useful information in forthcoming choices of actions. To exploit the contemporaneous assessment of the infinite tributary of rewards received, reduced in an approximately. The Bandit is attained from demonstrating difficulties, for example, an i -armed bandit, that is a slot engine with i arms, each subsequent in a mysterious, perchance divergent scattering of payoff s.

It is quite challenging to notice which arm (UAV beam) provides the utmost average return in technologies like determining the rates. However, by playing the numerous arms of the slot engine, the information on which arm is best may be obtained. Nonetheless, the observations taken to use to have information are similarly users' rewards. Striking equilibrium among gaining rewards with acquisition information a case, for instance, it is not garbed to uninterruptedly wrench the arm that has achieved best in the previous; subsequently, it might have to be situated that individual was impartial unfortunate with the preeminent arm.

Classically in this kind of problem as wireless communication to obtain, there is a time of gaining information, monitored by a period of narrowing down the arms, monitored by a period of playing the arm to be the best. More important modeling to illustrate these problems comes from operational trials in which there are p drones for a given mission.

It is assumed that response from the operation is instant so that the UAV efficiency based on the mission that the

present UAV does an operation that is acknowledged once the other UAV requisite is preserved. It is in fact not acknowledged exactly which unique of the operational missions is best; nonetheless decide which operations to give each drone, remembering that the primary goal is to serve as many UAVs as possible. This may require a drone mission procedure that is not the one that seems utmost at the existing interval to attain info that might be of monotonous to forthcoming UAVs.

Assume that y reward distributions be denoted by $D_1(v|\vartheta_1), \dots, D_y(v|\vartheta_y)$ where $\vartheta_1, \dots, \vartheta_y$ are parameters considering that the tenets are not identified exactly, nonetheless since combined erstwhile scattering is to be recognized and the study denotes it to be $P(\vartheta_1, \dots, \vartheta_y)$.

Primarily, action θ_1 will be chosen from a provided set $\{1, \dots, y\}$, observation 0_1 , the recompense for the initial phase, is taken from the scattering $D\theta_1$ might also be based on this information, action θ_2 of the same action space, and an observation 0_2 , taken from $D\theta_2$. Let us also

assume that given θ_m , the parameters $\theta_1, \dots, \theta_y, 0_m$ to be chosen from $D\theta_m$ autonomously of the previous. The choice instruction for this issue is a categorization $S = (\theta_1, \theta_2, \theta_3, \dots)$ of utilities adjusted to the interpretations; that is to say, it may be contingent on earlier engagements and interpretations putting in mind that the denotation of 0_m shows both past action and observation leading.

$$\theta_m(\theta_1, 0_1, \theta_2, 0_2, \dots, \dots, \theta_{m-1}, D_{m-1}) \quad (1)$$

There are possibilities discounted sequences, let us denote it to be K , $K = (\beta_1, \beta_2, \dots)$ such that the i th observation is discounted β_i were $0 \leq \beta_i \leq 1$ for. The maximum expected rewards are presented as $E \sum_i \beta_i \cdot 0_i$ since the complete

discounted return is $\sum_i \beta_i \cdot 0_i$. This delinquent is christened

to be i -armed MAB problem continuously yields the identical recognized amount that is, the circulation D connected with one of the arms is debauched at an identified relentless.

Therefore, to accomplish a finite rate for the predictable remuneration, Let us adopt

(a) each distribution, D_i will be $i = 1, \dots, y$, has finite first moment,

(b) also consider $\sum_i \beta_i \cdot p \infty$, additional dualistic substantial

circumstances of the discount sequence are

(c) the m -horizon unvarying reduction intended for which

$$\beta_1 = \dots = \beta_m = 1 \ \& \ \beta_{m+1} = \beta_{m+2} = \dots = 0,$$

(d) the geometric discount in which K , will be given by $K = (1, \beta, \beta^2, \beta^3, \dots)$, logically meaning that $\beta_i = \beta^{i-1}$ for $i = 1, \dots, \dots$.

The payoff is basically $\sum_1^m 0_i$, the entirety of the very initial m clarifications. The problem converts a unique through a finite perspective which can in the method be disentangled by backward generation. Around is an interlude alteration in which the approaching m phases organized accomplishment at the twitch excluding for the alteration from the previous delivery to the advanced dissemination. The extravagance predominantly difficulties with symmetrical discount and autonomous arms, that is to say, prior scatterings P for that $\theta_1, \dots, \theta_y$ are autonomous accordingly that surveillance on the single-arm will not affect the acquaintance of the dissemination of slightly supplementary arm.

As an introductory to the explanation of the i -armed bandit with symmetrical reduction and autonomous arms, the study prerequisite to comprehend a modestly MAB as the nature of wireless complications is. The single or multi-armed bandit is really a bandit problem issue, but a single arm may be less useful during the dissemination of takings, besides so plays only an insignificant role. Firstly (1), the study further depicted how the MAB can be associated with a preventing rule problem.

Take responsibility that a prearranged arm has a concomitant categorization of unsystematic variables, x_1, x_2, x_3, \dots with recognized joint distribution satisfying $SUP_m X_m^+ p \infty$. For the new arm, the returns are assumed firstly not considered from a known distribution with expectation λ . The discount sequence in the case taken to be geometric $P = (1, \beta, \beta^2, \dots)$ $0 \leq \beta \leq 1$ was looking for a decision rule $\varphi M = (\theta_1, \theta_2, \theta_3)$ to maximize

$$\varphi(M) = E \left(\sum_1^{\infty} \beta^{i-1} 0_i \mid \varphi M \right) \quad (2)$$

The benefit of constructing this remark so that study might nowadays assume that the resolution imperative of φM ensures not be contingent over 0_i when $\theta_i = 2$, since 0_i is acknowledged to stand λ . Consequently, the study dons that the additional arm stretches a continuous reappearance of λ collectively interval it is dragged surveyed by (2).

First theorem: In case it is primarily optimum to habit the second arm in the sense that $SUP_{\varphi M} \varphi^* = \sup \{ \varphi(\varphi M), \text{ where } \theta_i = 2 \}$, then it is optimal to use the second arm continuously and therefore the

$\varphi M^* = \lambda / (1 - \beta)$. Second theorem: Consider that $\forall(\beta)$ designate the optimum degree of return for by means of the first arm at the concession β . Let us prove all the assumptions in the fast lemma. Firstly, in the case $\epsilon \in 0$ to obtain the decision rule of M in that rule $\theta_1 = 2$, lastly $\varphi(M) \geq \varphi^* - \epsilon$ this is given computed as

$$\begin{aligned} \varphi(M) &= \lambda + \beta E \left(\sum_2^{\infty} \beta^{i-2} 0_i \mid M \right) = \lambda + \beta E \left(\sum_2^{\infty} \beta^{i-1} 0_{i+1} \mid M \right) \\ &= \lambda + \beta E \left(\sum_2^{\infty} \beta^{i-1} 0_i \mid M \right) = \lambda + \beta E \left(\sum_2^{\infty} \beta^{i-2} 0_i \mid M^1 \right) \leq \lambda + \beta \varphi^* \end{aligned}$$

Whereas the rule M shifted by 1, and $0_i = 0_{i+1}$. Consequently, we have $\varphi^* - \epsilon \leq \lambda + \beta \varphi^*$ subsequently $\epsilon \in 0$ subjective meaning, this implies

$$\varphi^* \leq \lambda / (1 - \beta) \quad (3)$$

Nonetheless, this value obtained at (3) is achievable by using the second arm at the respective phase. This is also understood that theorem is likewise effective in the n -uniform reduction categorization. It is considered to be discount categorization P is supposed to be consistent in case it has a cumulative failure degree, i.e. in the case

$\beta_m / \sum_m^{\infty} \beta_i$ is not decreasing on its definition domain.

Notably, the above theorem is not factual for roughly reduction classifications is readily comprehended a case in the example in case $P = \{0, 1, 1, 0, \dots\}$ meaning that P is regular yet x_i is exchangeable like at 10, other at 0 and $\lambda = 0$ formerly the solitary optimum stratagem is to track an initial wrench of that second arm with a wrench of the first arm. Accurately what assets of P is compulsory for the above theorem appears to be unidentified.

Considering lemma 2 nearby exist an optimum instruction for this problem, it is correspondingly the rule that customs second arm at completely stages, or the statute compatible to the terminating rule $L \geq 1$ that is optimum for the terminating rule delinquent with payoff and communicated as

$$H_m = \sum_1^m \beta^{i-1} x_i + \lambda \sum_{m+1}^{\infty} \beta^{i-1} \quad (4)$$

$$\forall(\beta) = \underset{L \geq 1}{SUP} \frac{E \left(\sum_1^L \beta^{i-1} x_i \right)}{E \left(\sum_1^L \beta^{i-1} \right)} \quad (5)$$

Therefore, the second arm (since arms have the same behavior, we consider arms to be the beams of the UAV) is optimal initially in case, $\lambda \geq \forall(\beta)$ let us prove the use first theorem and (4), it may perhaps contain the courtesy to pronouncement rules M specified by a discontinuing

time L which characterizes the previous interval that first arm is recycled (5). The payoff's exhausting L tends to be

$$E\left(\sum_1^L \beta^{i-1} x_i + \lambda \sum_{L+1}^{\infty} \beta^{i+1}\right)$$

Where we considered $L=0$ producing $\lambda/(1-\beta)$. This implies that the second arm is optimal firstly in case all stopping rules are taken to be $L \geq 1$.

$$E\left(\sum_1^L \beta^{i-1} x_i + \lambda \sum_{L+1}^{\infty} \beta^{i-1}\right) \leq \lambda/(1-\beta)$$

$$E\left(\sum_1^L \beta^{i-1} x_i \leq \lambda E \sum_{L+1}^{\infty} \beta^{i-1}\right); E\left(\sum_1^L \beta^{i-1} x_i / E \sum_{L+1}^{\infty} \beta^{i-1}\right) \leq \lambda$$

This is equivalent to $\forall(\beta) \leq \lambda$. The value $\forall(\beta)$ contingent only on β and on the distribution of the returns from first arms x_1, x_2, \dots , everywhere, the situation signifies the indifference fact: the rate λ for the second arm in the solitary of MAB that indifferent amongst preparatory off's on the first arm and indicating the second arm wholly the segments. Let us yield to the i -armed outlaw by arithmetical reduction and self-determining arms obligating revenues symbolized by s

First arm $x(1,1) \dots x(1,2), \dots, \dots, \dots$, (6)

Second arm $x(2,1), \dots, s x(2,2), \dots, \dots$ (7)

Arm i , $x(i,1), \dots, x(i,2), \dots (i,3), \dots, \dots$

Suppose, the variables are reliant on flanked by commotions and that the first complete instants exist and are consistently constrained, $\sup_{i \geq 1, t \geq 1} E|x(i, t)| \leq \infty$ the deduction is β , where $0 \leq \beta \leq 1$ so also considering (6) and (7), let us pursue a decision rule $\varphi = \theta_1, \theta_2, \theta_3, \dots$. Therefore, to maximize the total discounted return it will be given by

$$\varphi(M) = E\left(\sum_{t=1}^{\infty} \beta^{t-1} 0_t | M\right) \quad (8)$$

For every arm (beam), computation of return ought to be articulated as

$$\forall_i = \sup_{L \geq 1} E \sum_{t=1}^L x(i, t) / E \sum_{t=1}^L \beta^{t-1}$$

$i = 0, 1, 2, \dots$. Here, we suppress β in the notation for $\forall \beta$ nice going to behold constant throughout.

Firstly (8), we proofed firstly the special case that around objectively two arms ($i = 2$) besides wherever altogether the arbitrary variables are decadent. We designate the outlays after the first arm to be $x(1), x(2), \dots$, the second arm to be $z(1), z(2), \dots$ among others. The stated two arms are bounded sequences of real numbers. For x , it will be expressed as

$$\forall_x = \sup_{i \geq 1} \sum_1^i \beta^{t-1} x(t) / \sum_1^i \beta^{t-1} \text{ and } z$$

$$\forall_z = \sup_{i \geq 1} \sum_1^i \beta^{t-1} z(t) / \sum_1^i \beta^{t-1}$$

Subsequently $x(t)$ assumed bounded, the series of $\sum_1^m \beta^{t-1} x(t)$ joins, consequently that there exists a value of i , possibly ∞ at which the supremum in the definition of \forall_x is taken on as well us \forall_z . Suppose that j is this value of i so that $1 \leq j \leq \infty$ considering this first lemma that the sequence of (6) is non-random and bounded. In case

$$\forall_x = \sum_1^j \beta^{t-1} x(t) / \sum_1^j \beta^{t-1}, \text{ then for all } i \leq j$$

$$\sum_{t=i}^j \beta^{t-1} x(t) \geq \forall_x \sum_{t=i}^j \beta^{t-1} \quad (8)$$

And for j finite and $i < j$, leading to

$$\sum_{t=j+1}^i \beta^{t-1} x(t) \leq \forall_x \sum_{t=j+1}^i \beta^{t-1} \quad (9)$$

At this stage, we can now proof both (8) and (9) as

$$\sum_{t=i}^j \beta^{t-1} x(t) \geq \forall_x \sum_{t=i}^j \beta^{t-1}, \text{ and } \sum_{t=j+1}^i \beta^{t-1} x(t) \leq \forall_x \sum_{t=j+1}^i \beta^{t-1}$$

Subtracting the latter from the former contributes to (8) when i is less than or equal to j and gives (9) when this leads the equations to be simulated.

Let us use a Bernoulli MAB approach demonstrated as a point to (\forall, ψ) given that a given medical device M obtains a reward probability $\{\theta_1, \dots, \theta_M\}$. At respectively interval phase t , we consider an accomplishment of a unique slit considered medical mechanism and obtain a remuneration of r . notably, \forall are regularity of movements, independently mentioning to the interaction with one slot medical machine. The percentage of achievement is the predictable return, $Q(a) = E[r | a] = \theta$.

In case the achievement a_t at the interval phase t is on the i^{th} medical machine, then $Q(a_t) = \theta_i$. the ψ is a reward function. In the situation of Bernoulli MAB, during the study, distinguish a reward r in a *stochastic* approach was considered as well. At the stretch footstep t $r_t = \psi(a_t)$, (at) may return reward 1 with a probability of 0 or other with it will be given by $Q(a_t)$. The aim here grows into to

exploit the cumulative rewards computed as $\sum_{t=1}^T r_t$.

Uncertainly the study demonstrated that it distinguishes the optimum accomplishment with the superlative reward, and then the aim is identical to minimalize the prospective regret without picking the optimal action.

This mainly means that the optimum reward probability θ^* of the optimal action a^* will be depicted by

$$\theta^* = Q(a^*) = \max_{a \in \mathcal{V}} Q(a) = \max_{1 \leq i \leq M} \theta_i$$

The defeat utility ξ is provided by the entire regrets that cannot be selected as the finest action up to the time step T





$$\xi = E \left[\sum_{t=1}^T (\theta^* - Q(a_t)) \right]$$

Another approach besides bandit, the ϵ -greedy algorithms exists to yield the superlative accomplishment utmost of the stretch, nonetheless guarantees unsystematic consideration infrequently. The accomplishment connotation is expectably bequeathing to the involvement by averaging the rewards interrelated to the objective achievement a experimental to the current t . Given that 1 is a binary pointer utility and $N_t(a)$ is the stretch of the stroke a .

$$\hat{Q}_t(a) = \frac{1}{Z_t(a)} \sum_{\tau=1}^t r_\tau \mathbb{1}[a_\tau = a] = Z_t(a) = \sum_{\tau=1}^t r_\tau \mathbb{1}[a_\tau = a]$$

The ϵ -greedy algorithm stress $\hat{a}_t^* = \operatorname{argmax} a \in \mathcal{V} \hat{Q}_t(a)$ es that means that a small probability ϵ considers a random action.

Table 1: Margin specifications

Notations	Specification
Beta-Parameters ($\alpha = 1, \beta = 1$), α β	Expected Reward Probability to be 50% successful beam Unsuccessful beam
Reward probabilities	{0.0, 0.1, 0.2, ..., 0.9}
τ	Extensive negligible inputs
Device -system bandwidth	1 GHz
M	10 slot medical machines
$\alpha = 1000$	Expected reward probability is 10%
$\beta = 9000$	
solver Execution Times	10000 steps
	ϵ -greedy algorithm
	UCB1 algorithm
	Thompson sampling
	Bayesian UCB Algorithm
1 st and 2 nd solver Times	500 steps
Simulator tool	MATLAB

Bayesian UCB / UCB1 algorithm tends to have a different approach to the ϵ -greedy algorithm, assumes any prior on the reward circulation, and consequently depend on the variation of Hoeffding for an actual oversimplify guesstimate. Thompson sampling partakes an unpretentious awareness nevertheless then all excessive for responding to the MAB problem of devices where we select action an affording to the possibility that is optimal. Considering $\pi(a | ht)$ being the prospect of enchanting achievement a particular the antiquity ht :

$$\pi(a | ht) = \mathbb{1}[Q(a) > Q(a'), \forall a' \neq a | ht]$$

$$= \mathbb{1} R | ht \mathbb{1}[a = \operatorname{argmax}_{a \in \mathcal{V}} Q(a)]$$

Importantly, it is also regular to accept that $Q(a)$ follows the distribution of Beta in For the Bernoulli bandit, as $Q(a)$ is fundamentally the achievement opportunity θ in Bernoulli. It's important to note that all bandits have operation besides the assumption of the behaviors like Bernoulli, multi-armed among others.

A naive method can be used is to continue playing with one alternative for numerous rounds to ultimately approximate the "accurate" recompense chance rendering to the common edict of huge records in computations. Nevertheless, this is moderately extravagant, and confidently does not assurance the superlative lasting recompense as anticipated.

At every time t , we model an expected reward, $\mathcal{Q}(a)$ since the prior distribution Beta (α_i, β_i) for each exploit. The superlative accomplishment is nominated among trials after the true recompense is experimental, informs the Beta dissemination consequently, which is fundamentally doing Bayesian implication to calculate the subsequent with the identified previous and the probability of attainment the experimented information as $aTSt = \operatorname{argmax} a \in \mathcal{V} \mathcal{Q}(a)$ and thus

$$\alpha_i \leftarrow \alpha_i + r_t \mathbb{1}[aTSt = a_i]$$

$$\beta_i \leftarrow \beta_i + (1 - r_t) \mathbb{1}[aTSt = a_i]$$

To diminish the brink p in time, consider extra assertive destined appraisal with additional rewards experiential. Customary of $p = t^{-4}$. This assumption and innovation are done by the UCB1 given u as the UCB, $u = U_t(a)$

$$U_t(a) = \sqrt{\frac{2 \log t}{Z_t(a)}}, a_t^{UCB1} = \operatorname{argmax}_{a \in \mathcal{V}} Q(a) + \sqrt{\frac{2 \log t}{Z_t(a)}}$$

This depicts that in Upper Confidence Bound algorithm; continuously inferior the avaricious achievement to exhaust the potentials the UCB given by

$$a_t^{UCB} = \operatorname{argmax}_{a \in A} \hat{Q}_t(a) + \hat{U}_t(a)$$

4- Simulation Results and Discussion

Within this section, the demonstrations of results and conversation of the conveyed simulation results in figure 2a, figure 2b, figure 3a, figure 3b, figure 4 and, figure 5 with the exhaustive explanation of the algorithm presented.

4-1- Model Description

The mUBS considerably use a predictable stand β of B is denoted as $|\mathbf{D}|$ distinctive, not non-orthogonal beams. The study adopted that the mUBS can solitary inferior a subsection of bm concurrently given that $bm \in \mathbf{N} \text{ } bm \leq B$, is considered to be immovable amounts. Some of the limitations identified on mmWave channel sparsity. The main reason for the mUBS is to choose a subsection of bm that will be maximizing the number of records efficaciously traditional by the imminent CR in the reportage capacity. The study further assumed that mUBS is not certain or no nothing about then environment. In situation, the simplicity of the system execution reduces as the operative necessities nothing to be configured like at each mUBS according to the environment. Therefore, the mUBS has to learn as the situation changes to select the subset of the beams. In this way, the UAV will be to account for every approaching CR in context to beams it emits. The model similarly takes in mind a discrete-interval situation, given that the mUBS modernizes it is beam-ray miscellany in the regular time setup/period in each a setup $t=0,1,\dots,T$, given that $T \in \mathbf{N}$ considered to be a finite horizon, the following three activities are applied.

- i. Given that set $W_t = \{CRu,i\} i=1,\dots,\vartheta_t$ of $CRt = |Nt|$ CR resembling buses, different IoTs mong others to the mUBS. The number of CRt of the CRt of the CRs fulfills the condition that $CRT \leq Vmax$, considering the $Vmax \in \mathbf{N}$ will be the maximum number of the supported CRs contained by the analysis capacity. At the time of the cataloging, mUBS will be having the capacity to receive info about the context it,i of every imminent automobile CRu,i will be a three-dimensional flight path taken from the bounded context space or coverage area $K = [0,1,2]^k$.
- ii. The mUBS chooses a detachment of bm beams. The study similarly denotes that the regular of selected beams in the period t by $St = \{st, j\} j=1,\dots, bm \subseteq \beta$. Before the CR in the W_t will be cognizant approximately the nominated beams complete CR interface.

- iii. At the time, when the CRu will be within the range of mUBS coverage area according to google map and mUBS will be in a position to transmit data to any CR within the coverage area. Observation will be considered on the amount of data $A_{od,j}(xt, i,t)$ CR CRu will be productively be received through selected beams $A_{od,j}, j=1,\dots, bm$, till the expiration of t.

Considered the denotation of the random variable $rb(x)$ the beam enactment of the b under the perspective of the x. It will be meaning that the extent of data $rb(x)$ a CR with the perspective $x \in X$ self-control be reception from the mUBS using the $b \in \beta$. We adopt that this indiscriminate adaptable is circumscribed based on $[0,1, M_{Aod}]$, given M_{Aod} will be the determined aggregate of the data that expected by CR. M_{Aod} will be bounded by the determined frequency of the broadcast channel. We denoted the estimated value of the beam presentation of beam b in the context x with $\mu_b(x)$. The mUBS ambitions at choosing a subdivision of the bm which will exploit the anticipated obtain data at the CR. Therefore the optimal subdivision in interval

$$t \nabla_t^*(K_t) = \{ \nabla_{t,j}^*(X_t) \} j=1,\dots, bm \subseteq \beta.$$

Therefore the set $\nabla_t^*(K_t)$ will be depending on $X_t = \{xt, i\} i=1,\dots, Vt$ and bm satisfy.

$$\nabla_t^*(K_t) \in b \in \beta \setminus \left(\bigcup_{k=1}^{j=1} \{ \nabla_{t,k}^*(K_t) \} \right) \sum_{i=1}^{\vartheta_t} \mu_b(kt, i)$$

Noting that $j=1,\dots, bm$. In case the mUBS will be knowing the expected beam performance $\mu_b(x)$. For every CR perspective $x \in K$ and respectively $bm b \in \beta$, it will be selected optimum subcategory of the beam for every set of pending CR according to (1). To obtain an expected amount the data will be received over the sequence from 1 to the time T.

$$\sum_{t=1}^T \sum_{i=1}^{\vartheta_t} \sum_{j=1}^{bm} E \left[r \nabla_{t,j(x)}^*(xt, t) \right] = \sum_{t=1}^T \sum_{i=1}^{\vartheta_t} \sum_{j=1}^{bm} \mu \nabla_{t,j(x)}^*(xt, i) \quad (10)$$

The mUBS does not recognize the coverage area, it will be learning the expected performance $\mu_b(x)$. Over a given period (10).

To absorb these concerns, the mUBS has to attempt out diverse bm of miscellaneous CR context terminated interval also ensuring that the beams will be proved in being good. Lastly, the learning algorithm will be done some times for CR in the coverage area in the context of X_t , selecting the St of bm .

The algorithm I Pseudocode of Proposed Reinforcement algorithm

1. Input: $T, PT, K(t)$
2. Prime perspective divider: Create divider P_T of perspective astronomical $[0, 1, 2]^k$ into $(P_T)^x$ hyper-cubes of duplicate amounts
3. Prime stands: Aimed by entirely $b \in B$ and entirely the $h \in P_T$ stand $N_{b,h} = 0$
4. Prime guesses: Aimed at totally $b \in B$ and wholly $h \in P_T$, set $\mu'_{b,h} = 0$
5. for each $t = 1, \dots, K$ do
 6. Perceive car settings $K_t = \{k_i, i\} i = 1, \dots, W_t$
 7. Discovery $H_t = \{h_i, i\} i = 1, \dots, W_t$ in that $k_{t,i} \in p_{t,i} \in P_T, i = 0, 1, 2, \dots, W_t$
 8. Calculate the usual of under-explored grins $B_{H_t}^{ue}(t)$ trendy (5)
 9. if $B_{H_t}^{ue}(t) \neq 0$ -formerly Assessment
 10. $u = acreage(B_{H_t}^{ue}(t))$
 11. if $u \geq m$ then
 12. choice $s_t, 1, \dots, s_{t,m}$ randomly from $B_{H_t}^{ue}(t)$
 13. else
 14. choice $s_t, 0, 1, 2, \dots, s_{t,u}$ equally u beams from $B_{H_t}^{ue}(t)$
 15. choice $s_{t,u} + 1, \dots, s_{t,m}$ equally the $(m, -u)$ bm from (6)

$$b'_1 H_t(t), \dots, b_{m-u} H_t(t)$$
 16. else if
 17. else
 18. select $s_t, 1, \dots, s_{t,m}$ for instance the m bm

$$b'_1 H_t(t), \dots, b_{m-u} H_t(t)$$
 as a (7)
 19. end if
 20. Detect the acknowledged date $r_{j,i}$ of every car $W_{t,i} i = 1, \dots, W_t$ in every bm $s_{t,j}, j = 0, 1, 2, \dots, m$
 21. for $i = 1, \dots, W_t$ do
 22. for $j = 1, 2, \dots, m$ do

$$\mu'_{s_{t,j}, h_{t,i}} = \frac{\mu'_{s_{t,j}, h_{t,i}} + r_{j,i}}{N_{s_{t,j}, h_{t,i}} + 1} \text{ and } N_{s_{t,j}, h_{t,i}} = N_{s_{t,j}, h_{t,i}} + 1$$
 23.
$$\mu'_{s_{t,j}, h_{t,i}} = \frac{\mu'_{s_{t,j}, h_{t,i}} + r_{j,i}}{N_{s_{t,j}, h_{t,i}} + 1} \text{ and } N_{s_{t,j}, h_{t,i}} = N_{s_{t,j}, h_{t,i}} + 1$$
 24. end for
 25. end for
 26. end for

The variety of erudition motivation is depending on the history of the beams to be selected. The predictable magnitude of data expected by the vehicle will be certain as follows in case we consider the selection $St, t=1, \dots, T$ of the algorithm.

$$\sum_{t=1}^T \sum_{i=1}^{\vartheta_t} \sum_{j=1}^{bm} E[r_{st,j}(xt, i)]$$

$$\sum_{t=1}^T \sum_{i=1}^{\vartheta_t} \sum_{j=1}^{bm} E[\mu_{st,j(kt)}(kt, i) - r_{st,j}(kt, i)] \quad (11)$$

Therefore, the anticipated metamorphosis in the aggregate of acknowledged data accomplished and an algorithm will be the "regrets of learning" taken to be R considering both (10) and (11).

$$R(T) = E \left[\sum_{t=1}^T \sum_{i=1}^{\vartheta_t} \sum_{j=1}^{bm} (r_{t,j(kt)}^* - r_{st,j(kt)}) \right]$$

$$R(T) = \left[\sum_{t=1}^T \sum_{i=1}^{\vartheta_t} \sum_{j=1}^{bm} (\mu_{t,j(kt)}^* - E[r_{st,j(kt)}]) \right] \quad (12)$$

4-2- Learning Technique

The prototypical bm assortment in an mUBS by way of a fast 3-dimension semi-online learning problem as depicted in the algorithm below since it allows identification of the best beams independently in a given interval although secretarial for energetic traffic and setting vicissitudes exhausting a condition in figure 2.

Hence, the mUBS necessities to recognize the superlative rays by prudently choosing subdivisions of bm over time. These tactic cataracts below the grouping of appropriate MAB problems. These difficulties moreover contain side evidence that disturbs the recompenses of the activities. A contextual MAB method escalates the mUBS doesn't objective absorb or choice which bm is unbeatable on steady, then over as an alternative it achievements extra info almost impending CR to ascertain that bms are the superlative under a prearranged traffics as the algorithm explains called the algorithm I Pseudocode of Proposed Reinforcement algorithm.

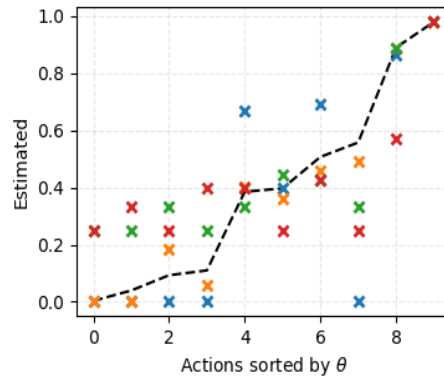


Fig. 2(a) Reward probability against the different estimated probability.

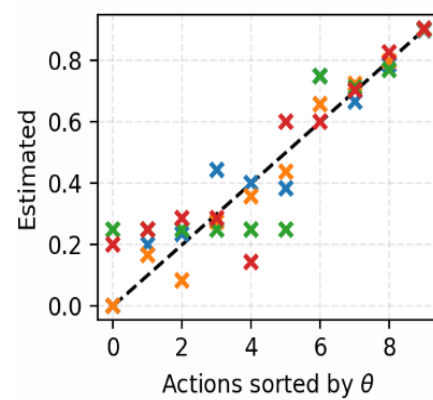


Fig. 2(b) Reward probability against the different estimated probability.

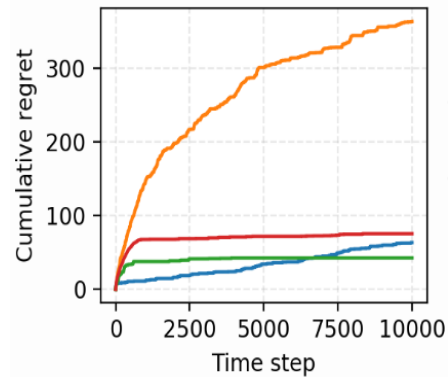


Fig. 3(a) Time steps against the dissimilar cumulative regrets.

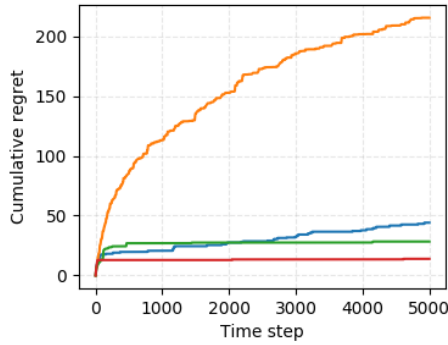


Fig.3(b) Time steps against the dissimilar cumulative regrets.

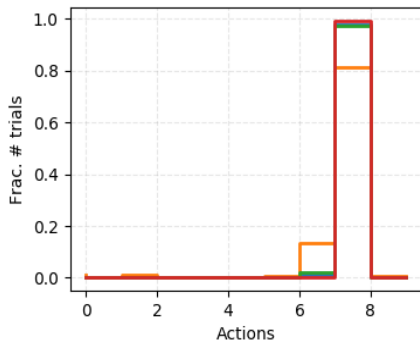


Fig. 4 A fraction Action is selected during the 5000 Execution.

4-3- Benchmark Algorithms and Metrics

In this sub section, it depicts the provision an exhaustive performance assessment by associating the proposed 3D-FML to numerous other algorithms. The succeeding particularizes on every yardstick (i.e., Optimal, 2-diamentional machine learning algorithm (2DFML), Dimensions typically measured are quality, time and cost. Where in this perspective, it shows the parameters that have been compared with) scheme:

- The Optimal

This category of the algorithm has a priori information about the predictable bm to be assessed $\mu_b(x)$ of every bm $b \in \beta$ in every circumstance $x \in X$ in addition henceforward contributes an upper bound to the additional procedures that satisfies all the properties in (1).

- The Upper Confidence Bound

This is an irregular of the conventional learning procedure [29], which it is applied and adjusted to the standard of the use-case. The algorithm learns from earlier perceived bm assessment, nevertheless deprived of enchanting keen on account context data. In every time, we chose to adopt UCB among other bandits chooses bm beams with the uppermost predictable UCB on their anticipated bm performance.

- The Random

Within this kind of the algorithm, it chooses random bm in separately period randomly. This algorithm may possibly explore every bm at least once or not. Formerly, the algorithm twigs over the bm with the utmost of the data received depending on the desire of the communication.

- The 2D-FML

This is another algorithm that has been currently availed to arena, where it uses reinforcement learning approach where the vehicles are known when they reach the coverage area of the base station [28].

4-4- Geometric Assessment

Firstly, an evaluation of a generic scenario is presented. Secondly, scrutinize the influence of numerous constraints, that is to say, the number of bms selected, the incidence of blockages, the arrival rate of the vehicles, and the underlying traffic patterns. Least otherwise stated, a consideration of the case where for instance as further channel is classified in table 2:

- The percentages at which the permanent blockages and temporary blockages block every one relates to at least the 20 % of all routes.
- The pattern of the traffics (for instance, the automobile (car, vehicle) on the arrival rate and the path chances) alternate per the distinctive traffic patterns demonstrated the by map presented.
- Subsequently Map simply delivers the distinctive everyday pattern of the traffic for 0.71 day (beginning at 06:00 to exactly 22:00); the algorithms were executed over a 0.71 day. Each execution was frequent over 22 times so that the presentation of results demonstrated 94% assurance intervals within the figures.

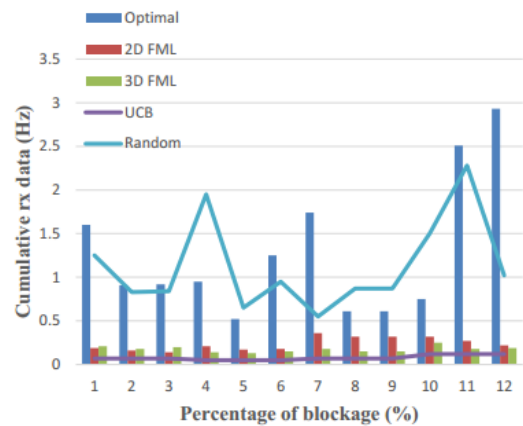


Fig. 5 The Impact of blockages on the accumulative data received for the arrival rate of specified λ and time (t).

4-5- The Impact of Blockages

Fig. 5 confirms that an increase in the received data given that predefined arrival percentage of λ in the instance of any given rate m nominated bms per interval like for instance 10%, 30%, 60%, 80%, 90% of the permanent obstructions in the structure. Evidently, as the percentages of the permanent obstructions in the system do increase, as the accumulative of data received decrease. On behalf of slightly percentages of the permanent blockage, the proposed 3D-FML overtakes wholly non-optimal algorithms. The accumulative of the data received attained by 3D-FML deceits amongst 16.75%, and 18.32% complex than that succeeded by the next-best algorithm upper confidence bound. Furthermore, the 3D-FML's algorithm accomplished results diverge from that of an optimal simply by at utmost 3.76%.

4-6- The Live daily traffic patterns

Traffics has different patterns per different users, The daily traffics is occasionally a inaccuracy, as the peak retro repeatedly continues additional sixty minutes and the "rush hour" denotes to the measurements of traffics, not the promptness of its stream. A rush-hour maybe 6:00pm to 10:00 am and 4:00pm to 8:00 pm depending on the set up of the city. Traffic periods could fluctuate from city to city, from area to area, district to district, and seasonally. Application developers availed applications to alert the user on the proceeding pattern depending on the place, Waze.

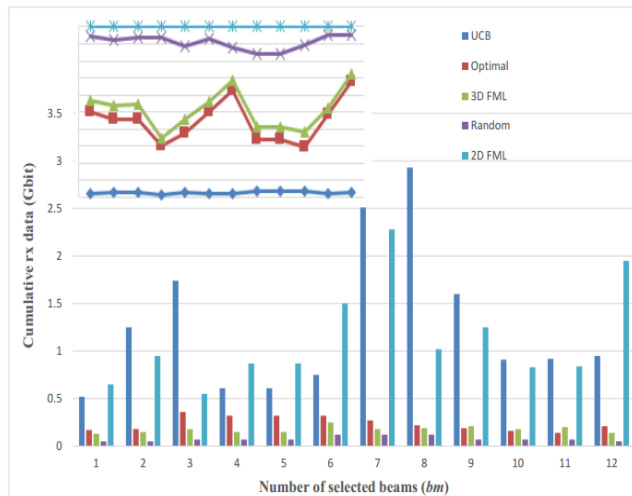


Fig. 6 Impact on the received bm data for arrival rate on cumulative received specified data.

The most well-known traffic app is free and perhaps its biggest draw is the real-time traffic information provided by users includes Google Maps, INRIX, MapQuest, Apple Maps, Traffic Spotter, USA Traffic Cameras, TomTom GPS Navigation Traffic among others. The national highway traffic safety administration recorded that utmost misfortunes transpire throughout "rush hour," altered period of time. And according to the national highway traffic safety administration, the Saturday is the supreme treacherous day of the week to drive, predominantly since there are more compartments and more drunk drivers on the highway than any additional day.

4-7- The Impact of the numeral Selected Beams

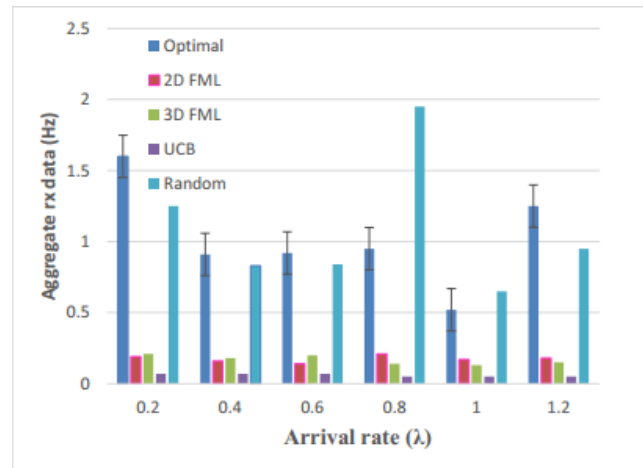


Fig.7 Impact on the arrival rate λ of data based on the cumulative received data.

Analysis of the influence of the amount of beams selection bm per retro on the accumulative of the data received. Fig. 3 displays the accumulative of the data received accomplished with an arrival rate λ for diverse $bm \in \{0,1,2,4,8,9,\dots\}$. As the quantity of concurrently beam increases selected, the accumulative of the data received increases. Nonetheless, the greater the quantity of bms , the greater is the hardware hurdle and energy consumptions at the mUBS. For diverse value of bm , the accumulative of the data received achieved by 3D-fast machine leaning is amid 11.75% and 20.88% upper than that attained by the next-best algorithms like upper confidence bound and merely up to 5.76% is the lower than that attained by an Optimal as figure 7 depicts.

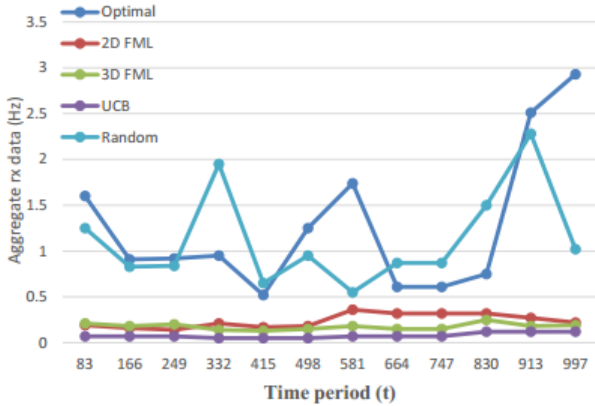


Fig. 8 The aggregates data received for arrival ratio with λ , and the β .

4-8- The Impact of arrival rate

The impact of arrival rate depends on some considerable factors, labor-intensive formation and struggle- dynamic assessments to necessitate abundant more than 1.16667 hours. Additionally, combat-driving examinations might only internment possessions of permanent blockages, nonetheless not from the temporary obstacles. The subplot in Fig. 4 depicts the ordinary of the expected data (rx) through arrival rate λ . Regular data (rx) tends to the data terminated altogether the automobiles in the systems up to this historical. These figures demonstrate fast machine learning's rapid learning and adaptation capabilities. Precisely, the three dimensional fast machine learning completes over 89% of performance of optimum in thirty nine times. This illustrates how quick 3D- fast machine learning congregates to near-optimal of the beam selections.

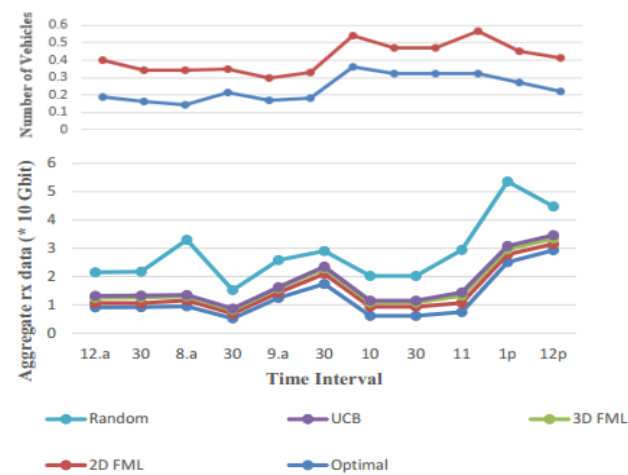


Fig. 9 Data for 172800 minutes of a-live day-to-day Traffic pattern.

4-9- The Average Received data

The variations in the figure results from the numeral of vehicles in the systems. Explicitly, the aggregated of the data received rises with the quantity of automobiles, the impacts of automobile arrival rates and traffics. As predictable, an optimal elasticities of the additional upper bounds because of a priori consociate of the anticipated bm performance. The proposed 3D-FML algorithm outclassed the additional algorithms as identified for example upper confidence bound, and the random. Observation achieved comes out to be that Fast Machine Learning's assessments rapidly tactics that of an optimum contained by the first hundred periods although the additional algorithms accomplish at least 22% eviler than 3D-fast machine learning. Throughout the comparisons, mean absolute deviation for all the compared algorithms and the proposed is in precise demonstrated in the table 3.

Table 2: Margin specifications

Parameters	Values and Notations
Carrier Wave	28 Gigahertz
Scheme Bandwidth	1 Gigahertz
Transmit Power	30 dBm
Path loss Model (dB)	$32.4 + 17.31 \log_{10} d(m) + 20 \log_{10} (f_c (GHz)) + \xi$
Noise figure	4 dB at mmBS; 7 dB vehicle
UAV's beam width	36°
Simulator tool	MATLAB
Thermo Noise	$-174 dBm / Hz$

The result illustrated in this paper was contained using the parameters put in table 2 compared to models theorized [28]; the exploration is rummage-sale to acclimatize to structural subtleties, for instance, the arrival of obstructions and vicissitudes in traffic configurations. The algorithm recognizes obstructions by estimating the cumulative acknowledged data of each automobile for the individually designated bm where the current models have not that trait. Furthermore, the algorithm acclimates the traffic's arrays through learning affiliation amongst the path of advent and, the acknowledged dAsadiata. Results, it chooses the beams, which exploit the inclusive system aptitude. Consequently, it affords additional extra to the infrastructures with sophisticated traffic and henceforward, attends an outstanding numeral of UAV.

Table 3: Mean Absolute Deviation results

Parameters	Mean Absolute Deviation					
	1	2	3	4	5	6
Optimal	1.60	0.91	0.92	0.95	0.52	1.25
3D FML	0.19	0.16	0.14	0.21	0.17	0.18
2D FML	0.21	0.18	0.20	0.14	0.13	0.15
UCB	0.07	0.07	0.07	0.05	0.05	0.05
Random	1.25	0.83	0.84	1.95	0.65	0.95
Parameters Extension	7	8	9	10	11	12
Optimal	1.74	0.61	0.61	0.75	2.51	2.93
3D FML	0.36	0.32	0.32	0.32	0.27	0.22
2D FML	0.18	0.15	0.15	0.25	0.18	0.19
UCB	0.07	0.07	0.07	0.12	0.12	0.12
Random	0.55	0.87	0.87	1.50	2.28	1.02

The focal awareness linked within figure 3 shows that it's conceivable to putrefy a multifarious supervisory of MAB problem into a categorization of uncomplicated pronouncements, where each verdict of the sequence is solved utilizing the MAB. The mean of the first arm can be obtained in various conducts including ranked bandit approach retains the agreeable traits of preparatory the attention by a sample of the intergalactic and then concentrating gradually on the most propitious capacity, at the diverse weighing machine, permitting to the assessments eventually undertaking an indigenous pursuit about the global optimal functions.

The Thompson sampling riggings the impression of probability matching this is due to the reward approximations $\hat{Q}(a)$ are appraised from subsequent deliveries, any of these prospects is corresponding to the possibility that the correspondent action is optimum, comfortable on pragmatic *hty*. It's observed that exploration is needed since information is valuable, this means that no exploration is much effective using greedy algorithms, random exploration on the ϵ -greedy algorithm, and current exploration is seen in Upper confidence bounds and Thompson illustrated in figure 2(a) and 2(b).

A multi-armed bandit approach to wireless communication problems can be a sign of an efficient optimization method compared to traditional statistics as a state vogueish figure 3(a), besides vogueish figure 3(b).

4-10- Heuristic approaches in instigating MAB.

Other concepts like heuristic approach in instigating MAB tryouts are approachable to countenance stretchy

enticement circulations that may knob the varieties of concerns that ascend intangible claims, the principles of the preconception's apparels are known nevertheless the identities of special cases of reward distributions.

The contributions of the utility optimization are a period of categorized enthusiastic processes anticipated for broad-spectrum exploration creations with diverse algorithmic instantiations contingent on whether the estimates are noisy, the enactment of the algorithms is contingent on the "local" behavior of the function around its global optimal communicated in terms of the quantity of near-optimal states phlegmatic with some geometric as presented in figure 4. In case the indigenous unevenness of the utility is recognized formerly solitary can enterprise identical effective optimizations.

The stochastic MAB problem remains a significant typical for studying the regrets, exploration, and exploitation adjustment in reinforcement learning as well as observed figure 5. While numerous procedures for the problem are well-understood hypothetically, empirical authorization of their efficiency is generally limited.

4-11- The several issues in wireless transportations

The model the problem as that MAB delinquent where numerous issues in wireless transportations obligate been treated by the use of multi-armed bandits, a decision-making individual has to choose a subsection of schedules of unidentified recompenses to activity the recompense terminated time. Importantly a MAB method is appropriate for our problematics since an umBS outflow an inadequate customary of beams instantaneously.

The Algorithms matches the issues of mmWave UAV communication on numerous facades are critically identified for example (a) the model distinguishes perpetual structures like edifices, and recurrently congested extents owing to transitory obstructions for example base stations or structure situates haunted by huge automobiles exhausting operational knowledge; (b) the model regulate traffic designs to have structure aptitude maximization by providing grander coverage (like in this case distribution of additional *bms*) in situations with heftier traffics.

4-12- The significance of Mobile Base station

This is significant since wave BS may communicate instantaneously terminated an inadequate number of *bm*. This restraint is subject to the hardware traits, the millimeter Wave frequency sparsity, and the beam form practice lastly (b) on the list it conjectures traffics from the perspective (like the automobile's path of the entrance) and chooses the superlative beams. The mainstreams of boulevards have dissimilar road traffic arrays unbiased by

the interval of the diurnal. Although inferring these arrangements are obtainable of the latitude of this paper, the model strategies algorithms to recognize and absorb from such configurations.

4-13- The effect of user speed vehicle

On the focus on animate everyday traffic configuration: The higher the speed of vehicle as per the UCB, the lower the computation rate, where the higher the speed of vehicle, the lower optimum computation done per the random and optimum estimation per the emblematic circulation arrangement. The 2D-FML algorithm slightly performs close to the 3D-FML model presented due to the fact that 3D models perform better than the 2D the figures of demonstration illustrated since in areas of hard to reach increases connectivity, and mobility conditions. Because of a regular of consequence, Googles' distinctive traffics do not confiscation the instantaneous deviations in overcrowding designs that are perceptible in conscious traffic tale. Therefore, the model depicted the tentative sentient traffic accounts for Google's positioning within a period of two days and half an hour.

5- Conclusions

This paper presented a first semi-online algorithm that selects the best beams emitted from the UAV since there are gaining popularity nowadays in social networking and service delivery like now commercial UAVs are becoming smarter. The paper stresses that the applicability of 5G research is in its initial stages whereby few countries have embraced it due to international technological conflicts. The paper clearly demonstrated that it is applicable to use UAVs as a base station in hard to reach areas by the use of reinforcement learning as per the details.

Furthermore, the paper depicted that UAV's advancement will lead to increased illegal surveillance easier, fly into private property, and get video, photos, or both that are not possible for camera-wielding beings that cannot mainly be controlled on the web. The papers did not consider an international policy on the height at which UAV can fly, never consider any category of UAVs. Also, it used a geometric dataset that is not of real-time. In the future, the proposed model is to be subjected to other algorithms besides those ones that have been used in perspective to beam orthogonality, location recording, and situation of co-located automobiles, interference and amount of selected beams.

References

- [1] R. Torkamani and R. A. Sadeghzadeh, "Wavelet-based Bayesian Algorithm for Distributed Compressed Sensing.", *Journal of Information Systems and Telecommunication*, Vol. 7, No. 2, April-June 2019.
- [2] M. Jaderyan and H. Khotanlou, "SGF (Semantic Graphs Fusion): A Knowledge-based Representation of Textual Resources for Text Mining Applications." , *Jour. of Information Systems and Telecommunication*, Vol. 7, No. 1, January-March 2019.
- [3] N. Sharma, A. S. Arora, A. P. Singh, and J. Singh, "The Role of Infrared Thermal Imaging in Road Patrolling Using Unmanned Aerial Vehicles," in *Unmanned Aerial Vehicle: Applications in Agriculture and Environment*, Springer, 2020, pp. 143–157.
- [4] C. Qu, W. Gai, M. Zhong, and J. Zhang, "A novel reinforcement learning-based grey wolf optimizer algorithm for unmanned aerial vehicles (UAVs) path planning," *Appl. Soft Comput.*, p. 106099, 2020.
- [5] F. Al-Turjman, H. Zahmatkesh, and R. Daboul, "Optimized Unmanned Aerial Vehicles Deployment for Static and Mobile Targets' Monitoring," *Comput. Commun.*, vol. 149, pp. 27–35, 2020.
- [6] Y.-F. Liu, X. Nie, J.-S. Fan, and X.-G. Liu, "Image-based crack assessment of bridge piers using unmanned aerial vehicles and three-dimensional scene reconstruction," *Comput.-Aided Civ. Infrastruct. Eng.*, 2020
- [7] S. Lee et al., "Intelligent traffic control for autonomous vehicle systems based on machine learning," *Expert Syst. Appl.*, vol. 144, p. 113074, 2020.
- [8] E. dos Santos Moreira, R. M. P. Vanni, D. L. Função, and C. A. C. Marcondes, "A Context-Aware Commu. Link for Unmanned Aerial Vehicles," in *2010 Sixth Advanced International Conference on Telecommunications*, 2010, pp. 497–502.
- [9] D. He, S. Chan, and M. Guizani, "Communication security of unmanned aerial vehicles," *IEEE Wirel. Commun.*, vol. 24, no. 4, pp. 134–139, 2016.
- [10] A. Abdessameud and A. Tayebi, "Formation control of VTOL unmanned aerial vehicles with communication delays," *Automatica*, vol. 4, no. 11, pp. 2383–2394, 2011.
- [11] W. Gaofeng, G. Xiaoguang, Z. Kun, and F. Xiaowei, "Multi-objective Placement of Unmanned Aerial Vehicles as Communication Relays Based on Clustering Method," in *2019 Chinese Control And Decision Conference (CCDC)*, 2019, pp. 1462–1467.
- [12] H. Nawaz, H. M. Ali, and M. H. Mahar, "Swarm of Unmanned Aerial Vehicles Communication Using 802.11 g and 802.11 n," *IJCSNS*, vol. 19, no. 4, p. 289, 2019.
- [13] G. S. Voronkov, E. A. Smirnova, and I. V. Kuznetsov, "The method for synthesis of the coordinated group DPCM codec for unmanned aerial vehicles communication systems," in *2019 International Conference on Electrotechnical Complexes and Systems (ICOECS)*, 2019, pp. 1–4.
- [14] H. Harmon, "The Use of Unmanned Aerial Vehicles in Public Safety Communication and Optimization of Coverage (Case Study-Cook Islands)," PhD Thesis, Auckland University of Technology, 2017
- [15] S. Fekri-Ershad and F. Tajeripour, "Multi-resolution and noise-resistant surface defect detection approach using new

- version of local binary patterns,” *Appl. Artif. Intell.*, vol. 31, no. 5–6, pp. 395–410, 2017.
- [16] S. Fekri-Ershad and F. Tajeripour, “Impulse-Noise resistant color-texture classification approach using hybrid color local binary patterns and pullback-leibler divergence,” *The Computer Journal*, vol. 60, no. 11, pp. 1633–1648, 2017.
- [17] J. A. Adams et al., “Cognitive task analysis for developing unmanned aerial vehicle wilderness search support,” *Journal of cognitive engineering and decision making*, vol. 3, no. 1, pp. 1–26, 2009.
- [18] H. Ayaz et al., “Monitoring expertise development during simulated UAV piloting tasks using optical brain imaging,” in *2012 IEEE Aerospace Conference*, 2012, pp. 1–11.
- [19] S. O. Murphy, C. Sreenan, and K. N. Brown, “Autonomous Unmanned Aerial Vehicle for Search and Rescue Using Software Defined Radio,” in *2019 IEEE 89th Vehicular Technology Conference (VTC2019-Spring)*, 2019, pp. 1–6.
- [20] H. El-Sayed, M. Chaqfa, S. Zeadally, and D. Putkal, “A Traffic-Aware Approach for Enabling Unmanned Aerial Vehicles (UAVs) in Smart City Scenarios*,” *IEEE Access*, pp. 1–1, 2019.
- [21] M. Meng et al., “BeamRaster: A Practical Fast Massive MU-MIMO System With Pre-Computed Precoders,” *IEEE Transactions on Mobile Computing*, vol. 18, no. 5, pp. 1014–1027, May 2019.
- [22] W. Shafik and S. A. Mostafavi, “Knowledge Engineering on Internet of Things through Reinforcement Learning,” *Int. J. Comput. Appl.*, vol. 975, p. 8887.
- [23] S. M. Matinkhah, W. Shafik, and M. Ghasemzadeh, “Emerging Artificial Intelligence Application: Reinforcement Learning Issues on Current Internet of Things,” in *2019 16th international Conference in information knowledge and Technology (ikt2019)*, p. 2019.
- [24] W. Shafik, S. M. Matinkhah, and M. Ghasemazade, “Fog-Mobile Edge Performance Evaluation and Analysis on Internet of Things,” *J. Adv. Res. Mob. Comput.*, vol. 1, no. 3.
- [25] B. T. Borst, S. Hegde, N. Proutière, A., 2004. Wireless data performance in multi-cell scenarios. *ACM SIGMETRICS Performance Evaluation Review*, 32(1), pp.378-380.
- [26] Jose, J., Ashikhmin, A., Marzetta, T.L. and Vishwanath, S., 2009, June. Pilot contamination problem in multi-cell TDD systems. In *2009 IEEE International Symposium on Information Theory* (pp. 2184-2188). IEEE.
- [27] Kazi, B.U., and Wainer, G., 2020. Coordinated multi-cell cooperation with a user-centric dynamic coordination station. *Computer Networks*, 166, p.106948.
- [28] A. Asadi, S. Müller, G. H. Sim, A. Klein, and M. Hollick, “FML: Fast machine learning for 5G mmWave vehicular communications,” in *IEEE INFOCOM 2018-IEEE Conference on Computer Communications*, 2018, pp. 1961–1969.
- [29] M. M. Ferdous, S. G. Anavatti, M. Pratama, and M. A. Garratt, “Towards the use of fuzzy logic systems in rotary wing unmanned aerial vehicle: a review,” *Artif. Intell. Rev.*, vol. 53, no. 1, pp. 257–290, 2020.
- [30] W. Shafik, M. Matinkhah, M. Asadi, Z. Ahmadi, and Z. Hadiyan, “A Study on Internet of Things Performance Evaluation,” *J. Commun. Technol. Electron. Comput. Sci.*, vol. 28, pp. 1–19, 2020.

Wasswa Shafik received a B.Sc. degree in Computer Engineering from Ndejje University Kampala, Uganda in 2016. He is currently pursuing his MSc degree in Computer Engineering (Networks Option) at Yazd University, Iran. He is a research member of the Intelligent Connectivity Research laboratory at Yazd University. His research interests include Smart Grid, Smart Cities, Cyber Security, 5G & Beyond, and Machine Learning.

S. Mojtaba Matinkhah is an assistant professor of Computer Engineering at Yazd University, Iran. He received his B.Sc. degree in software engineering from the University of Isfahan, Iran, an M.Sc. degree in Computer Security from Tehran Polytechnic, and his Ph.D. in Computer Networks from Tehran Polytechnic. He was a scholar visitor of Mississippi State University, the USA in 2011. He is currently the head of the Intelligent Connectivity Research laboratory in Yazd. His main interests are 5G, IoT, Smart Grids, post-quantum security, Cloud-Fog-Edge computing.

Mohammad Ghasemzadeh is an Associate Professor of Computer science and Engineering, working at Yazd University in Iran, since 1996. Received his BSc degree in computer science and Engineering from Shiraz University, Shiraz, Iran in 1989, the MSc degree in artificial intelligence and robotics from Amirkabir University of Technology (Tehran Polytechnic) in 1995. For his PhD research program, he worked at the University of Trier and at the University of Potsdam both in Germany, from Feb. 2002 to Feb. 2006. He completed his PhD degree in “Theoretical Computer Science”, in Nov. 2005. He also spent his sabbatical as guest/postdoc researcher at HPI in Germany, from Feb. to Nov. 2016. His main interests are algorithms, intelligent systems, machine learning, big-data analytics and soft computing.

Investigate Network Simulation Tools in Designing and Managing Intelligent Systems

Fatemeh Fakhar*

Department of Computer Engineering and Information Technology, Payame Noor University, Iran
Fakhar.mshd@gmail.com

Received: 04/Jul/2019

Revised: 24/Oct/2019

Accepted: 08/Dec/2019

Abstract

Network simulation is a technique that models network behavior by performing transaction calculations between different network entities and using mathematical formulas and taking observations of network products. A network simulator is a software program have been applied to analyze the performance of a computer network without the presence of a real network. Hardware equipment, equipment configuration, communication, and routing protocols and network traffic modeled in simulation software and the behavior of the network and its components examined from different dimensions. The user can also customize the simulation software according to their needs. Simulation software has different uses, and the user can use these tools to model their network by recognizing this software. In terms of research, it is difficult to create a network, especially large networks, in a real-time scenario, and it is not easily possible to carry out it in the real world, and it is very costly. So, simulators help network developers to control whether the network can work in real-time or not, or whether it is efficient enough. This reduces the time and cost of network application testing. Today, simulation technology is successfully used to model, design and manage a variety of intelligent systems. Numerous tools have been created in this regard. In this article, we review and compare important network simulators such as CloudSim, GloMoSim, GNS3, NS-2, Opnet, OMNet ++, NetSim, NS-3, AVRORA, Packet Tracer, QualNet, J-Sim, REAL and OptSim and their results. These comparisons express from several perspectives in the tables.

Keywords: Simulation; network simulator; network simulation; Network simulation languages; comparison.

1- Introduction

Today, there are huge industrial and economic activities around the world, and factories, manufacturing, and industrial centers, with the workforce and availability of the raw materials they need, sometimes run around the clock. In the meantime, the most important thing that industry owners and service providers pay particular attention to is to optimize their activities and actions so that they can use the smallest materials and components and the required equipment to give the highest quality products or services to give. In another case, the owners of capital and groups that are in the early stages of entering these types of activities need to have an image of what they want to invest in, as well as the amount of utility (return) of that specific activity and general knowledge of the cycle of their desired activity. The first suggestion for this goal is a small (laboratory) sampling, which, of course, requires first costs for a project that we do not now have a comprehensive understanding of its problems. Network simulation means virtual simulation. The purpose of the simulation is to find problems in existing networks or find unexpected interactions on a network that has not yet been built [1].

By locating or preventing existing failures, reliability improved and costs reduced. Other reasons for using the simulator include the lack of hardware and the consequent lack of hardware reform problems or other problems, such as the complex configuration of real-world equipment [2]. In the field of computer research and network communication, simulation is an important technique, because network behavior modeled by calculating the interconnection between different network components using mathematical formulas. Simulation can also be modeled by real or virtual recording and periodic real-time observation of real-world networks. Once the data is obtained through observations of simulation experiments, the network behavior and supported protocols viewed and analyzed in a series of off-line testing experiments. Also, all types of environmental features modified in a controlled way to test how the network can work under the combination of different parameters or configuration conditions. Comparing simulators and choosing one of them will help organizations and groups achieve the goal of each project.

In [1]: Some of the simulators are dedicated to a wireless network, some of them are dedicated to a wired network or both types of networks. Because of wide variations in operating systems, hardware requirements, programming software requirements, output features, and scalability, it is very difficult to choose a suitable simulator for a specific

job and had one Table for Comparison of 14 simulators based on 7 general information of simulators.

In [28],[30] presented a comprehensive survey on current network simulators and introduced their main features, consider their advantages and disadvantages, but introduced a few simulators (OPNET, NS2, NS3, OMNeT++), therefore it is not a good survey for those who find the Suitable network simulators for their researches.

In [29] introduced the main features of a different network simulator and considered their advantages and disadvantages (NS-2, NS-3, OPNET, OMNeT++, J-Sim, QualNet).

In [31][32]: One method of analyzing systems is the simulation. The basic idea is that if a system modeled, the corresponding results analyzed by changing the characteristics of this model. Because the process of modifying the model is cheaper than actual implementation, a range of situations analyzed at low-cost. thus, introduced the main features of the different network simulator and considered their advantages and disadvantages (NS-2, NS-3, OPNET, OMNeT++, J-Sim, QualNet, PeerSim).

One of the interesting areas in computer networks is to give the ability to evaluate and measure ideas, protocols, architectures and answer various questions, before the physical implementation of the network and its implementation. In terms of research, it is difficult to create a network, especially large networks, in a real-time scenario, and it is not easily possible to carry out it in the real world, and it is very costly. So, simulators help network developers to control whether the network can work in real-time or not, or whether it is efficient enough. This reduces the time and cost of network testing.

For this reason, we came up with a model of a list of widely used network simulators and comparing their properties, one of the ways to optimize the choice and finally speed up network access. Therefore, in this study, many(23) simulators from several different perspectives including accessibility, support, components, simulation, platform, visual/visual, supported protocols, testing, main applications, prominent features and advantages, disadvantages and constraints will compare and we outline the results for a better choice of network developers in the tables [3].

2- What is a Computer Network?

A computer network consists of two or more computers and accessories such as printers, scanners, and the like, which are directly used for the common use of hardware and software, data sources and connected devices. All hardware and software are available in the Source Network. In computer networks, according to the type of computer configuration, each computer can simultaneously use its resources, including tools and data,

with other computers at the same time. The reasons for using the network are as follows [4]:

- 1 - **Common use of resources:** Common use of a source of information or computer equipment, regardless of the geographic location of each resource, refers to the use of common resources.
2. **Reducing costs:** Focusing on resources and sharing them, avoiding their distribution in different units, and the specific use of each user in an organization, will reduce costs. This feature refers to network support providers in the network, This means that we can give various sources of information and systems in the second version of the network and support and if you do not have access to one of the sources of information on the network "due to system failure", use backup copies. Support for servers in the network increases the system's continuous activity and readiness.
4. **Reducing the time:** Another goal of creating computer networks is to set up strong links between remote users, meaning there is no geographical limitation of information exchange to reduce the time of information exchange and the use of their resources.
5. **Ability to develop:** A local network expanded without changing the structure of the system and become a larger network. Here, the cost of developing a system for the cost of facilities and equipment needed to expand the network considered.
6. **Communications:** Users can exchange their messages through existing innovations such as e-mail or other messaging systems; even file transfers are possible.

3- Simulation

Simulation technology and software are one of the most powerful methods and tools available to managers, industry engineers, system analysts, and so on, which enables them to make systems, in hands, before making any decision about any production system, service, Modeling and simulating them, performing or working them, and making necessary statistical surveys in all its dimensions to make better decisions, with the goal of reducing costs and increasing profit (or efficiency).

Using simulation, a range of dynamic (dynamic) issues analyzed in the areas of manufacturing, support, and services. The simulation allows for modeling the flow of materials and goods, human resources and information in the organization, and analyzing the system by simulating and adjusting different scenarios, 3D animations, and ... It concerned with potential improvements [5].

4- Network Simulator

A network simulator is a piece of software or hardware that predicts the behavior of a computer network without a real network. A network simulator is a software program that imitates the function of a computer network. In simulators, the computer network modeled with devices and traffic, and then its efficiency analyzed and analyzed. Usually, users can customize the simulator to fulfill their own analytical needs. Simulators generally support well-known protocols that used today, such as wireless LAN(WLAN), WiMAX (Worldwide Interoperability for Microwave Access), UDP (User Datagram Protocol), and TCP (Transmission Control Protocol) [6].

Most commercial user interface graphical use simulations. Some network simulators need the comments of scripts and commands (network parameters). The network parameters define the network status (place of nodes, links (and events), data transfer, link failure, etc.). The most important output of the simulators is the tracking file. Tracking files can document any simulation event that analyzed and analyzed. Some simulators have added functions to capture data directly from the environment at different times of day, week, and month to show the average, worst and best modes. Network simulators give other tools for facilitating visual analysis of trends and potential bottlenecks. Simulation of the network is difficult, such as, when there is large congestion, the average occupancy estimate due to high variance is difficult. To estimate the network buffer overflow, the time required to respond increased [7]. Specific techniques, such as variant control, important sampling, etc., which extends the simulation speed.

5- Application of Network Simulators

Network simulators solve a lot of needs, faster and cheaper than the network simulator compared to the cost and timing of launching the test bed for a large project that includes computers and routers and data links. The simulators allow engineers and researchers to simulate scenarios that are difficult and expensive to carry out in real hardware with several ninety and new protocols tested on the network. Network simulators are useful because they allow researchers to test network protocols or change protocols in controlled and renewable environments. A kind of network simulator includes a range of network technologies and can help users build complex networks of simple blocks such as nodes and links, and are hierarchical networks with different types of nodes such as computers, hubs, network bridges, routers, switches, links and mobile units [8].

6- Network Simulation Tools and Soft wares

An important part of the development of each system is evaluation of its performance with regard to delays and delays in different real-life scenarios. In many cases, the performance and applicability evaluation of the network through simulation experiments, which also requires a suitable environment and simulation tools. Several tools and techniques have been created in this regard. For example, an event-driven simulation technique used, which is the basis of many new simulators. In a group of types of simulator software in communication networks classified into three main categories [9]:

1. general Purpose Simulation Language
2. communication Oriented Simulation Language
3. Network simulation software

Understanding the performance of these tools is an important step towards developing an overall method for generating network simulators [10]. Therefore, the following features of the simulators have discussed below.

7- Network Simulation Tools

This group includes full packets that Communication-Oriented Simulator can simulate communication networks without the need for coding and usually through graphical interfaces, presence of simulated elements corresponding to the actual elements (routers, switches, ...). In addition to enhancing accuracy, it improves ease and speed in the simulation process and is very suitable for unfamiliar users with programming technology. COMNET III, BoNes PlanNet, NETWORK II and L.NET III are examples. [14] The group is successful. Altogether, professional users, especially those dealing with specific networks, prefer simulation languages with difficulty working with them, and, in contrast to users who deal with the simulation topic in a cross-sectional fashion, packets like Prefer the instrument. Network simulation tools should have the following five characteristics [15][16]:

1. **Flexibility in modeling:** the user must be able to add new types of common network resources such as nodes, links, and protocols to the emulator's suite.
2. **Ease of Modeling:** the existence of graphical interfaces and the possibility of structured modeling, in the form of complex models based on simple models, as well as the ability to reuse modules are features that accelerate the simulation process.
3. **Fast execution of models:** processing time in large simulations is important for networks with a large number of nodes, which requires proper memory management.
4. **Animation:** graphical display of network elements that are exchanging messages with each other to solve simulation errors and understand how it works. In

some simulation software, the simulator runs simultaneously with the implementation of the simulator and in some others after it performed in the form of Playback.

5. **Ability to Re-run and Repeat Simulation:** The purpose of the simulation is mainly to investigate the effect of one or more parameters (average packet length or buffer capacity) on its efficiency, and the repeatability is a necessary condition for this software.

In general, it should note that the creation of an accurate and valid network simulator requires the use of simulation technology along with network knowledge and protocols. Of course, along with the above features, there will be some capabilities on the value of each simulator, among which are the following [16]:

- ✓ Presence of Built-in Modules corresponding to network elements and protocols.
- ✓ existence of a Random-Number Generator and, in more advanced forms, the ability to create quantities with different random distributions, because most occurrences in a simulation process are of the type of random processes.
- ✓ Support for users with timely upgrades (especially for new protocols) with full documentation.
- ✓ Providing reports on network performance parameters (output rate, efficiency, transmission delay and so on) in the form of figures and curves, together with the possibility of performing statistical operations on the results of other positive features of a simulator [17].

8- Compare and Evaluate Tools and Languages for Network Simulation

With the growing spread of research and simulation tools, many tools have developed, and since familiarity with single simulators is time-consuming, choosing the right simulator is very important. Therefore, researchers have evaluated and compared the various network simulation tools to help users decide on the usefulness simulator. Since the model correction process is cheaper than real implementation, a range of scenarios analyzed at low-cost. Therefore, several important simulators have investigated and compared in Table 1-1 to Table 1-3.

9- Conclusions

In the network research area, it is very costly to deploy a complete test bed containing multiple networked computers, routers and data links to validate and verify a certain network protocol or a specific network algorithm. The network simulators in these circumstances save a lot of money and time in accomplishing this task. Network

simulators are also particularly useful in allowing the network designers to test new networking protocols or to change the existing protocols in a controlled and reproducible manner.

In this paper, we present a comprehensive survey on current network simulators. We introduce their main features, consider their advantages and disadvantages, and discuss the current and future developments. We hope this survey to be a good reference source for those who feel difficult to find the appropriate network simulators for their research or practical requirements.

This article reviews and compares the tools and simulators of the network. The research shows that this kind of simulator is due to network control to decide whether the network is capable of working in real-time or not and has capacity to reduce the time and cost required to test the functionality of the network. In this paper, we tried to compare and compare tabular characteristics and application of the most (23) common simulators. Since the use of a network simulator is effective in the performance of a project and other important issues in laboratory research, depending on the specific characteristics of each simulator, the use of its types can vary depending on the application. Among the most important goals and achievements of research are:

- Network simulation means its virtual simulation. The goal is to simulate problems on existing networks or to find unexpected interactions on a network that has not yet been built. By locating existing problems and preventing them from occurring, it is possible to improve reliability and reduce costs.
- Another reason for using a simulator is the lack of need for hardware and the consequent lack of equipment repair problems or problems such as the complex configuration of real-world equipment.
- Computer simulation is used to study and study most systems such as transportation, hospital, industrial systems, manufacturing, traffic, warehouse, etc. due to its practicality and having its own advantages.
- we have done a comparative study on different network simulators and investigated them from different perspectives, including accessibility, applications, visibility, etc. This research focuses on networking tools and simulators by the features presented in the above tables. Because it seeks to compare tools and simulators that speed up work, reduce errors, increase the level of abstraction and cut complexity. In the table below, these features and parameters are also compared with short.

Table 1-1: Comparison of network simulators based on general characteristics of simulators

Features	NS-2	NS-3	Opnet	OMNET++	Netsim	Glomosim	NCTUns
First version	1996	2008	1986	1997	2002	1998	1999
Use license	open source	open source	Commercial	open source (For students and educational purposes) and commercial (for commercial purposes)(Dedicated	open source	open source
Language required (components)	OTCL (higher level) and C ++ (lower level)	C ++ (higher level) and Python (lower level)	C C++	C ++ and NED ¹ (Model Structure)	C++ Java	ParSEC ²	C++
General simulator	✓	✓	✓	✓	✓	✓	✓
Availability (accessibility)	Open source software (including source code and tests and examples), Excellent	Open source software (including source code, tests, and examples), Excellent	Commercial software (has no public access), good	Free software only for scientific and nonprofit use, good	Commercial simulator (includes free demo), Excellent	Open source simulator (included below .bin)	good
Easy to use	Hard	Hard	simple	simple	simple	Hard	Hard
Guidance and support	<ul style="list-style-type: none"> • Have a good guide • Public access to a license or license to use the software • Source code and examples 	<ul style="list-style-type: none"> • Have an excellent guide • Access to shared code page • Access to library documents • Access to API documents 	<ul style="list-style-type: none"> • Have an excellent guide • Mailing list (maintenance license required) • Source code and examples 	<ul style="list-style-type: none"> • Have an excellent guide including a programming guide • Has a library of simulation classes 	<ul style="list-style-type: none"> • Simulation engine • Graphical applications • My SQL databases • A laboratory menu includes training, routing protocols 	<ul style="list-style-type: none"> • Ability to expand and add new code • Predefined library components • Public access to the certificate 	<ul style="list-style-type: none"> • continuously supported, maintained and im-proved • New functions and network types are continuously added to NCTUns to enhance its functions, speed and capabilities
Learning time	large	moderate	large	moderate	low	moderate	low
Download and install time	moderate	long time to download and install all the necessary packages and software support	moderate	<ul style="list-style-type: none"> • Very simple and time consuming. • Easily available 	Easily available	Easy to download and install	Esay
Support for nodes mobility	✓	✓	✓	✓	✓	—	✓
Graphical interface support (Visual / Visibility)	<ul style="list-style-type: none"> • Poor • Lack of graphics output in the original version of the simulator • Added visualizer to eliminate the text-based interface 	<ul style="list-style-type: none"> • Good • View simulation result using NAM animator 	<ul style="list-style-type: none"> • Excellent • Advanced graphic interface 	<ul style="list-style-type: none"> • Good • Graphical user interface for execution • Support for strong graphical interface 	Have graphical applications	<ul style="list-style-type: none"> • Limited • Visual programming for the visual design of the ParSEC program 	<ul style="list-style-type: none"> • easy navigation GUI environment • GUI program contains four main components: <ul style="list-style-type: none"> - Topology Editor - Node Editor - Performance Monitor - Packet Animation Player Among these four components, the Node Editor is relevant to module developers.
Simulation scale	small	small	large	small	Big enough	large	moderate

¹ Network Description

² Parallel Simulation Environment for Complex Systems

Features	NS-2	NS-3	Opnet	OMNET++	Netsim	Glomosim	NCTUns
Number of nodes supporters	Up to 3000	-	210 to 290 in all topologies	-	-	up to 10000	Up to 4096 nodes
Parallelism	-	-	✓	MPI/PVM	-) (SMP/Beowulf	-
Supported platforms (platform)	Windows (CYGWIN), Linux MINT/UBUNTU / FEDORA / MINT / UNIX	Windows (CYGWIN), Linux MINT/UBUNTU /Free BSD X86/ FEDORA / MAC OS	Hewlett-Packard, Sun-4 SPARCVarious, Solaris 2.6, 7 8Microsoft Windows NT 4.0/Windows 2000Required System Patches-	Windows, Unix-based, Mac OS X 10.6 and 10.7	Windows (7, Vista) , windows XP	Windows, Linux, Sun SPARC Solaris	FreeBSD, Fedora, Red hat, Ubuntu, and Debian
Analysis tool	✓	✓	✓	✓	✓	—	✓
Support for network visualization	✓	✓	✓	✓	✓	poor	poor
Possibility to define and change the scenario	✓	✓	✓	✓	✓	poor	✓
Create a Trace File	✓	✓	✓	✓	✓	✓	✓
Support for design and implementation protocols (both simulated with wire and wireless)	✓	✓	✓	✓	✓	✓	✓
protocols supported (OSI' protocols)	<ul style="list-style-type: none"> •TCP/ UDP •FTP / Telnet and other web protocols in fixed and variable bit rate users •802.11 and TDMA² MAC layer protocols •Routing and multiprocessing protocols and, in general, all layers 	<ul style="list-style-type: none"> •Routing protocols •Management Protocol •MAC layer routing protocols •TCP / UDP / Wimax 	<ul style="list-style-type: none"> •TCP / IP /ATM •Frame relay •Protocol 802.11 •Wireless protocols 	<ul style="list-style-type: none"> •UDP / IP and ICMP •Transfer Protocol and MAC protocols •Stop and wait protocol •Some local protocols for wireless sensor networks 	<ul style="list-style-type: none"> •Aloha/ Slotted Aloha / Ethernet •CSMA/CD •Fast Ethernet/Gigabit Ethernet •Token Ring/Token bus in Wlan •ATM ,TCP/ RIP •OSPF, BGP, MPLS, ZigBee 802.15.4 •Wimax / Wireless Sensor Networks 	<ul style="list-style-type: none"> •Provide a protocol stack that includes models for channel, radio, MAC, network, transmission, and higher layers. •Simultaneous protocols and wireless networks 	<ul style="list-style-type: none"> •IEEE 802.3 CSMA/CD MAC •IEEE 802.11 (b) CSMA/CA MAC •IEEE 802.11(e) QoS MAC •IEEE 802.11(b) wireless mesh network routing protocol •DVB-RCS satellite MAC and PHY •spanning tree protocol, IP Mobile IP, Diffserv (QoS), RIP, OSPF, UDP, TCP, RTP/RTCP/SDP, HTTP, FTP, Telnet, Bit Torrent, etc.
Ability to interact with the actual system	✓	✓	✓	✓	✓	✓	✓
Ability to communicate with other components	✓	✓	✓	✓	✓	poor	✓
Fast simulation	moderate	moderate	Excellent	moderate	good	moderate	poor

¹ Open Systems Interconnection
² Time Division Multiple Access

Features	NS-2	NS-3	Opnet	OMNET++	Netsim	Glomosim	NCTUns
Event type simulation (Simulation Mode)	<ul style="list-style-type: none"> Discrete event Sync, single-threaded, queue-based off-the-shelf event, absolutely definitive, version Available Distributed and Parallel 	Discrete event (Separate event)	<ul style="list-style-type: none"> Discrete event Synchronized, single-threaded, queue-based, and fast-ended, absolutely definitive Multi-threaded, queuing-based event, distributed simulation and HLA1 Simultaneous 32-bit and 64-bit simulation core 	<ul style="list-style-type: none"> Discrete event (Separate event) Object-oriented and component-based modules 	<ul style="list-style-type: none"> Random discrete event Multi-threaded with Net Engine simulation core 	<ul style="list-style-type: none"> Discrete event Simulate a separate parallel event Use of synchronous and asynchronous algorithms Multi-threaded 	Kernel Inspection Method
Network Type (Available Module)	<ul style="list-style-type: none"> Wired / wireless / wireless sensor / ADHOC / MANET / Wired cum Wireless / SDN / VANET / Security /Vertical Handover 	<ul style="list-style-type: none"> Wired / wireless / wireless sensor / ADHOC / MANET / Wired cum Wireless / SDN / VANET / Device to Device Communication 	<ul style="list-style-type: none"> Wired / wireless / wireless sensor / ADHOC / MANET / Wired cum Wireless / SDN / VANET / Radio Network 	<ul style="list-style-type: none"> Wired / wireless / wireless sensor / ADHOC / MANET / Wired cum Wireless / SDN / VANET / WBAN / Under water sensor network / Social sensor network 	<ul style="list-style-type: none"> Wired and Wireless Sensor Network (Wireless LAN, WiMAX) 	<ul style="list-style-type: none"> Wired, Wireless, AdHoc networks but currently have wireless support 	<ul style="list-style-type: none"> Wired, Wireless, Adhoc and Wireless Sensor Networks
purpose of the test (Ease of testing a computer program)	<ul style="list-style-type: none"> Testing the main platform of this simulator (VINT project) Apply daily validation tests to NS2 and send to NSNAM 	A unit test to inform the user of the correct simulator functionality (in parallel with the help of WAF)	<ul style="list-style-type: none"> Testing technology design in real scenarios Test user scenarios for the correct functioning of the new network to join the central network 	Test and debug simulation of users by a strong graphical user interface	Test quick connectivity in K-NetSim (Test connections between network equipment and protocols.)	Test routing algorithms	<ul style="list-style-type: none"> Test and use as an emulator (very useful as function and the performance of real-world devices can be tested under various simulated network).
purpose of creating the simulator (Main application cases)	<ul style="list-style-type: none"> Especially designed for network simulation and research protocols General simulator Use in network research and simulation of IP networks Uses for scalable wireless and wired network simulation 	<ul style="list-style-type: none"> The main goal is to use research, training and its sustainability. Use in research and education issues in wireless and wired networks 	<ul style="list-style-type: none"> The GUI provides flexibility to make complex scenarios easier Design, deployment and management of infrastructure, equipment and network applications Use in the design and study of communication networks, equipment, protocols and applications and wireless environments 	<ul style="list-style-type: none"> Comprehensive analysis and component and powerful simulation of the discrete event at the university provide research General simulator Wired and wireless communication networks, protocols and queues in the network Modeling distributed hardware systems Validation of hardware architecture Assessing the performance aspects of complex software systems Modeling and simulating event systems - distinct 	<ul style="list-style-type: none"> This is a leading simulation software for modeling and simulating protocols, which allows us to analyze computer networks with great depth, power and flexibility. Use in experiments and research in network labs Development in order to study security technology and support for cyberspace exercises 	<ul style="list-style-type: none"> It provides modular simulation for the protocol column and free for scientific research, but is not updated regularly. Uses in wireless and wired mobile networks, and in particular the simulation of parallel wireless networking 	<ul style="list-style-type: none"> Integrated protocol module and can be used for both simulations MANET and VANET. supports remote and concurrent simulations use any real-life UNIX network configuration and monitoring tools.
features and benefits	<ul style="list-style-type: none"> Easy to add new protocols A large number of available protocols available to the public Ability to add new protocols and public access to them 	<ul style="list-style-type: none"> NS3 is not a NS2 extension, a new simulator. open source Virtualization support Ability to add new protocols 	<ul style="list-style-type: none"> Opnet communicates with other simulators Motorcycle engine simulator discrete event Simulated rechargeable wireless support Integration, debugging and 	<ul style="list-style-type: none"> Powerful GUI (easier tracking and debugging) Power consumption calculations Has a compiler for the NED topology description language Network graphic editor Command line interface to run 	<ul style="list-style-type: none"> Simple and understandable user interface Learning concepts and not engaging users by choosing unnecessary devices Implement password recovery Implement Telnet between 	<ul style="list-style-type: none"> Quick integration of models developed in different layers Use in a parallel development environment Ability to develop Run by a large set of synchronization protocols 	<ul style="list-style-type: none"> easy navigation GUI environment. Realistic network traffic can be generated by realistic life applications to generate more realistic simulation results evaluated easily the

¹ High Level Architecture

	<ul style="list-style-type: none"> Object design and the ability to create protocols and use them Features in sensor networks including sensor channels, battery models, low-power protocols Perform simulations at the closed level, resulting in more precise results 	<ul style="list-style-type: none"> The software core written in C++ and the Script interface in the Python language Attention to realism and the approach of design to real systems Software integration Customize the output without rebuilding the network core 	<p>GUI-based analysis</p> <p>Hierarchical Modeling Environment</p> <ul style="list-style-type: none"> The fastest discrete event simulator engine among upcoming industrial solutions A complete library of Opnet models Grid computing support for distributed simulation Integration, debugging, graphical user interface and analysis 	<p>the simulation</p> <p>Simulation management facilitating tools</p> <ul style="list-style-type: none"> Creating an Infrastructure for writing various simulations 	<p>devices</p> <ul style="list-style-type: none"> Simple display of routing table Build and upgrade the lab by the builders Simulate network traffic using virtual packet technology 	<ul style="list-style-type: none"> Run in two types of shared or distributed memory systems 	<p>performance of any real-life application under various simulated network conditions.</p>
Disadvantages and Limitations	<ul style="list-style-type: none"> supports only two MAC wireless protocols, 802.11 and one TDMA protocol You do not need to get to know writing programming language Supports two wireless MAC protocols: 802.11 and TDMA Lack of user-friendly package due to text-based interface Need advanced skills to do the right simulation Unavailable customization Lacks a functional model 	<ul style="list-style-type: none"> The Python link does not work in Cygwin. Only IPv4 is supported Lack of available models Lacks a graphical interface for creating topology Has a laboratory-level visibility capability 	<ul style="list-style-type: none"> A commercial product Memory consumption models Minor training Limit the accuracy of the results to the sampling resolution Inefficient in the absence of the event for a long time Missing laboratory guides 	<ul style="list-style-type: none"> The number of protocols is not large enough. Compatibility problem (not portable) OMNeT++ is a bit slow due to the implementation of long simulation and high memory consumption Lack of sufficient number of available protocols 	<ul style="list-style-type: none"> Supports 47 different Cisco devices Supports 200 devices on network topology 	<ul style="list-style-type: none"> Limited to IP network due to low level design assumptions No support for adding new protocols Limited to package formats and energy models It provides the Random Waypoint mobility model, which may not be suitable for all types of simulations 	<ul style="list-style-type: none"> Connection through dispatcher with simulation server is not much stable. The programming is not supported by the NCTUns. So, the parameters are needed to set via GUI. The manipulation of every node has to be done node by node or all nodes in the same time.

Table 1-2: Comparison of network simulators based on general characteristics of simulators

Features	QualNet	CloudSim	Tossim	J-Sim	REAL	Avrora	Packet tracer
First version	2000	2009	2003	2002	2008	2007	2006
Use license	Commercial (Separate license for academics and others)	open source for Cloud Computing	open source	open source	open source	open source cycle-accurate simulator	open source
Language required (components)	C++	Java	Python, C++, NesC	Java	Scenario (and C) Modulated Structure .NET) Java	Java	Cisco Input / Output Instructions Javascript/CSS
General simulator	✓	for Cloud Computing	Special for WSN	✓	✓	Only for WSN	✓
Availability (accessibility)	Commercial version of the GloMoSim simulator, Excellent	Open source simulator (includes package, class, and example)	Good	Open source software (with available code and examples), moderate	Open source simulator	Open source simulator (online access to documents)	Free and open source software
Easy to use	moderate	hard	very simple	simple	simple	moderate	moderate

Features	QualNet	CloudSim	Tossim	J-Sim	REAL	Avrora	Packet tracer
Guidance and support	<ul style="list-style-type: none"> • Excellent Documentation and Standards Library • Certificate of Product Extensibility • Annual support and maintenance contract • Access to new releases and software certification 	<ul style="list-style-type: none"> • Have an excellent guide • Access to API1 documents • Source code and examples • Certification access • Access to .jar NS-3 NetSim files 	<ul style="list-style-type: none"> • online documents • provides images instead of text commands • compiles TinyOS code directly 	<ul style="list-style-type: none"> • Have a good guide • Public access to the certificate • Source code and examples 	<ul style="list-style-type: none"> • Ability to use new modules • Accessibility from the main server • Excellent guide • Online documents 	<ul style="list-style-type: none"> • Access archive to previous projects • Lacks active support • Access a large amount of information through the help option • Online documents 	<ul style="list-style-type: none"> • In-service tutorials • Public access to the certificate
Learning time	Very simple	simple	very simple	moderate	simple	moderate	moderate
Download and install time	short time	free or open source, available for public download.	Esay	Easy to download and install	Easy as Available Site	Esay	Esay
Support for nodes mobility	✓	-	✓	✓	<ul style="list-style-type: none"> • Graphical interface written in Java to create faster simulation scenarios 	<ul style="list-style-type: none"> • has a control flow graph for graphical representation of program instructions to understand the structure and understand how the code is compiled. 	<ul style="list-style-type: none"> • Graphical interface with multi-language support, has two logical topologies and physics along with improved multi-user environments
Graphical interface support (Visual / Visibility)	<ul style="list-style-type: none"> • Excellent Graphical design of a network model using libraries by the QualNet animator 	<ul style="list-style-type: none"> • itself do not have a graphical user interface, extensions such as CloudReports offer a GUI for CloudSim simulations • Graphical interface for reading network topologies • Have Limited graphical tools (Via CloudAnalyst) 	✓	<ul style="list-style-type: none"> • Good • 2 user graphical interfaces and use Console 	✓	No graphical interface	<ul style="list-style-type: none"> • a built-in GUI interface is supported, with a possibility to trace and store all events. • Different languages supported for the GUI. • User-Friendly Interface
Simulation scale	large	small	small	small	large	Very large	large
Number of nodes supporters	500 to 20000	small	thousands of nodes	Maximum 1000 for a tree	Up to 30000	thousands of nodes up to 10,000 nodes	1000 to 3000
Parallelism	(SMP/Beowulf) ✓	not consider parallel experiments	-	Based on RMI	✓	✓	✓
Supported platforms (platform)	MAC os, Unix, Windows, Linux, Solaris, DOS	LINUX (ubunt), Windows 7	Linux Operating Systems or on Cygwin on Windows	Windows XP, Vista & 7, MAC OS X, Linux Matlab	Unix ,BSD 4.3 ,Solaris ,Ultrix , Free BSD ,Sun OS ,IRIX	Mica2 and Mica Z independent operating systems	Windows) XP· Vista·7 ,(Linux)ubuntu ·fedora(
Analysis tool	✓	✓	✓	✓	✓	✓	Linux ,Windows
Support for network visualization	✓	✓	✓	✓	✓	✓	✓
Possibility to define and change the scenario	✓	✓	difficult	✓	✓	allow users to create a new simulation type and choose the type of simulation to perform, depending on the number and orientation of the nodes	✓
Features	QualNet	CloudSim	Tossim	J-Sim	REAL	Avrora	Packet tracer

¹ application programming interface

Create a Trace File	✓	✓	configuration difficult to construct	✓	✓	✓	✓
Support for design and implementation protocols (both simulated with wire and wireless)	✓	✓	✓	✓	✓	✓	✓
protocols supported (OSI¹ protocols)	<ul style="list-style-type: none"> • Has a protocol stack • Routing protocols • Message protocol and wireless, wired and widespread networks 	Open flow	<ul style="list-style-type: none"> • includes two different models: Radio models for all kinds of transmission aspects and ADC models 	<ul style="list-style-type: none"> • Routing protocols such as RPSG2 and AODV3 • Protocols used in wireless lines and single networks 	<ul style="list-style-type: none"> • Flow control protocols (TCP and...) • Transition layer protocols 	<ul style="list-style-type: none"> • MAC layer protocols in wireless networks 	<ul style="list-style-type: none"> • BGP, EIGRP, EIGRPv6, OSPF, OSPFv6, RIP, RIPng, • DHCP, DHCPv6, FTP, HTTP, HTTPS, RADIUS, POP3, SMTP, SNMP, SSH, Telnet, TACACS. • SCCC, TCP, UDP. • ARP, CAPWAP, HSRP, HSRPv6, ICMP, ICMPv6, IP, IPv6, NDP. • Bluetooth, CDP, CTP, H.323, LACP, LLDP, PAgP, STP, USB, VTP.
Ability to interact with the actual system	✓	✓	configuration difficult to construct	✓	✓	✓	Emulate real network interface
Ability to communicate with other components	✓	✓	capture a wide range of network interactions	✓	✓	✓	✓
Fast simulation	Excellent	Fast (seconds)	good	Poor	good	Very fast performs as much as 20 times faster than most other simulators with equivalent accuracy	Fast
Event type simulation (Simulation Mode)	<ul style="list-style-type: none"> • Discrete event • Distributed and parallel simulation 	Discrete event (Separate event)	Discrete bit-level event	<ul style="list-style-type: none"> • Varied numerical method • Sync, single-threaded, fully-definite, multi-threaded, based on processes with Real-time, non-deterministic, component-oriented architecture 	threading directly and intolerably	Discrete event (Separate event)	<ul style="list-style-type: none"> • Use the time-out model • Discrete-event
Network Type (Available Module)	Wired / wireless / wireless sensor / ADHOC / MANET / Wired cum Wireless / SDN / VANET	Cloud Computing	Simulation of a TinyOS -based Wireless Sensor Network	Wired and Wireless Sensor Networks	Wired and Wireless Sensor Networks	sensor network simulation communicate via the radio using the software stack provided in TinyOS	.Packet level
purpose of the test (Ease of testing a computer program)	<ul style="list-style-type: none"> • Network communication testing • Test network connectivity features by users • Design protocols • Creating and mobilizing network scenarios • performance evaluation 	<ul style="list-style-type: none"> • Testing cloud networks due to limited test bed scales • Service testing of user generated networks in a controlled environment • Use of test methods and faults to fix defects • Has a library package test • Modeling and simulating infrastructures and cloud computing 	<ul style="list-style-type: none"> • simple and powerful testing for Wire-less Sensor Network • observes interaction difficult to live-capture 	<ul style="list-style-type: none"> • Design, create and test the unit of each component in this simulator • Modeling systems to apply distinct changes in objects • Provide all Java language features • Use in wireless and sensor environments 	<ul style="list-style-type: none"> • Testing scenarios in non-stress conditions • Output scenario testing of this simulator in a large amount of workload 	<ul style="list-style-type: none"> • Test the program before placing on the hardware by the user • Use GDB debugging for development and testing 	<ul style="list-style-type: none"> • Test in the Activity Wizard section of this simulator • Test in the Activity Wizard section of this simulator
Features	QualNet	CloudSim	Tossim	J-Sim	REAL	Avrora	Packet tracer

¹ Open Systems Interconnection
² Greedy Perimeter Stateless Routing
³ Ad-hoc On-demand Distance Vector

purpose of creating the simulator (Main application cases)	A unique platform for designing new protocols and their performance analysis	<ul style="list-style-type: none"> well known simulator for cloud computing can extended easily but currently it does not consider parallel experiments or lifecycles of VMs. 	Provides a simple and powerful simulator for wireless sensor networks	Especially for the components of the public network	Use to study dynamic behavior of flow control and congestion in networks	<ul style="list-style-type: none"> Use in wireless sensor networks can support thousands of nodes simulation can save much more execution time Enable validation of time-dependent properties of large-scale networks 	<ul style="list-style-type: none"> Use CCNA and CCNP training courses to create an unlimited number of network equipment and debug these networks without the use of real switches and routers
features and benefits	<ul style="list-style-type: none"> Powerful GUI It shows great scalability, it's time to simulate logically. Can support real-time speed to enable loop-free software, network simulation and hardware in loop modeling Designing a new prototype and optimizing the old model Network Performance Analysis Features speed, expandability, reliability, portability, expandability Remove the implementation limit of just one protocol at a time 	<ul style="list-style-type: none"> Supports modeling and simulation of large-scale cloud computing data centers, energy-conscious computing resources, federal clouds Support for dynamic insertion, stop and resume simulation Supports user-defined systems Optimizing the cost of access to resources focusing on improving profit 	<ul style="list-style-type: none"> very simple but powerful emulator for WSN support thousands of nodes simulation (very good feature: because it can more accurately simulate the real-world situation) Besides network, it can emulate radio models and code executions. This emulator may be provided more precise simulation result at component levels because of compiling directly to native codes 	<ul style="list-style-type: none"> Supports energy modeling, with the exception of radio power consumption, Support for mobile wireless networks and sensor networks. Customer-referenced architecture Supports energy modeling with the exception of radio power consumption scalable Simplified equipment models 	<ul style="list-style-type: none"> ability to add different modules Ability to test and configure the simulator as needed Custom design and production Uses for analyzing TCP, FIFO, and ... protocols. 	<ul style="list-style-type: none"> Managing networks with 10,000 nodes (at speeds of around 20 times the speed of other simulators and precision with accuracy) Validation of independent network time features A scalable and accurate simulator for hardware platforms in sensor programs Implementation of sensor network along with precise timing Ability to connect programs with radio help flexibility and portability 	<ul style="list-style-type: none"> Powerful simulation Interoperability, writing and simulation evaluation A combination of real simulation and configuration experience Supports HSRP1
Disadvantages and Limitations	<ul style="list-style-type: none"> A commercial product Hard Install in Linux Reduce personalization due to some tools 	<ul style="list-style-type: none"> Limit testing to the test bed scale Reproduce the problem of results 	<ul style="list-style-type: none"> Only for applications of TinyOS not good for the performance metrics of other new protocols. every node has to run on NesC code, a programming language that is event-driven, component-based and implemented on TinyOS only for type of homogeneous applications motes-like nodes are the only thing that TOSSIM can simulate 	<ul style="list-style-type: none"> Low simulation performance. Only MAC protocols for 802.11 wireless networks are provided. Extra cost at runtime Low performance Supports a MAC wireless protocol: 802.11 Unnecessary overhead at runtime 	<ul style="list-style-type: none"> The limitation of simulation time to cases such as protocols 	<ul style="list-style-type: none"> Failure to model the Clock output change Inability to model mobility is 50% slower than TOSSIM. 	<ul style="list-style-type: none"> Inappropriate for modeling production networks <ul style="list-style-type: none"> Has technical limitations other than Cisco3 not support modeling

Table 1-3: Comparison of network simulators based on general characteristics of simulators

Features	DRMsim	SSFnet	GrooveNet	Trans	GNS3	JiST	OptSim
First version	2003	2004	2001	2002	2006	2004	2005
Use license	open source	open source	open source	open source	open source	open source	open source
Features	DRMsim	SSFnet	GrooveNet	Trans	GNS3	JiST	OptSim
Language required (components)	Java	Java/C++	C++	Java/C++	Dynamics	C/C++/python/Java	C++ /C

¹ Hot Stand by Routing Protocol

General simulator	Only for WSN	✓	✓	Only for WSN	✓	✓	✓
Availability (accessibility)	good	good	medium	Poor	Open and free simulator	Very good	Business simulator up to 1998 (now free in academic cases)
Easy to use	moderate	Hard	simple	moderate	simple	easy to use	moderate
Guidance and support	<ul style="list-style-type: none"> made to compromise on the quality of the code written so that extensibility and reusability are maximized document availability moderate 	<ul style="list-style-type: none"> simulate Core Internet protocol models (IP, BGP4, OSPF, TCP, UDP), Sockets and various workload-generating client-server application models and Protocol validation tests using SSFNet. 	<ul style="list-style-type: none"> a hybrid simulator for geographic routing that address the need for a robust, easy-to-use realistic network and traffic simulation. is an opportunistic broadcast protocol with minimal handshaking between sending and receiving parties with little or no shared state information among neighboring vehicles 	document availability poor	<ul style="list-style-type: none"> Have an excellent guide Limited Certification 	JiST simulations are written in Java, compiled using a regular Java compiler, and run over a standard, unmodified virtual machine.	<ul style="list-style-type: none"> A wide library of models, standard components and parameters accurate
Learning time	good	large	good	low	low	good	low
Support for nodes mobility	✓	✓	✓	✓	-	✓	
Graphical interface support (Visual / Visibility)	<ul style="list-style-type: none"> Topology Generators like Brite, Inet, GLP etc. Import external topologies. E.g. CAIDA maps 	<ul style="list-style-type: none"> facilitates topological network component addressing and automated IP address allocation (CIDR compliant) 	<ul style="list-style-type: none"> coded in C++ and Matlab provides GUI for drawing structures and graph. a cross- platform GUI in Qt 	<ul style="list-style-type: none"> open source GUI simulation tool integrates traffic and network simulators (SUMO and ns2) to generate realistic simulations of Vehicular Ad-hoc networks (VANETs) 	<ul style="list-style-type: none"> Simple and general graphical interface (includes everything) 	limited Graphical interface	<ul style="list-style-type: none"> Visual simulation of communication systems Interface with 3D tools, Liekki design software, visual vector analysis Inner visual connections provides feedback for efficient simulations.
Backup the graphical interface	✓	✓	✓	✓	Excellent	✓	
Simulation scale	large	Very large	large	large	small	Excellent	small
Number of nodes supporters	Up to 10000 nodes	Up to 100000 nodes	-	Up to 3000 nodes	Up to 1000 nodes	large	Up to 1000 nodes
Parallelism	-	✓	✓	-	✓	Java, win, Mac, Linux, Unix	-
Supported platforms (platform)	UNIX, Linux, Mac OS	<ul style="list-style-type: none"> Parallel execution under Linux, Solaris, and Windows NT using JDK1.2 and higher. Other platforms may support parallelism as well. 	<ul style="list-style-type: none"> Linux Operating Systems with kernel version 2.6 (Ubuntu-12.4 tested and SUSE). requires Qt 3.x graphic library. 	Linux, Windows (trace-generation mode)	Router platforms for Cisco and Linux, Windows and Mac	Linux, Android 4.1+, iOS 8+ and Microsoft Windows.	Windows, Unix, Linux
Analysis tool	✓	✓	✓	✓	✓	✓	✓
Support for network visualization	✓	✓	✓	✓	✓	✓	✓
Features	DRMsim	SSFnet	GrooveNet	Trans	GNS3	JiST	OptSim
Possibility to define and change the scenario	✓	✓	✓	✓	✓	✓	✓

Create a Trace File	<ul style="list-style-type: none"> compromise on the quality of the code written so that extensibility and reusability are maximized. 	✓	✓	<ul style="list-style-type: none"> Mobility trace generation for ns2 from TIGER and Shape file maps (using the net convert tool from SUMO). 	✓	✓	✓
protocols supported (OSI¹ protocols)	<ul style="list-style-type: none"> Network and Data Link Layer RIP, BGP and NSR Supports high end routing protocols like BGP, DSDV up to 16000 nodes. 	<ul style="list-style-type: none"> with various highly scalable Internet protocols like IP, TCP, UDP, BGP, OSPF etc. and large network elements like Routers, Switches, Links, LAN's etc 	<ul style="list-style-type: none"> support three types of simulated nodes: <ul style="list-style-type: none"> vehicles which are capable of multi-hopping data over one or more DSRC channels fixed infrastructure nodes mobile gateways capable of vehicle-to-vehicle and vehicle-to-infrastructure communication. 	<ul style="list-style-type: none"> Physical, Network and Data link Layer Support 	OSPF ² , Ethernet and STP ³ protocols	used to implement IEEE 802.15.4 MAC-layer protocol	<ul style="list-style-type: none"> Random or fixed access protocols Reservation protocols High-speed communication protocols MAC protocol TCP / IP (most of the transmission and network protocol protocols)
Ability to interact with the actual system	✓	✓	✓	✓	✓	✓	✓
Ability to communicate with other components	✓	<ul style="list-style-type: none"> SSFNet models provide components for simulating networks at IP level and above and include models for hosts, router, links and a framework for modeling protocols. 	✓	✓	✓	✓	✓
Fast simulation	Fast	good	moderate	good	Fast	moderate	Fast
Event type simulation (Simulation Mode)	Discrete event	Discrete event	Hybrid simulator	Discrete event	Discrete event	discrete event	Separating the domain time and frequency range
Available module	Simulate dynamic path model	Modeling and simulating Internet and network protocols	Realistic simulation	Vehicle Adhoc Networks (VANETs)	<ul style="list-style-type: none"> a discrete-event network simulator for Internet systems targeted primarily for research educational use 	discrete event simulators, called virtual machine-based simulation	especially designed to be used in Monte Carlo simulation and Project Portfolio applications.
purpose of the test (Ease of testing a computer program)	<ul style="list-style-type: none"> careful analysis of the data structures in the network model as well as on the granularity and time management of the simulation model 	<ul style="list-style-type: none"> simulated Protocol validation tests using SSFNet. 	<ul style="list-style-type: none"> support multiple network interfaces for real vehicle-to-vehicle and vehicle-to-infrastructure communication:5.9GHz DSRC, IEEE 802.11a/b/g, 1xRTT and EVDO cellular interfaces. Communication over TCP or UDP sockets. All real vehicles communicate with DSRC or 802.11 with each other mobile gateways communicate with infrastructure nodes over the cellular interface 	<ul style="list-style-type: none"> has two distinct modes of operation: <ul style="list-style-type: none"> network centric application centric 	<ul style="list-style-type: none"> Virtual network testing functions exactly similar to the Cisco packet Tracer software. 	<ul style="list-style-type: none"> JiST is a high-performance discrete event simulation engine that runs over a standard Java virtual machine. It is a prototype of a new general-purpose approach to building discrete event simulators, called virtual machine-based simulation, that unifies the traditional systems and language-based simulator designs. 	<ul style="list-style-type: none"> Laboratory testing of equipment such as agile Test component features such as L-I curves for laser models and so on

¹ Open Systems Interconnection

² Open Shortest Path First

³ Spanning Tree Protocol

<p>purpose of creating the simulator (Main application cases)</p>	<ul style="list-style-type: none"> • Quick simulation of routing schemes in a wide dynamic network • study focusing on dynamic compact rout • evaluate the main performance metrics of routing schemes especially, the metrics related to the scalability and dynamic properties of these schemes (main goal) 	<ul style="list-style-type: none"> • Unique to increase scalability of modeling, traffic patterns, bandwidth and so on • basically, designed for simulating various network scenarios like network topology, protocols and traffic and also enables simulation of Wide Area Network like Internet. 	<ul style="list-style-type: none"> • Unique for its capacity to analyze performance and scalability of communication protocols between vehicles • capable of communication between simulated vehicles, real vehicles and between real and simulated vehicles. • Supports communication between real and simulated vehicles such that vehicles in the vicinity of each other are able to exchange packets 	<ul style="list-style-type: none"> • Provides the TraCI interface for generating range by connecting SUMO and ns2 with Google Earth • Realistic Joint Traffic and Network Simulator for VANETs 	<ul style="list-style-type: none"> • A tool for configuring and checking the network to participate in Cisco Exams • a discrete-event network simulator for Internet systems • targeted primarily for research • educational use 	<ul style="list-style-type: none"> • The JiST approach is inherently flexible, capable of transparently performing important cross-cutting program transformations and optimizations. • This transparency is a key benefit: simulation code that runs on JiST need not be written in a domain-specific language invented specifically for writing simulations, nor need it be littered with special-purpose system calls and call-backs to support runtime simulation functionality 	<p>Visual simulation of communication systems at the signal emission level</p>
<p>features and benefits</p>	<ul style="list-style-type: none"> • DRMsim is dedicated for a routing model simulation. • efficient graph structures and algorithms. • capability to import external topologies (e.g., CAIDA maps). • “100% pure Java” application which makes it executable on most platforms. • free of cost for research purposes. 	<ul style="list-style-type: none"> • Scalable high-performance Java simulation platform • simple, standardized syntax for high-level model description DML. • Allow Management of global traffic patterns • Support high performance network simulation • free or available at nominal cost for research purposes. 	<ul style="list-style-type: none"> • able to support hybrid simulation. • Support multiple vehicles, trip and mobility models over a variety of network link and physical layer models. • provide well-defined model interfaces that make easy to add different network models. • implement multiple rebroadcast policies to investigate the broadcast storm problem. • The well-defined graphical user interface makes it easy to auto-generate simulations 	<ul style="list-style-type: none"> • facilitate automated generation of random vehicle routes. • Map cropping and speed rescaling (only for TIGER maps) • Google Earth visualization of simulations (only for TIGER maps) • Provide TraCI interface for mobility trace generation by coupling SUMO and ns2. 	<ul style="list-style-type: none"> • An excellent complementary tool for networking • High-quality design of complex network topologies • Real-time simulated virtual network communication <p>Receive packets using the wireshark protocol</p>	<ul style="list-style-type: none"> • The JiST approach is inherently flexible, capable of transparently performing important cross-cutting program transformations and optimizations. 	<ul style="list-style-type: none"> • Has a virtual lab with over 600 components and fiber • Quick learning curve • Has several implementation engines including both time domain and frequency separation cases • Has MATLAB interfaces, functional programming and ... • Luna Optical Vector Analyzer
<p>Disadvantages and Limitations</p>	<ul style="list-style-type: none"> • only the routing protocol can be simulated • does not make use of parallel/distributed discrete-event simulation techniques. It optionally relies on distribution for the parallel execution of simulation batches. • packaged to be used on a machine must have at least 4G of memory. If memory is insufficient, the performance of the software may decrease or may corrupt the simulation process 	<ul style="list-style-type: none"> • convergence may occur in presence of long-range correlated traffic. • Understanding of scaling conditions: some emergent phenomena can be seen in sufficiently large networks, with sufficiently many traffic flows. • Need to understand relations between different abstraction levels. • Need to predict internet behavior under alternative-futures scenarios. 	<ul style="list-style-type: none"> • As GrooveNet is Open Source, every online document and online support is not available all the time 	<ul style="list-style-type: none"> • The development of TraNS is suspended. Hence, TraNS does not support the latest version of both Sumo and ns2. • Although free version of this type is available, no proper documentation is available for TraNs simulator. 	<ul style="list-style-type: none"> • Requires Cisco Input / Output images • Change the required CPU resources dynamically • not support various protocols such as Border gateway protocol and Multi-Protocol Label Switching (MPLS) 	<ul style="list-style-type: none"> • not simulate all services and functions like tunneling 	<ul style="list-style-type: none"> • Limit on the Demo of this simulator, such as the failure to modify Schematic simulations, the failure to create individual simulator schematics, the lack of storage of graphical objects, the absence of user-defined models or model details

References

- [1] Mo. Humayun Kabir, S. Islam, Md. J. Hossain and S. Hossain, "Detail Comparison of Network Simulators", International Journal of Scientific & Engineering Research, Volume 5, Issue 10, October-2014, pp 203-218
- [2] Dr. L. RAJA, "STUDY OF VARIOUS NETWORK SIMULATORS", International Research Journal of Engineering and Technology (IRJET), Vol. 5 Issue 12, Dec 2018
- [3] Irin Dorathy and M. Chandrasekarab, "Simulation tools for mobile ad hoc networks: a survey", Journal of Applied Research and Technology, Vol. 16, no. 5, oct 2018
- [4] Abdul Wahid Khorasani and Shahrooz Entezami, "Training simulation of operations with Arena 9", Marve publishing ISBN 9789642878161, 2006
- [5] Klaus Wehrle, Mesut Günes and James Gross, "Modeling and tools for network simulation" is a comprehensive and comprehensive book for the basic simulation of a network. journal springer, ISBN 978-3-642-12331-3, 2010
- [6] Asmussen, Søren, Glynn, Peter W. "Stochastic Simulation: Algorithms and Analysis". Springer. Series: Stochastic, Vol 57, 2007
- [7] Ahmad Bakhtiari Shahri and Morteza Sarangzai javan, "Simulation of computer networks by software packet tracer, along with the training of network applied concepts", open source publishing, 2009
- [8] Reza Ebrahimi Ateni, Amir Hasani Karbasi and Saman Taheri, "Specialized training simulation of computer and telecommunication networks with OPNET", Kian publishing, ISBN 9-021-307-600-978, 2016
- [9] Johann Márquez-Barja, Carlos T. Calafate, Juan-Carlos Cano, Pietro Manzoni, "An overview of vertical handover techniques: Algorithms, protocols and tools" Elsevier Journal of Computer Communications Volume 34, Issue 8, 1 June 2011, Pages 985–997
- [10] Mohammad Ashjaei, Author Vitae, Moris Behnam Author Vitae, Thomas Nolte, Elsevier, "SEtSim: A modular simulation tool for switched Ethernet networks" Journal of Systems Architecture Volume 65, April 2016, Pages 1–14
- [11] Johann Márquez-Barja, Carlos T. Calafate, Juan-Carlos Cano, Pietro Manzoni, "An overview of vertical handover techniques: Algorithms, protocols and tools" Elsevier Journal of Computer Communications Volume 34, Issue 8, 1 June 2011, Pages 901–936
- [12] Ivan Minakov, Roberto Passerone, Alessandra Rizzardi, Sabrina Sicari, "Comparative Study of Recent Wireless Sensor Network Simulators" Journal ACM Transactions on Sensor Networks (TOSN) TOSN Homepage archive Volume 12 Issue 3, August 2016 Issue-in-Progress Article No. 20
- [13] Nazmus Saquib, Md. Sabbir Rahman Sakib, Al-Sakib Khan Pathan, "ViSim: A user-friendly graphical simulation tool for performance analysis of MANET routing protocols" Elsevier Journal of Mathematical and Computer Modelling Volume 53, Issues 11–12, June 2011, Pages 2204–2218
- [14] Joel Helkey, Author Vitae, Lawrence Holder Author Vitae, Behrooz Shirazi, "Comparison of simulators for assessing the ability to sustain wireless sensor networks using dynamic network reconfiguration", Elsevier Journal of Sustainable Computing: Informatics and Systems Volume 9, March 2016, Pages 1–7
- [15] Ali Maqousi and Tatiana Balikhina, "chapter 24- Wire and Wireless Local Area Networks Simulation: OPNET Tutorial, In: "Simulation in Computer Network Design and Modeling: Use and Analysis", 2012, DOI: 10.4018/978-1-4666-0191-8.ch024, IGI Global publishing, pages 490-515
- [16] Umar Toseef and Manzoor Ahmed Khan, "OPNET Simulation Setup for QoE Based Network Selection", In: "Simulation in Computer Network Design and Modeling: Use and Analysis", 2012, DOI: 10.4018/978-1-4666-0191-8.ch006, IGI Global publishing, page 100-139
- [17] Fahmy H.M.A. "Simulators and Emulators for WSNs". In: "Wireless Sensor Networks. Signals and Communication Technology", Springer, Singapore, pp 381-491, 3 March 2016
- [18] Hossam Mahmoud Ahmad, "Signals and Communication Technology", In: "Wireless Sensor Networks", pp 381-491, Springer publisher, March 2016
- [19] S. Zhu, G. Schaefer, "Network simulation tools for supporting teaching in computer networks", In: "Simulation in Computer Network Design and Modelling: Use and Analysis", IGI Global publishing, pp. 479-489, 2012
- [20] Alfonso Ariza and Alicia Triviño, "Chapter 7- Simulation of Multihop Wireless Networks in OMNeT++", In: "Simulation in Computer Network Design and Modeling: Use and Analysis", DOI: 10.4018/978-1-4666-0191-8.ch007, IGI Global publishing, pages 140-157, 2012
- [21] Sattar J Aboud, "Chapter 21- Evaluation of Simulation Models, In: "Simulation in Computer Network Design and Modeling: Use and Analysis" DOI: 10.4018/978-1-4666-0191-8.ch021, IGI Global publishing, pages 442-458, 2012
- [22] Jafar Ababneh, Hussein Abdel-Jaber, Firas Albalas, Amjad Daoud, "Chapter 22- Analyzing and Evaluating Current Computer Networks Simulation Models", In: "Simulation in Computer Network Design and Modeling: Use and Analysis", DOI: 10.4018/978-1-4666-0191-8.ch022, IGI Global publishing, pages 459-478, 2012
- [23] Shao Ying Zhu and Gerald Schaefer "Chapter 23 -Network Simulation Tools for Supporting Teaching in Computer Networks", In: "Simulation in Computer Network Design and Modeling: Use and Analysis", DOI: 10.4018/978-1-4666-0191-8.ch023, IGI Global publishing, pages 459-478, 2012
- [24] Xiaolong Li, Meiping Peng, Jun Cai, Changyan Yi, Hong Zhang, "OPNET-based modeling and simulation of mobile Zigbee sensor networks" Peer-to-Peer Networking and Applications, March 2016, Volume 9, Issue 2, pp 414–423
- [25] Atta ur Rehman Khan, Sardar M. Bilal, Mazliza Othman, "A Performance Comparison of Network Simulators for Wireless Networks", IEEE International Conference on Control System, Computing and Engineering, Jul 2013
- [26] Sobeih, W. Chen, J.C. Hou, L. Kung, N. Li, H. Lim, H. Tyan, "J-Sim: a simulation and emulation environment for wireless sensor networks", IEEE Wireless Communications, 2006, Vol 13 p 104-119
- [27] H. Sundani, H. Li, V. K. Devabhaktuni, M. Alam, P. Bhattacharya, "Wireless Sensor Network Simulators A

- Survey and Comparisons", International Journal of Computer Networks (IJCN), 2011, Vol.2 p252-260.
- [28] Jianli Pan,"A Survey of Network Simulation Tools: Current Status and Future Developments",report, November 24, 2008
- [29] Arvind T, "A Comparative Study of Various Network Simulation Tools", International Journal of Computer Science & Engineering Technology (IJCSET), ISSN: 2229-3345, Vol. 7 No. 08 Aug 2016
- [30] Suraj G. Gupta, Mangesh M. Ghonge, Parag D. Thakare, Dr. P. M. Jawandhiya,"Open-Source Network Simulation Tools: An Overview", ISSN: 2278 – 1323 International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 2, Issue 4, April 2013
- [31] abdi Vajihe, Farzana Mohammedi and Javidan raze "A comparative study of computer network simulation software", Second National Conference on Computer, Tehran, Iran, June 2015
- [32] azadi Elham, azady elaheh and challenger "Review and compare network simulation tools and languages", The 8th Symposium on Advances in Science and Technology (8thSASTech), Mashhad, Iran,2015
- [33] Anand Nayyar, Rajeshwar Singh "A Comprehensive Review of Simulation Tools for Wireless Sensor Networks (WSNs)", Journal of Wireless Networking and Communications,5(1):19-47, January 2015 DOI: 10.5923/j.jwnc.20150501.03
- [34] Jaganath. M, Vasanth. R, Malarselvi. G "An Exhaustive Consideration Of Wired and Wireless Network Simulators",International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-8 Issue-2S4, July 2019
- [35] Rahul Mangharam, Daniel Weller, Raj Rajkumar , Priyantha Mudalige and Fan Bai "GrooveNet: A Hybrid Simulator for Vehicle-to-Vehicle Networks",Third Annual International Conference on Mobile and Ubiquitous Systems: Networking & Services, 2006
- [36] Mehdi Mekni, Bernard Moulin,"A Survey on Sensor Webs Simulation Tools ", The Second International Conference on Sensor Technologies and Applications, September 2008
- [37] Muhammad Imran ; Abas Md Said ; Halabi Hasbullah,"A survey of simulators, emulators and testbeds for wireless sensor networks", International Symposium on Information Technology, june 2010
- [38] Sunghyun Yoon, Young Boo Kim"A Design of Network Simulation Environment Using SSFNet"First International Conference on Advances in System Simulation, Sept. 2009
- [39] Bilal Ahmed, Asad Waqar Malik, Taimur Hafeez and Nadeem Ahmed,"Services and simulation frameworks for vehicular cloud computing: a contemporary survey", EURASIP Journal on Wireless Communications and Networking, 2019
- [40] Toby Schneider ; Henrik Schmidt,"NETSIM: A Realtime Virtual Ocean Hardware-in-the-loop Acoustic Modem Network Simulator",Fourth Underwater Communications and Networking Conference (UComms), Aug.2018
- Fatemeh Fakhar** received the B.S. degree in Computer Engineering from Azad University, Mashhad Branch, Iran in 2005, and M.S. degree in software Computer Engineering from Azad University, Mashhad Branch, Iran, in 2011. Currently she is a faculty member of Payame Noor University - Ahvaz Center. Her research interests include e-commerce, Biometrics and its security applications, WSN networks, Database systems, Database security, Simulations, Web mining and Data mining

A New Capacity Theorem for the Gaussian Channel with Two-sided Input and Noise Dependent State Information

Nima S. Anzabi-Nezhad*

Department of Electrical Engineering, Quchan University of Technology, Iran
nima.anzabi@qiet.ac.ir

Ghosheh Abed Hodtani

Department of Electrical Engineering, Ferdowsi University of Mashhad, Iran
ghodtani@gmail.com

Received: 10/Oct/2019

Revised: 24/Dec/2019

Accepted: 08/Jan/2020

Abstract

Gaussian interference known at the transmitter can be fully canceled in a Gaussian communication channel employing dirty paper coding, as Costa shows, when interference is independent of the channel noise and when the channel input is designed independently of the interference. In this paper, a new and general version of the Gaussian channel in presence of two-sided state information correlated to the channel input and noise is considered. Determining a general achievable rate for the channel and obtaining the capacity in a non-limiting case, we try to analyze and solve the Gaussian version of the Cover-Chiang theorem mathematically and information-theoretically. Our capacity theorem, while including all previous theorems as its special cases, explains situations that can not be analyzed by them; for example, the effect of the correlation between the side information and the channel input on the capacity of the channel that can not be analyzed with Costa's "writing on dirty paper" theorem. Meanwhile, we try to exemplify the concept of "cognition" of the transmitter or the receiver on a variable (here, the channel noise) with the information-theoretic concept of "side information" correlated to that variable and known at the transmitter or at the receiver. According to our theorem, the channel capacity is an increasing function of the mutual information of the side information and the channel noise.

Keywords: Communication channel capacity; Gaussian channel capacity; correlated side information; two sided state information; interference cancellation; dirty paper coding.

1- Introduction

Side information channels have been extensively studied since the initiation by Shannon [1] and the subsequent study by Kusnetsov-Tsybakov [2]. The capacity of a channel with side information (CSI) known non-causally only at the transmitter and only at the receiver has been determined by Gel'fand-Pinsker (GP) [3] and Heegard-El Gamal [4] respectively.

Considering the GP theorem for the Gaussian channel, Costa [5] obtained an interesting result, i.e., the channel capacity in presence of Gaussian interference known non-causally at the transmitter is the same as the case without interference. Having extended the results of Gelfand-Pinsker, Cover-Chiang [6] established a general capacity theorem for the channel with two-sided state information. There are many other important researches in the literature, e.g. [7]-[10]. The results for the single user

channel have been generalized possibly to multi user channels, at least in special cases [11]-[18].

Our Work: In this paper, we analyze the Gaussian channel with two-sided input and noise dependent state information as additive interference known at the transmitter and the receiver. The problem has three important aspects:

Information theoretic point of view: Gel'fand-Pinsker (GP) theorem [3] obtains the capacity of the channel with side information known non-causally at the transmitter. Investigating the GP theorem for channels with continuous alphabets in a special situation, Costa [5] obtained a Gaussian version of the GP theorem. As seen in (Fig. 1), the side information S_1 is considered as an additive interference and known non-causally at the transmitter. In the channel, the noise Z is independent of (X, S_1) and moreover the input X can be designed with any arbitrary correlation with the side information S_1 . Costa shows that employing dirty paper coding (DPC) in which X is designed independently of S_1 , the interference S_1 can be fully canceled, and so the channel capacity surprisingly is the capacity of the channel without interference.

The results of this paper has been presented, partially, in Iran Workshop on Communication and Information Theory, IWCIT 2015, as invited talk.

Cover-Chiang [6] analyze the channel with two-sided and correlated state information non-causally known at the transmitter and receiver and obtain the capacity theorem as the extended version of GP theorem.

The Gaussian version of the GP and Cover-Chiang theorems are open problems in information theory. In addition to Costa's "writing on dirty paper", there are many other important researches in the literature, e.g [10] and its references that studied the problem in special cases. In [10], the channel with *one-sided* additive interference (known at the transmitter) is analyzed in which the interference and noise have arbitrary joint distribution and the noise is dependent on the channel input and interference. The authors obtained a *lower bound* for the capacity of the channel.

In this paper, we try to analyze the Gaussian channel with two-sided state information as additive Gaussian interference (S_1, S_2) and known non-causally at the transmitter and receiver and dependent on the Gaussian channel noise Z and input X (Fig. 2). The random variables (X, S_1, S_2, Z) are *arbitrarily correlated*, and so the channel can be considered as a more general Gaussian version of the Cover-Chaing channel. We prove a general achievable rate (lower bound) for the channel (lemma 1), then, we obtain an upper bound for the capacity of the channel in the case that the channel input, the side information, and the channel noise, form the Markov chain $X \rightarrow (S_1, S_2) \rightarrow Z$ (lemma 2). We show the coincidence of the lower and upper bounds under this circumstance, and so establish our capacity theorem for the channel (Theorem 1). The theorem includes Costa's "writing on dirty paper" [5] (when Z is independent of (X, S_1, S_2) and X is independent of S_1 as Costa' channel) and the lower bound proved in [10] (with Gaussian noise and interference, ignoring S_2 , and X independent of S_1) as its special cases. Our theorem shows that as in [5], the effect of known interference at the transmitter and the receiver can be fully canceled employing "dirty paper-like coding" scheme.

Practical point of view: Enormous developments of wireless communications make the spectrum into one of the most precious resources in modern communications. Costa shows that employing dirty paper coding (DPC), it is possible to fully cancel known interference at the transmitter without consuming additional power, and therefore DPC is one way to utilize the spectrum efficiently by reusing it. However, Costa's "writing on dirty paper" is not applicable for the situation there exist interference known at the receiver (S_2 in Fig. 2) or random variables (X, S_1, S_2, Z) are correlated or the channel input X can not be designed independently of S_1 . One example of these situations is the cognitive interference channel in which the transmitted sequence of one transmitter is a known interference for the other transmitter and these two sequences may be dependent on

each other (for example when the sources are dependent). Some other communication scenarios in which the channel input and the side information may be correlated and the related investigations can be found in [9] and [19]. In [9] the problem of optimum transmission rate under the requirement of minimum mutual information $I(S_1^n; Y^n)$ is investigated.

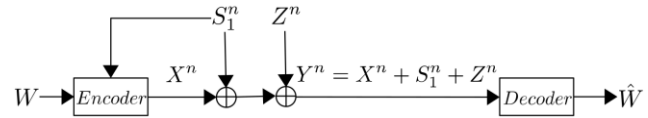


Fig. 1. Gaussian channel with additive interference known non-causally at the transmitter.

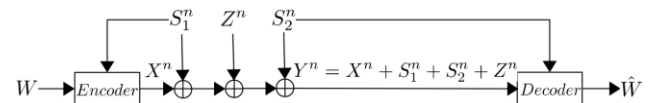


Fig. 2. Gaussian channel with side information at the transmitter and at the receiver.

Cognition on the channel noise: In this paper, we try to describe and analyze the knowledge ("cognition") of the transmitter and receiver about the channel noise Z . We consider the side information known at the transmitter S_1 (at the receiver S_2) and correlated to the channel noise Z , as the cognition of the transmitter (receiver) on the channel noise. This cognition can be perfect or imperfect. It is expected that if the side information S_1 is more correlated with Z (that means greater mutual information $I(S_1; Z)$), the transmitter acquires more knowledge about the channel noise, so employing the proper coding scheme, achieves more data rate.

Regarding the side information S_1 and S_2 as the cognition of the transmitter and the receiver on the channel noise Z , some conditions between random variables are sensible and sound. For example, it is sensible to assume that the knowledge that the transmitter got about the channel noise, gained just via the side information (S_1, S_2) . This condition can be expressed by the equation $I(X; Z|S_1, S_2) = 0$ or Markov chain $X \rightarrow (S_1, S_2) \rightarrow Z$, which is assumed in obtaining the upper bound of the capacity. Our theorem shows that the channel capacity is an increasing function of the mutual information between the side information and the channel noise $I(S_1, S_2; Z)$.

From this point of view, the subject of the knowledge is the channel noise and the transmitter and receiver acquire this knowledge via side information (S_1, S_2) . Therefore, our theorem is indeed an analysis of the effect of uncertainty about the channel noise on the capacity of the channel. In [20] and [21], we analyze the problem in some different and more limited situations.

The problem of partial channel state information studied extensively, e.g. [22]-[25]. In these papers, the subject of knowledge is the state information itself. In [25] the imperfect known state information (as a channel

interference) is partitioned to one perfect known and one unknown part. In [24] partially known two-sided state information is viewed as a disturbed state information by Gaussian noise and then the channel sensitivity to small perturbation is analyzed.

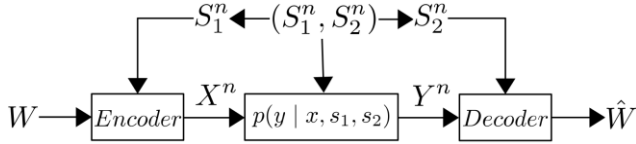


Fig. 3. Channel with side information available non-causally at the transmitter and at the receiver

In section 2, we briefly review the Cover-Chiang and the Gel'fand-Pinsker theorems and then introduce scrutiny of the Costa theorem. In section 3, we define our Gaussian channel thoroughly and present the capacity of the channel. In Section 4, we present some corollaries of the capacity theorem, some numerical comparisons and explain that how the capacity theorem can exemplify the "cognition" of the transmitter and or the receiver on the channel noise. The proofs of lower and upper bounds of the capacity are given in Section 5. Section 6 contains the conclusion. A Lemma which is used in our proofs, is given in the Appendix.

2- A Review of Previous Related Works

To clarify our approach in subsequent sections, in this section we first briefly review the Cover-Chiang capacity theorem for channels with side information available at the transmitter and at the receiver. We then review the Gel'fand-Pinsker (GP) theorem which is a special case of Cover-Chiang theorem when side information is known only at the transmitter. Finally, the Costa theorem ("writing on dirty paper" theorem), which is the Gaussian version of the GP theorem, is investigated.

Cover-Chiang Theorem

Fig.3 shows a channel with side information known at the transmitter and at the receiver where X^n and Y^n are the transmitted and the received sequences respectively. The sequences S_1^n and S_2^n are the side information known non-causally at the transmitter and at the receiver respectively. The transition probability of the channel, $p(y|x, s_1, s_2)$, depends on the input X , the side information S_1 and S_2 . It can be shown that if the channel is memoryless and the sequences (S_1^n, S_2^n) is independent and identically distributed (i.i.d.) random variables under $p(s_1, s_2)$, then the capacity of the channel is [6]:

$$C = \max_{p(u, x|s_1)} [I(U; S_2, Y) - I(U; S_1)] \quad (1)$$

where the maximum is over all distributions:

$$p(y, x, u, s_1, s_2) = p(y|x, s_1, s_2)p(u, x|s_1)p(s_1, s_2) \quad (2)$$

and U is an auxiliary random variable. It is important to note that the Markov chains:

$$S_2 \rightarrow S_1 \rightarrow UX \quad (3)$$

$$U \rightarrow XS_1S_2 \rightarrow Y \quad (4)$$

are satisfied for all distributions in (2).

Gel'fand-Pinsker Theorem

Gel'fand-Pinsker (GP) theorem [3], can be considered as a special case of (1), when there is no side information known at the receiver ($S_2 = \phi$). The capacity of the channel is:

$$C = \max_{p(u, x|s_1)} [I(U; Y) - I(U; S_1)] \quad (5)$$

for all distributions:

$$p(y, x, u, s_1) = p(y|x, s_1)p(u, x|s_1)p(s_1). \quad (6)$$

Costa's "Writing on Dirty Paper"

Costa [5] examined the Gaussian version of the channel with side information known at the transmitter (Fig. 1). As can be seen, the side information is considered as an additive interference at the receiver. Costa showed that the channel, surprisingly, has the capacity $\frac{1}{2} \log \left(1 + \frac{P}{N} \right)$, which is the same for channels with no interference S_1 . Costa derived this capacity by using the results of Gel'fand-Pinsker theorem extended to random variables with continuous alphabets. In this subsection, we first introduce the Costa assumptions and then present a proof for this theorem in such a way that it enables us to introduce our channel and develop our theorem in subsequent sections.

The channel is defined by continuous random variables $(Y, X, U, S_1) \sim f(y, x, u, s_1)$ with following properties:

- S_1^n is a sequence of Gaussian i.i.d. random variables with distribution $S_1 \sim \mathcal{N}(0, Q_1)$.
- The output is given by $Y^n = X^n + S_1^n + Z^n$, where Z^n is the sequence of white Gaussian noise with zero mean and variance N i.e. $Z \sim \mathcal{N}(0, N)$ and independent of (X, S_1) . The sequence S_1^n is non-causally known at the transmitter. The transmitted sequence X^n is assumed to have the power constraint $E\{X^2\} \leq P$.

It is readily seen that the distributions $f(y, x, u, s_1)$ having the above three properties are in the form of (6). We denote the set of all these $f(y, x, u, s_1)$'s with \mathcal{F}_C . Although for the Costa channel described above, no restriction has been imposed on the correlation between X and S_1 , in the Costa theorem, the maximum rate corresponds to independent X and S_1 , and U in form of linear combination of X and S_1 . We define \mathcal{F}'_C as a subset of \mathcal{F}_C with elements $f'(y, x, u, s_1)$ having the following properties as well as the properties mentioned before:

- X is a zero mean Gaussian random variable with the maximum variance P and independent of S_1 .
- The auxiliary random variable U takes the linear form $U = \alpha S_1 + X$.

It is clear that the set \mathcal{F}'_C and their marginal and conditional distributions are subsets of corresponding \mathcal{F}_C 's.

Achievable rate for Costa channel: From (5), when extended to memoryless channels with discrete time and continuous alphabets, we can obtain an achievable rate for the channel. The capacity of Costa channel can be written as:

$$C_{Costa} = \max_{f(u,x|s_1)} [I(U;Y) - I(U;S_1)] \quad (7)$$

where the maximum is over all $f(y,x,u,s_1)$'s in \mathcal{F}_C . Since $\mathcal{F}'_C \subseteq \mathcal{F}_C$ we have:

$$C_{Costa} \geq \max_{f'(u,x|s_1)} [I(U;Y) - I(U;S_1)] \quad (8)$$

$$= \max_{f'(u|x,s_1)f'(x|s_1)} [I(U;Y) - I(U;S_1)] \quad (9)$$

$$= \max_{\alpha} [I(U;Y) - I(U;S_1)] \quad (10)$$

The expression in the last bracket is calculated for distributions $f'(y,x,u,s_1)$ in \mathcal{F}'_C described above. Thus, defining $R(\alpha) = I(U;Y) - I(U;S_1)$, $\max_{\alpha} R(\alpha)$ is an achievable rate for the channel. $R(\alpha)$ and $\max_{\alpha} R(\alpha)$ is calculated as:

$$R(\alpha) = \frac{1}{2} \log \left(\frac{P(P+Q_1+N)}{PQ_1(1-\alpha)^2 + N(P+\alpha^2Q_1)} \right), \quad (11)$$

and

$$\max_{\alpha} R(\alpha) = R(\alpha^*) = \frac{1}{2} \log \left(1 + \frac{P}{N} \right) \quad (12)$$

where

$$\alpha^* = \frac{P}{P+N}. \quad (13)$$

Both $R(\alpha^*)$ and α^* are independent of Q_1 and then of S_1 .

Converse part of Costa theorem: From (5) we can also obtain an upper bound for the channel capacity. We have:

$$I(U;Y) - I(U;S_1) = -H(U|Y) + H(U|S_1) \quad (14)$$

$$\leq -H(U|Y, S_1) + H(U|S_1) \quad (15)$$

$$= I(U;Y|S_1) \quad (16)$$

$$\leq I(X;Y|S_1) \quad (17)$$

where inequality (15) follows from the fact that conditioning reduces the entropy and (17) follows from Markov chain $U \rightarrow XS_1 \rightarrow Y$ which is correct for all distributions $f(y,x,u,s_1)$ in the form of (6), including the distributions in the set \mathcal{F}_C . Hence we can write:

$$C_{Costa} = \max_{f(u,x|s_1)} [I(U;Y) - I(U;S_1)] \quad (18)$$

$$\leq \max_{f(x|s_1)} [I(X;Y|S_1)] \quad (19)$$

$$= \max_{f(x|s_1)} [H(Y|S_1) - H(Y|X, S_1)] \quad (20)$$

$$= \max_{f(x|s_1)} [H(X+Z|S_1) - H(Z|X, S_1)] \quad (21)$$

$$\leq \max_{f(x|s_1)} [H(X+Z) - H(Z)] \quad (22)$$

$$= \frac{1}{2} \log \left(1 + \frac{P}{N} \right), \quad (23)$$

where the inequality (22) is due to the fact that conditioning reduces the entropy and $H(Z|X, S_1) = H(Z)$ (because the channel noise Z is independent of (X, S_1) in

the channel, as defined above). The maximum in (22) is obtained when X and Z are jointly Gaussian with $E\{X^2\} = P$ because when the variance is limited, Gaussian distribution maximizes the entropy. From (12) and (23) it is seen that the lower and the upper bounds of the capacity coincide, and therefore the channel capacity is equal to $\frac{1}{2} \log \left(1 + \frac{P}{N} \right)$. It is also concluded that for the Costa channel, the optimum condition which leads to the capacity is when $X \sim \mathcal{N}(0, P)$ and independent of S_1 .

We can explain the Costa theorem more, as follows: Let consider $Y = X + S_1 + S'_1 + Z$ with independent Gaussian interference S_1 with variance Q_1 , S'_1 with variance Q'_1 and Z with variance N . If the transmitter knows nothing about this interference, then we take $U = X$ and $C = \frac{1}{2} \log \left(1 + \frac{P}{N+Q_1+Q'_1} \right)$. If S_1 is known at the transmitter, then we take $U = X + \alpha S_1$ and we have $C = \frac{1}{2} \log \left(1 + \frac{P}{N+Q'_1} \right)$ and if S_1 and S'_1 are both known at the transmitter, then $U = X + \alpha S_1 + \beta S'_1$ and $C = \frac{1}{2} \log \left(1 + \frac{P}{N} \right)$.

3- Capacity of The Gaussian Channel with Two-sided Noise and Channel Input Dependent Side Information

In this section, first, we introduce a new Gaussian channel with side information known non-causally at the transmitter and side information known non-causally at the receiver both as Gaussian additive interference at the receiver [26]. Then we present a theorem that obtains the capacity of the channel. The theorem can be considered as a Gaussian version of the Cover-Chiang unifying theorem.

3-1- Definition of The channel

Consider the Gaussian channel depicted in Fig. 2. The side information at the transmitter S_1 and at the receiver S_2 is considered as additive interference at the receiver. Our channel has three differences with Costa's one as follows:

i) In our channel, a specified correlation coefficient ρ_{XS_1} between X and S_1 , exists. Let the capacity of the channel be C . The channel with no restriction on ρ_{XS_1} (as is in the Costa channel) has the capacity C_1 [21]:

$$C_1 = \max_{\rho_{XS_1}} C \quad (24)$$

ii) To investigate the effect of the side information known at the receiver, we suppose that in our channel there exists Gaussian side information S_2 known non-causally at the receiver which is correlated to both X and S_1 .

iii) We allow the channel input X and the side information S_1 and S_2 to be correlated to the channel noise Z .

Remark: Note that assuming the input random variable X correlated to S_1 and S_2 with specific correlation coefficients, does not impose any restriction on X 's distribution and it can be proved that the distribution of X is *free to choose* [21].

1) *Definition of the channel:* The channel is defined by continuous random variables $(Y, X, U, S_1, S_2) \sim f(y, x, u, s_1, s_2)$ with following properties:

- (S_1^n, S_2^n) are i.i.d. sequences with zero mean and jointly Gaussian distributions with variance $Q_1 = \sigma_{S_1}^2$ and $Q_2 = \sigma_{S_2}^2$ respectively. The sequences S_1^n and S_2^n are non-causally known at the transmitter and at the receiver respectively.

- The output sequence $Y^n = X^n + S_1^n + S_2^n + Z^n$, where Z^n is the sequence of white Gaussian noise with zero mean and variance N .

- Random variables (X, S_1, S_2, Z) have the covariance matrix \mathbf{K} :

$$\mathbf{K} = E \begin{bmatrix} X^2 & XS_1 & XS_2 & XZ \\ XS_1 & S_1^2 & S_1S_2 & S_1Z \\ XS_2 & S_1S_2 & S_2^2 & S_2Z \\ XZ & S_1Z & S_2Z & Z^2 \end{bmatrix} = \begin{bmatrix} \sigma_X^2 & \sigma_X\sigma_{S_1}\rho_{XS_1} & \sigma_X\sigma_{S_2}\rho_{XS_2} & \sigma_X\sigma_Z\rho_{XZ} \\ \sigma_X\sigma_{S_1}\rho_{XS_1} & \sigma_{S_1}^2 & \sigma_{S_1}\sigma_{S_2}\rho_{S_1S_2} & \sigma_{S_1}\sigma_Z\rho_{S_1Z} \\ \sigma_X\sigma_{S_2}\rho_{XS_2} & \sigma_{S_1}\sigma_{S_2}\rho_{S_1S_2} & \sigma_{S_2}^2 & \sigma_{S_2}\sigma_Z\rho_{S_2Z} \\ \sigma_X\sigma_Z\rho_{XZ} & \sigma_{S_1}\sigma_Z\rho_{S_1Z} & \sigma_{S_2}\sigma_Z\rho_{S_2Z} & \sigma_Z^2 \end{bmatrix} \quad (25)$$

where ρ_{XS_1} is the correlation coefficient between X and S_1 , and so on.

In this channel, the Gaussian noise Z is not necessarily independent of the additive interference S_1 and S_2 and the input X . Moreover X^n is assumed to have the constraint $\sigma_X^2 \leq P$. Except σ_X , all other parameters in \mathbf{K} have *fixed values* specified for the channel and are considered as *the definition of the channel*.

- We assume that (X, U, S_1, S_2) form the Markov Chain $S_2 \rightarrow S_1 \rightarrow UX$. As mentioned earlier in (3), this Markov chain is satisfied by all distributions $f(y, x, u, s_1, s_2)$ in the form of (2) in Cover-Chiang capacity theorem. This Markov chain results in the weaker Markov chain $S_2 \rightarrow S_1 \rightarrow X$ and this implies that (for the proof see Lemma 3 in the Appendix):

$$\rho_{XS_2} = \rho_{XS_1}\rho_{S_1S_2} \quad (26)$$

We assume that the set of all these distributions $f(y, x, u, s_1, s_2)$ denoted with \mathcal{F} . It is readily seen that all distributions $f(y, x, u, s_1, s_2)$ in \mathcal{F} are in the form of (2). Therefore we can apply the extended version of Cover-Chiang theorem for random variables with continuous alphabets to this channel.

2) *The channel in optimum situation:* We will show that the optimum distribution resulting in maximum transmission rate, is obtained when random variables (Y, X, U, S_1, S_2) are distributed under $f^*(y, x, u, s_1, s_2)$ with following *additional* properties:

- The random variables (X, S_1, S_2) are jointly Gaussian distributed and X has zero mean and the maximum variance P , i.e. $X \sim \mathcal{N}(0, P)$.

- As in the Costa theorem [5]:

$$U = \alpha S_1 + X. \quad (27)$$

but here X and S_1 are correlated.

We assume that the set of all these distributions $f^*(y, x, u, s_1, s_2)$ denoted with \mathcal{F}^* . It is clear that the set \mathcal{F}^* and their marginals and conditional distributions are subsets of corresponding \mathcal{F} 's.

3) *Some necessary definitions:* Suppose \mathbf{K}_{opt} is the covariance matrix for random variables (X, S_1, S_2, Z) in optimum situation of the channel having all properties mentioned above; defining:

$$E_{0,i} = E\{XS_i\} = \sigma_X\sigma_{S_i}\rho_{XS_i}, \quad i = 1, 2 \quad (28)$$

$$E_{0,3} = E\{XZ\} = \sigma_X\sigma_Z\rho_{XZ} \quad (29)$$

$$E_{1,2} = E\{S_1S_2\} = \sigma_{S_1}\sigma_{S_2}\rho_{S_1S_2} \quad (30)$$

$$E_{i,3} = E\{S_iZ\} = \sigma_{S_i}\sigma_Z\rho_{S_iZ}, \quad i = 1, 2 \quad (31)$$

we can write \mathbf{K}_{opt} , and its determinant D and its minors D_1 to D_{13} as:

$$\mathbf{K}_{opt} = \begin{bmatrix} P & E_{0,1} & E_{0,2} & E_{0,3} \\ E_{0,1} & Q_1 & E_{1,2} & E_{1,3} \\ E_{0,2} & E_{1,2} & Q_2 & E_{2,3} \\ E_{0,3} & E_{1,3} & E_{2,3} & N \end{bmatrix}, \quad D = \det(\mathbf{K}_{opt}) \quad (32)$$

$$D_1 \triangleq \begin{vmatrix} Q_1 & E_{1,2} & E_{1,3} \\ E_{1,2} & Q_2 & E_{2,3} \\ E_{1,3} & E_{2,3} & N \end{vmatrix}, \quad D_2 \triangleq \begin{vmatrix} Q_1 & E_{1,2} \\ E_{1,2} & Q_2 \end{vmatrix} \quad (33-1)$$

$$D_3 \triangleq \begin{vmatrix} E_{0,1} & Q_1 & E_{1,2} \\ E_{0,2} & E_{1,2} & Q_2 \\ E_{0,3} & E_{1,3} & E_{2,3} \end{vmatrix}, \quad D_4 \triangleq \begin{vmatrix} P & E_{0,1} \\ E_{0,1} & Q_1 \end{vmatrix} \quad (33-2)$$

$$D_5 \triangleq \begin{vmatrix} P & E_{0,2} & E_{0,3} \\ E_{0,2} & Q_2 & E_{2,3} \\ E_{0,3} & E_{2,3} & N \end{vmatrix}, \quad D_6 \triangleq \begin{vmatrix} Q_2 & E_{2,3} \\ E_{2,3} & N \end{vmatrix} \quad (33-3)$$

$$D_7 \triangleq \begin{vmatrix} P & E_{0,1} & E_{0,2} \\ E_{0,2} & E_{1,2} & Q_2 \\ E_{0,3} & E_{1,3} & E_{2,3} \end{vmatrix}, \quad D_8 \triangleq \begin{vmatrix} P & E_{0,2} \\ E_{0,2} & Q_2 \end{vmatrix} \quad (33-4)$$

$$D_9 \triangleq \begin{vmatrix} P & E_{0,1} & E_{0,2} \\ E_{0,1} & Q_1 & E_{1,2} \\ E_{0,2} & E_{1,2} & Q_2 \end{vmatrix}, \quad D_{10} \triangleq \begin{vmatrix} E_{0,1} & E_{0,2} \\ E_{1,2} & Q_2 \end{vmatrix} \quad (33-5)$$

$$D_{11} \triangleq \begin{vmatrix} E_{0,1} & E_{1,2} & E_{1,3} \\ E_{0,2} & Q_2 & E_{2,3} \\ E_{0,3} & E_{2,3} & N \end{vmatrix}, \quad D_{12} \triangleq \begin{vmatrix} E_{1,2} & Q_2 \\ E_{1,3} & E_{2,3} \end{vmatrix} \quad (33-6)$$

$$D_{13}^N \triangleq \begin{vmatrix} 1 & \rho_{S_1S_2} & \rho_{S_1Z} \\ \rho_{S_1S_2} & 1 & \rho_{S_2Z} \\ \rho_{S_1Z} & \rho_{S_2Z} & 1 \end{vmatrix}, \quad D_{13} \triangleq \begin{vmatrix} E_{0,2} & Q_2 \\ E_{0,3} & E_{2,3} \end{vmatrix} \quad (33-7)$$

D_1^N , defined in (33), is the determinant of $\text{cov}(S_1, S_2, Z)$ when the variance of random variables are normalized to 1.

3-2- The Capacity Theorem for The channel

Theorem 1: The Gaussian channel defined in 3.1.1 (Fig. 2) when the channel input X , the side information (S_1, S_2) and the channel noise Z , form the Markov chain $X \rightarrow (S_1, S_2) \rightarrow Z$, has the capacity:

$$C = \frac{1}{2} \log \left(1 + \frac{P}{N} \frac{(1 - \rho_{XS_1}^2)(1 - \rho_{S_1 S_2}^2)}{D_1^N} \right) \quad (34)$$

$$= \frac{1}{2} \log \left(1 + \frac{P}{N} (1 - \rho_{XS_1}^2) \exp(2I(S_1, S_2; Z)) \right),$$

and the capacity (34) is achieved by employing the channel input X and the auxiliary random variable U as in optimum situation 3.1.2 and with

$$\alpha^* = \frac{(D_3 + D_9) - (D_7 + D_{11})}{D_1 + 2D_3 + D_9}, \quad (35)$$

the terms defined in (33).

Proof: The proof is given in Section 5.

Remark: In the Gaussian channel defined in this section, the side information (S_1, S_2) can be dependent on the channel noise. The capacity is proved with constraint of $X \rightarrow (S_1, S_2) \rightarrow Z$. As we explain in the next section, the Markov chain states that the knowledge the transmitter has got on the channel noise Z , is acquired via the side information (S_1, S_2) .

4- Interpretations of The Capacity Theorem

In this section, we examine the effect of the channel parameters on the channel capacity and explain them.

4-1-Cancellation of Interference

It is seen that with employing a DP like coding scheme, interference S_1 and S_2 can be fully canceled, as in Costa's writing on dirty paper.

4-2- The Effect of ρ_{XS_1}

Corollary 1: If there is no specific correlation between X and S_1 , as Costa's dirty paper, the capacity is achieved when the channel input X is designed independently of the known S_1 :

$$C = \max_{\rho_{XS_1}} \left(\frac{1}{2} \log \left(1 + \frac{P}{N} (1 - \rho_{XS_1}^2) \exp(2I(S_1, S_2; Z)) \right) \right)$$

$$= \frac{1}{2} \log \left(1 + \frac{P}{N} \exp(2I(S_1, S_2; Z)) \right). \quad (36)$$

Corollary 2: If we assume that the channel noise Z is independent of (X, S_1, S_2) , from (34), the capacity of the channel is:

$$C = \frac{1}{2} \log \left(1 + \frac{P}{N} (1 - \rho_{XS_1}^2) \right) \quad (37)$$

From (24), C is reduced to the Costa capacity $\frac{1}{2} \log \left(1 + \frac{P}{N} \right)$ by maximizing it with $\rho_{XS_1} = 0$.

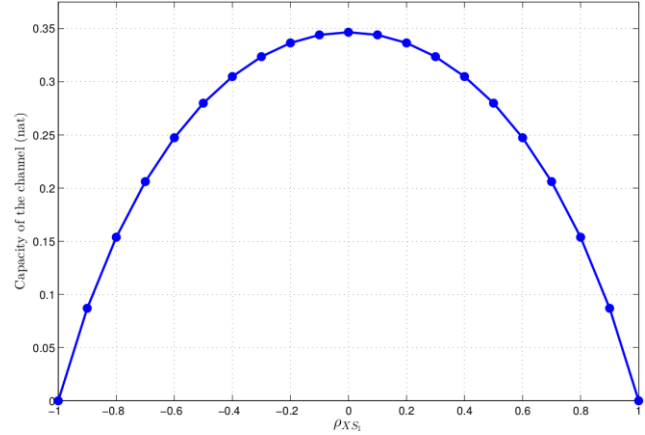


Fig. 4. Capacity of the channel with respect to ρ_{XS_1} when the channel noise Z is independent of (X, S_1, S_2) and with signal to noise ratio $\frac{P}{N} = 1$

Corollary 3: It is seen that in the case the side information S_2 is independent of the channel noise Z , the capacity of the channel is equal to the capacity when there is no interference S_2 . In other words, in this case, the receiver can subtract the known S_2^n from the received Y^n without losing any worthy information.

Corollary 4: The correlation between X and S_1 decreases the capacity of the channel. It can be explained as follows: by looking at $Y = X + S_1 + Z$ in our dirty paper like coding, mitigating the input-dependent interference effect, also mitigates the input power impact on the channel capacity as this fact is seen in (37) as $\sigma_X^2(1 - \rho_{XS_1}^2)$.

As an extreme and interesting case, when $S_1 = X$ (then $\rho_{XS_1} = 1$), according to the usual Gaussian coding, the capacity seems to be $\frac{1}{2} \log \left(1 + \frac{4P}{N} \right)$, which is the capacity when $2X$ is transmitted and $Y = 2X + Z$ is received. But as our theorem shows, the capacity paradoxically is zero. It can be explained as follows: the receiver, based on his information, ought to decode according to the dirty paper like coding. In DP like coding, with given known sequence $S_{1,0}^n$, we find an auxiliary sequence U^n like U_0^n jointly typical with $S_{1,0}^n$ [5]. Jointly typicality of $(U_0^n, S_{1,0}^n)$ is equivalent to:

$$\left| (U_0^n - \alpha^* S_{1,0}^n)^T S_{1,0}^n \right| \leq \delta, \quad \delta \text{ small} \quad (38)$$

where \cdot^T denotes the transpose operation and α^* is computed according to (35). If $X = S_1$, there exists no such U_0^n : since $X_0^n = U_0^n - \alpha^* S_{1,0}^n = S_{1,0}^n$, we have

$$\left| (U_0^n - \alpha^* S_{1,0}^n)^T S_{1,0}^n \right| = \|S_{1,0}^n\|^2 \quad (39)$$

where $\|S_{1,0}^n\|$ is the norm of the given known sequence $S_{1,0}^n$ and therefore (38) can not be true. In other words, in this case, encoding error occurs.

Fig. 4 shows the variation of the capacity C in (37) with respect to ρ_{XS_1} when $\frac{P}{N} = 1$. It is seen that when the correlation between the channel input and the side information known at the transmitter increases, the channel capacity decreases. The maximum capacity is gained when $\rho_{XS_1} = 0$, which is Costa's capacity. Fig. 5 shows the capacity C in (37) with respect to SNR for five values of ρ_{XS_1} .

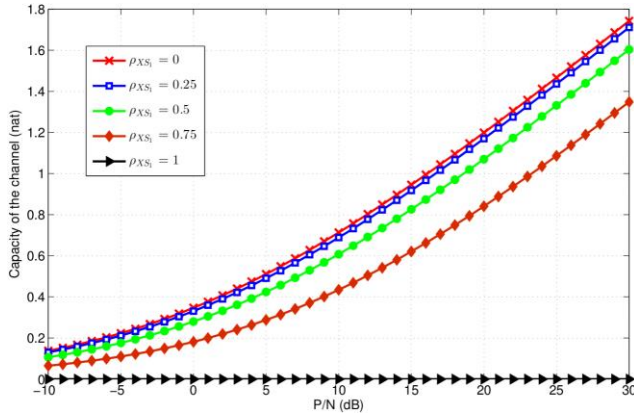


Fig. 5. Capacity of the channel with respect to SNR when the channel noise Z is independent of (X, S_1, S_2)

4-3- Cognition of Transmitter and Receiver on the Channel Noise

It is seen that the mutual information $I(S_1, S_2; Z)$ increases the channel capacity. For the sake of simplicity, we ignore the effect of ρ_{XS_1} and assume the channel capacity is given by (36).

If we suppose that $S_2 = \phi$, the capacity is given by:

$$C = \frac{1}{2} \log \left(1 + \left(\frac{P}{N} \right) \exp(2I(S_1; Z)) \right) \quad (40)$$

$$= \frac{1}{2} \log \left(1 + \frac{P}{N} \frac{1}{(1 - \rho_{S_1 Z}^2)} \right). \quad (41)$$

It is seen that more correlation between the side information S_1 and the channel noise Z , results in more capacity. It can be explained by the fact that the side information known at the transmitter and correlated with the channel noise Z , carries knowledge about Z for transmitter, so enhances the transmitter's ability to cancel the channel noise. When $\rho_{S_1 Z} = \pm 1$, the transmitter has got perfect knowledge about Z and the capacity reaches to infinite. Fig. 6 illustrates the capacity of the channel with respect to $\rho_{S_1 Z}$ when $\frac{P}{N} = 1$. Fig. 7 shows the capacity of the channel with respect to SNR for five values of $\rho_{S_1 Z}$.

The same situation comes about when $S_1 = \phi$: the side information S_2 known at the receiver and correlated with the Z , carries knowledge about Z for the receiver. More correlation between S_2 and Z results in more capacity. This shows the significance of the known additive interference at the receiver and the reason why subtracting S_2^n from received Y^n is a wrong decoding strategy.

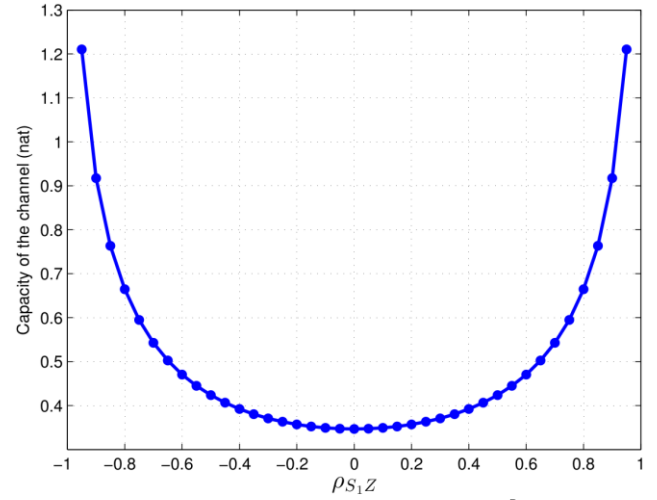


Fig. 6. Capacity of the channel with respect to $\rho_{S_1 Z}$ when $\frac{P}{N} = 1$

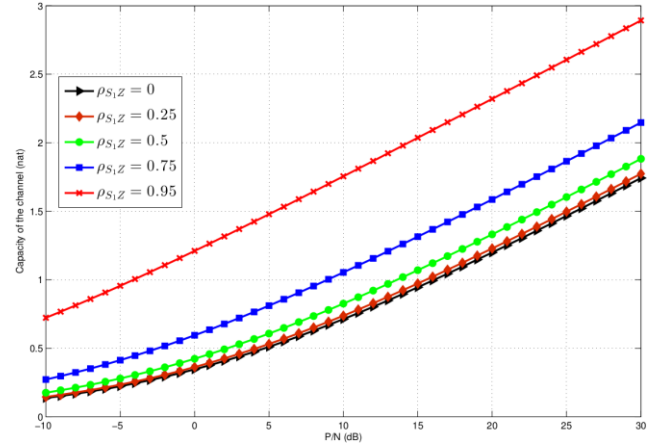


Fig. 7. Capacity of the channel with respect to SNR for five values of $\rho_{S_1 Z}$

If the side information at the transmitter and at the receiver (S_1, S_2) exists and is correlated to the channel noise Z , the capacity increases by $I(S_1, S_2; Z)$. Fig. 8 illustrates the capacity of the channel with respects to mutual information $I(S_1, S_2; Z)$ for five values of SNR.

If the signal to noise ratio is large enough, the capacity can be written as:

$$C = \frac{1}{2} \log \left(1 + \frac{P}{N} \right) + I(S_1, S_2; Z). \quad (42)$$

that shows the major effect of $I(S_1, S_2; Z)$ on the channel capacity.

Remark: The capacity of the channel and the effect of S_1 and S_2 on the capacity, reveals the cognitive role of the side information which is known at the transmitter or at the receiver and is correlated to the channel noise. As it is seen in this section, side information known at the transmitter (or receiver), carries the knowledge about the channel noise if it is correlated with the channel noise. If we regard the side information S_1 and S_2 as the cognition of the transmitter and the receiver on the channel noise Z , some conditions between random variables in our model are sensible and sound. For example, it is sensible to assume that the knowledge that the transmitter got about the channel noise, gained just via the side information (S_1, S_2) . This conditions can be expressed by equation $I(X; Z|S_1, S_2) = 0$ or Markov chain $X \rightarrow (S_1, S_2) \rightarrow Z$, which is assumed in our theorem.

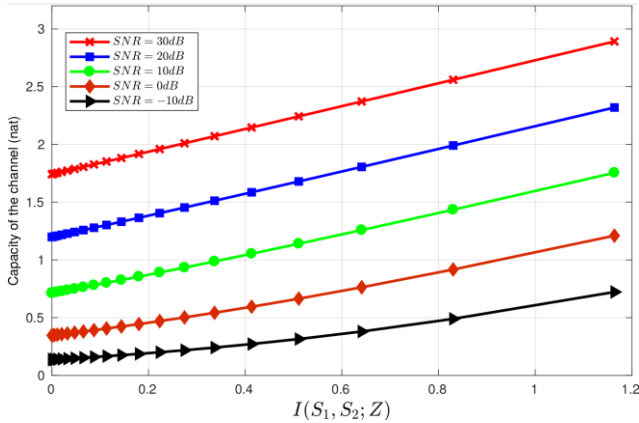


Fig. 8. Capacity of the channel with respect to $I(S_1, S_2; Z)$ for five values of SNR.

5- Proof of Theorem 1

To *prove* the theorem, first, we prove a general achievable rate for the channel. Then we obtain an upper bound for the capacity of the channel when we have the Markov chain $X \rightarrow (S_1, S_2) \rightarrow Z$. Then we show the coincidence of this upper bound with the lower bound of the capacity.

Lemma 1. Lower Bound of the Capacity: The channel defined in 3.1.1 has the lower bound R_G in (43), where D_1^N is defined in (33):

$$R_G = \frac{1}{2} \log \left(1 + \frac{[\sigma_X(1 - \rho_{X S_1}^2) - \sigma_Z(\rho_{X S_1} \rho_{S_1 Z} - \rho_{X Z})]^2 (1 - \rho_{S_1 S_2}^2)}{\sigma_Z^2 \left((1 - \rho_{X S_1}^2) D_1^N - (\rho_{X S_1} \rho_{S_1 Z} - \rho_{X Z})^2 (1 - \rho_{S_1 S_2}^2) \right)} \right) \quad (43)$$

Corollary 5: The lower bound (43) includes the lower bound obtained in [10]. If $\rho_{X S_1} = \rho_{S_1 S_2} = 0$ as in [10], then $D_1^N = 1 - \rho_{X Z}^2$ and we have

$$R_G = \frac{1}{2} \log \left(1 + \frac{(1 + \rho_{X Z} \frac{\sigma_Z}{\sigma_X})^2 \sigma_X^2}{1 - \rho_{X Z}^2 - \rho_{S_1 Z}^2 \frac{\sigma_X^2}{\sigma_Z^2}} \right), \quad (44)$$

which is the lower bound in [10].

Proof of Lemma 1: Using the extension of Cover-Chiang capacity theorem given in (1) for random variables with continuous alphabets, the capacity of our channel can be written as:

$$C = \max_{f(u, x|s_1)} [I(U; Y, S_2) - I(U; S_1)] \quad (45)$$

where the maximum is over all distributions $f(y, x, u, s_1, s_2)$ in \mathcal{F} defined in 3.1.1. Since $\mathcal{F}^* \subseteq \mathcal{F}$ we have:

$$C \geq \max_{f^*(u, x|s_1)} [I(U; Y, S_2) - I(U; S_1)] \quad (46)$$

$$= \max_{\alpha} [I(U; Y, S_2) - I(U; S_1)] \quad (47)$$

where the expression $I(U; Y, S_2) - I(U; S_1)$ in (47) is calculated for the distributions in \mathcal{F}^* defined in 3.1.2. Thus, defining $R(\alpha) = I(U; Y, S_2) - I(U; S_1)$, we have:

$$C \geq \max_{\alpha} R(\alpha) = R(\alpha^*), \quad (48)$$

therefore $R(\alpha^*)$ is a lower bound for the channel capacity.

To compute $R(\alpha^*)$, we write:

$$I(U; Y, S_2) = H(U) + H(Y, S_2) - H(U, Y, S_2) \quad (49)$$

and

$$I(U; S_1) = H(U) + H(S_1) - H(U, S_1), \quad (50)$$

For $H(Y, S_2)$ we have:

$$H(Y, S_2) = \frac{1}{2} \log \left((2\pi e)^2 \det(\text{cov}(Y, S_2)) \right) \quad (51)$$

where

$$\text{cov}(Y, S_2) = [e_{ij}]_{2 \times 2}$$

and

$$\begin{aligned} e_{11} &= P + Q_1 + Q_2 + N + 2E_{0,1} + 2E_{0,2} + 2E_{1,2} \\ &\quad + 2E_{0,3} + 2E_{1,3} + 2E_{2,3} \\ e_{12} &= e_{21} = E_{0,2} + E_{1,2} + Q_2 + E_{2,3} \\ e_{22} &= Q_2. \end{aligned}$$

After computing we have:

$$\det(\text{cov}(Y, S_2)) = D_2 + D_6 + D_8 + 2D_{10} - 2D_{12} - 2D_{13}, \quad (52)$$

where the terms are defined in (28)-(33).

For $H(U, Y, S_2)$ we have:

$$H(U, Y, S_2) = \frac{1}{2} \log \left((2\pi e)^3 \det(\text{cov}(U, Y, S_2)) \right) \quad (53)$$

where

$$\text{cov}(U, Y, S_2) = [e_{ij}]_{3 \times 3} \quad (54)$$

and

$$\begin{aligned} e_{11} &= P + \alpha^2 Q_1 + 2\alpha E_{0,1} \\ e_{12} &= e_{21} = P + (\alpha + 1)E_{0,1} + \alpha Q_1 + \alpha E_{1,2} + \alpha E_{1,3} \\ &\quad + E_{0,2} + E_{0,3}, \\ e_{13} &= e_{31} = \alpha E_{1,2} + E_{0,2} \\ e_{22} &= P + Q_1 + Q_2 + N + 2E_{0,1} + 2E_{0,2} + 2E_{1,2} \\ &\quad + 2E_{0,3} + 2E_{1,3} + 2E_{2,3} \end{aligned}$$

$$\begin{aligned} e_{23} &= e_{32} = E_{0,2} + E_{1,2} + Q_2 + E_{2,3} \\ e_{33} &= Q_2. \end{aligned}$$

After manipulations we have:

$$\det(\text{cov}(U, Y, S_2)) = \alpha^2 D_1 + 2\alpha(\alpha - 1)D_3 + 2(\alpha - 1)D_7 + (\alpha - 1)^2 D_9 + 2\alpha D_{11} + D_5 \quad (55)$$

For $H(S_1)$ and $H(U, S_1)$ we have:

$$H(S_1) = \frac{1}{2} \log((2\pi e)Q_1). \quad (56)$$

and

$$H(U, S_1) = \frac{1}{2} \log((2\pi e)^2 \det(\text{cov}(U, S_1))) \quad (57)$$

where

$$\text{cov}(U, S_1) = \begin{bmatrix} \alpha^2 Q_1 + P + 2\alpha E_{0,1} & \alpha Q_1 + E_{0,1} \\ \alpha Q_1 + E_{0,1} & Q_1 \end{bmatrix} \quad (58)$$

and its determinant:

$$\det(\text{cov}(U, S_1)) = D_4. \quad (59)$$

Substituting (51), (53), (56) and (57) in (49) and (50), we obtain $R(\alpha)$ as in (60):

$$R(\alpha) = \frac{1}{2} \log \left(\frac{D_4[D_8 + D_2 + D_6 + 2D_{10} - 2D_{12} - 2D_{13}]}{Q_1[(\alpha - 1)^2 D_9 + \alpha^2 D_1 + 2\alpha(\alpha - 1)D_3 + 2\alpha D_{11} + 2(\alpha - 1)D_7 + D_5]} \right) \quad (60)$$

The optimum value of α corresponding to maximum of $R(\alpha)$ is easily obtained as:

$$\alpha^* = \frac{(D_3 + D_9) - (D_7 + D_{11})}{D_1 + 2D_3 + D_9}, \quad (61)$$

Substituting α^* from (61) into (60) and using the equations (26), (28)-(31) and (33) we finally conclude that $R(\alpha^*)$ equals R_G in (43). Therefore R_G in (43) is a lower bound for the capacity of the channel defined in 3.1.1 (details of computations are omitted for the brevity).

Q.E.D

Lemma 2. Upper Bound of the Capacity: The capacity of the Gaussian channel defined in 3.1.1, when the channel input X , the side information (S_1, S_2) and the channel noise Z form the Markov chain $X \rightarrow (S_1, S_2) \rightarrow Z$, has the upper bound C in (34).

Proof of Lemma 2: First, we note that the Markov chain $X \rightarrow (S_1, S_2) \rightarrow Z$ and the Markov chain $S_2 \rightarrow S_1 \rightarrow X$ in (3), imply the weaker Markov chain $X \rightarrow S_1 \rightarrow Z$. And it can be proved that this Markov chain implies (for proof see the Appendix):

$$\rho_{XZ} = \rho_{XS_1} \rho_{S_1Z}. \quad (62)$$

For all distributions $f(y, x, u, s_1, s_2)$ in \mathcal{F} defined in 3.1.1, we have:

$$I(U; Y, S_2) - I(U; S_1) = -H(U|Y, S_2) + H(U|S_1) \quad (63)$$

$$\leq -H(U|Y, S_1, S_2) + H(U|S_1) \quad (64)$$

$$= -H(U|Y, S_1, S_2) + H(U|S_1, S_2) \quad (65)$$

$$= I(U; Y|S_1, S_2) \quad (66)$$

$$\leq I(X; Y|S_1, S_2) \quad (67)$$

where (64) follows from the fact that conditioning reduces entropy, (65) follows from Markov chain $S_2 \rightarrow$

$S_1 \rightarrow UX$ and (67) from Markov chain $U \rightarrow XS_1S_2 \rightarrow Y$ which are satisfied for any distribution in the form of (2), including the distributions in the set \mathcal{F} . From (1) and (67) we can write:

$$C = \max_{f(u, x|s_1)} [I(U; Y, S_2) - I(U; S_1)] \quad (68)$$

$$\leq \max_{f(x|s_1)} [I(X; Y|S_1, S_2)]. \quad (69)$$

From (69) it is seen that the capacity of the channel cannot be greater than the capacity when both S_1 and S_2 are available at both the transmitter and the receiver, which is physically predictable. To compute (69) we write:

$$I(X; Y|S_1, S_2) = H(Y|S_1, S_2) - H(Y|X, S_1, S_2) \quad (70)$$

$$= H(X + S_1 + S_2 + Z|S_1, S_2) \quad (71)$$

$$- H(X + S_1 + S_2 + Z|X, S_1, S_2) \quad (72)$$

$$= H(X + Z|S_1, S_2) - H(Z|X, S_1, S_2) \quad (73)$$

$$= H(X + Z|S_1, S_2) - H(Z|S_1, S_2) \quad (74)$$

$$= H((X + Z), S_1, S_2) - H(S_1, S_2, Z), \quad (74)$$

where (73) follows from the Markov chain $X \rightarrow (S_1, S_2) \rightarrow Z$. Hence, the maximum value in (69) occurs when $H((X + Z), S_1, S_2)$ is maximum. Since S_1, S_2 and Z are Gaussian, the maximum in (69) is achieved when (X, S_1, S_2) are jointly Gaussian and X has its maximum variance P ; in other words, $I(X; Y|S_1, S_2)$ is computed for distribution $f^*(y, x, s_1, s_2)$ defined in 3.1.2. Let $I^*(X; Y|S_1, S_2)$ be the maximum value in (69). We have:

$$C \leq I^*(X; Y|S_1, S_2) \quad (75)$$

To compute $I^*(X; Y|S_1, S_2)$, we first compute $H((X + Z), S_1, S_2)$ for distribution $f^*(y, x, s_1, s_2)$ defined in 3.1.2:

$$H((X + Z), S_1, S_2) = \frac{1}{2} \log \left((2\pi e)^3 \det(\text{cov}((X + Z), S_1, S_2)) \right) \quad (76)$$

where $\text{cov}((X + Z), S_1, S_2)$ is

$$E \left\{ \begin{bmatrix} (X + Z)^2 & (X + Z)S_1 & (X + Z)S_2 \\ (X + Z)S_1 & S_1^2 & S_1S_2 \\ (X + Z)S_2 & S_1S_2 & S_1^2 \end{bmatrix} \right\} = \begin{bmatrix} P + N + 2E_{0,3} & E_{0,1} + E_{1,3} & E_{0,2} + E_{2,3} \\ E_{0,1} + E_{1,3} & Q_1 & E_{1,2} \\ E_{0,2} + E_{2,3} & E_{1,2} & Q_2 \end{bmatrix} \quad (77)$$

and the determinant:

$$\det(\text{cov}((X + Z), S_1, S_2)) = D_1 + 2D_3 + D_9, \quad (78)$$

and the other term in (74):

$$H(S_1, S_2, Z) = \frac{1}{2} \log((2\pi e)^3 D_1) \quad (79)$$

Substituting (78) in (76), and from (79), we have:

$$I^*(X; Y|S_1, S_2) = \frac{1}{2} \log \left(1 + \frac{D_9 + 2D_3}{D_1} \right). \quad (80)$$

Rewriting (80) in terms of $\sigma_X, \sigma_{S_1}, \sigma_{S_2}, \sigma_Z, \rho_{XS_1}, \rho_{S_1Z}, \rho_{S_2Z}$ and $\rho_{S_1S_2}$ using (28)-(31) and (33) and taking into account two Markovity results (26) and (62), we finally conclude that (details of manipulations are omitted for the brevity):

$$I^*(X; Y|S_1, S_2) = \frac{1}{2} \log \left(1 + \frac{P(1-\rho_{XS_1}^2)(1-\rho_{S_1S_2}^2)}{D_1^N} \right). \quad (81)$$

Hence, C in (34) is an upper bound for the capacity of the channel when we have the Markov chain $X \rightarrow (S_1, S_2) \rightarrow Z$.

Q.E.D

For *completing* the proof of the capacity theorem, it is enough to compute the lower bound of the channel (43), when we have the Markov chain $X \rightarrow (S_1, S_2) \rightarrow Z$. Applying the equation (3) to (43), shows the coincidence of the upper and the lower bounds of the capacity of the channel in this case and considering:

$$I(S_1, S_2; Z) = \frac{1}{2} \log \left(\frac{1-\rho_{S_1S_2}^2}{D_1^N} \right) \quad (82)$$

the proof is completed.

Q.E.D

6- Conclusion

By fully *detailed* investigating the Gaussian channel in presence of two-sided input and noise dependent state information, we obtained a general achievable rate for the channel and established the capacity theorem. This capacity theorem first demonstrates the impact of the transmitter and receiver cognition on the capacity and second shows the effect of the correlation between the channel input and side information available at the transmitter and at the receiver on the channel capacity. Whereas, as expected, the cognition of the transmitter and receiver increases the capacity, the correlation between the channel input and the side information known at the transmitter decreases it.

7- Appendix

Lemma 3: Consider three zero mean random variables (X, S_1, S_2) with covariance matrix \mathbf{K} as:

$$\mathbf{K} = E \left\{ \begin{bmatrix} X^2 & XS_1 & XS_2 \\ XS_1 & S_1^2 & S_1S_2 \\ XS_2 & S_1S_2 & S_2^2 \end{bmatrix} \right\}$$

$$= \begin{bmatrix} \sigma_X^2 & \sigma_X\sigma_{S_1}\rho_{XS_1} & \sigma_X\sigma_{S_2}\rho_{XS_2} \\ \sigma_X\sigma_{S_1}\rho_{XS_1} & \sigma_{S_1}^2 & \sigma_{S_1}\sigma_{S_2}\rho_{S_1S_2} \\ \sigma_X\sigma_{S_2}\rho_{XS_2} & \sigma_{S_1}\sigma_{S_2}\rho_{S_1S_2} & \sigma_{S_2}^2 \end{bmatrix} \quad (83)$$

Suppose (S_1, S_2) are *jointly Gaussian* random variables. Then, if (X, S_1, S_2) form Markov chain $S_2 \rightarrow S_1 \rightarrow X$, (even if X is not *Gaussian*) we have:

$$\rho_{XS_2} = \rho_{XS_1}\rho_{S_1S_2} \quad (84)$$

or equivalently:

$$E\{S_1^2\}E\{XS_2\} = E\{XS_1\}E\{S_1S_2\} \quad (85)$$

Proof of Lemma 3: we can write:

$$\rho_{XS_2} = \frac{E\{XS_2\}}{\sigma_X\sigma_{S_2}} = \frac{E\{E\{XS_2|S_1\}\}}{\sigma_X\sigma_{S_2}} \quad (86)$$

$$= \frac{E\{E\{X|S_1\}E\{S_2|S_1\}\}}{\sigma_X\sigma_{S_2}} \quad (87)$$

$$= \frac{\rho_{S_1S_2}}{\sigma_X\sigma_{S_1}} E\{S_1E\{X|S_1\}\} \quad (88)$$

$$= \frac{\rho_{S_1S_2}}{\sigma_X\sigma_{S_1}} E\{XS_1\} \quad (89)$$

$$= \rho_{XS_1}\rho_{S_1S_2} \quad (90)$$

where (87) follows from the Markov chain $S_2 \rightarrow S_1 \rightarrow X$ and (88) follows from Gaussianness of (S_1, S_2) and the fact that $E\{S_2|S_1\} = \frac{\sigma_{S_2}\rho_{S_1S_2}}{\sigma_{S_1}}S_1$ and (89) follows from the general rule that for random variables A and B we have $E\{g_1(A)g_2(B)\} = E\{g_1(A)E\{g_2(B)|A\}\}$ [27, p.234].

Q.E.D

References

- [1] C. E. Shannon, "Channels with side information at the transmitter," IBM Journal of Research and Development, vol. 2, no. 4, pp. 289–293, oct. 1958.
- [2] A. V. Kosnetsov and B. S. Tsybakov, "Coding in a memory with defective cells," Probl. Pered. Inform., vol. 10, no. 2, pp. 52–60, Apr./Jun. 1974. Translated from Russian.
- [3] S. I. Gel'fand and M. S. Pinsker, "Coding for channel with random parameters," Probl. Contr. Inform. Theory, vol. 9, no. 1, pp. 19–31, 1980.
- [4] C. Heegard and A. El Gamal, "On the capacity of computer memory with defects," Information Theory, IEEE Transactions on, vol. 29, no. 5, pp. 731–739, sep 1983.
- [5] M. Costa, "Writing on dirty paper (corresp.)," Information Theory, IEEE Transactions on, vol. 29, no. 3, pp. 439–441, may 1983.
- [6] T. M. Cover and M. Chiang, "Duality between channel capacity and rate distortion with two-sided state information," Information Theory, IEEE Transactions on, vol. 48, no. 6, pp. 1629–1638, jun 2002.
- [7] S. Jafar, "Capacity with causal and noncausal side information: A unified view," Information Theory, IEEE Transactions on, vol. 52, no. 12, pp. 5468–5474, dec. 2006.
- [8] G. Keshet, Y. Steinberg, and N. Merhav, "Channel coding in the presence of side information," Found. Trends Commun. Inf. Theory, vol. 4, pp. 445–586, June 2008.
- [9] N. Merhav and S. Shamai, "Information rates subject to state masking," Information Theory, IEEE Transactions on, vol. 53, no. 6, pp. 2254–2261, june 2007.
- [10] I. Bergel, D. Yellin, and S. Shamai, "A lower bound on the data rate of dirty paper coding in general noise and interference," IEEE Wireless Communications Letters, vol. 3, no. 4, pp. 417–420, 2014.
- [11] Y. Steinberg, "Coding for the degraded broadcast channel with random parameters, with causal and noncausal side

- information,” *Information Theory, IEEE Transactions on*, vol. 51, no. 8, pp. 2867–2877, aug. 2005.
- [12] S. Sigurjonsson and Y.-H. Kim, “On multiple user channels with state information at the transmitters,” in *Information Theory, 2005. ISIT 2005. Proceedings. International Symposium on*, sept. 2005, pp. 72–76.
- [13] Y. H. Kim, A. Sutivong, and S. Sigurjonsson, “Multiple user writing on dirty paper,” in *Information Theory, 2004. ISIT 2004. Proceedings. International Symposium on*, june-2 july 2004, p. 534.
- [14] T. Philosof and R. Zamir, “On the loss of single-letter characterization: The dirty multiple access channel,” *Information Theory, IEEE Transactions on*, vol. 55, no. 6, pp. 2442–2454, june 2009.
- [15] Y. Steinberg and S. Shamai, “Achievable rates for the broadcast channel with states known at the transmitter,” in *Information Theory, 2005. ISIT 2005. Proceedings. International Symposium on*, sept. 2005, pp. 2184–2188.
- [16] R. Duan, Y. Liang, and S. Shamai, “State-dependent gaussian interference channels: Can state be fully canceled?” *IEEE Transactions on Information Theory*, vol. 62, no. 4, pp. 1957–1970, 2016.
- [17] Y. Sun, Y. Liang, R. Duan, and S. S. Shitz, “State-dependent z-interference channel with correlated states,” in *2017 IEEE International Symposium on Information Theory (ISIT)*, 2017, pp. 644–648.
- [18] Y. Sun, R. Duan, Y. Liang, and S. Shamai Shitz, “State-dependent interference channel with correlated states,” *IEEE Transactions on Information Theory*, vol. 65, no. 7, pp. 4518–4531, 2019.
- [19] Y.-C. Huang and K. R. Narayanan, “Joint source-channel coding with correlated interference,” in *Information Theory Proceedings (ISIT), 2011 IEEE International Symposium on*, 2011, pp. 1136–1140.
- [20] N. S. Anzabi-Nezhad, G. A. Hodtani, and M. Molavi Kakhki, “Information theoretic exemplification of the receiver recognition and a more general version for the Costa theorem,” *IEEE Communication Letters*, vol. 17, no. 1, pp. 107–110, 2013.
- [21] —, “A more general version of the costa theorem,” *Journal of Communication Engineering*, vol. 2, no. 4, pp. 1–19, Autumn 2013.
- [22] B. Chen, S. C. Draper, and G. Wornell, “Information embedding and related problems: Recent results and applications,” in *Allerton Conference, USA, 2001*.
- [23] A. Rosenzweig, Y. Steinberg, and S. Shamai, “On channels with partial channel state information at the transmitter,” *Information Theory, IEEE Transactions on*, vol. 51, no. 5, pp. 1817–1830, may 2005.
- [24] A. Zaidi and P. Duhamel, “On channel sensitivity to partially known two-sided state information,” in *Communications, 2006. ICC '06. IEEE International Conference on*, vol. 4, june 2006, pp. 1520–1525.
- [25] L. Gueguen and B. Sayrac, “Sensing in cognitive radio channels: A theoretical perspective,” *Wireless Communications, IEEE Transactions on*, vol. 8, no. 3, pp. 1194–1198, march 2009.
- [26] G. A. Hodtani, “The effect of transceiver recognition on the gaussian channel capacity,” in *2015 Iran Workshop on Communication and Information Theory (IWCIT)*, May 2015, pp. 1–3.
- [27] A. Papoulis and S. U. Pillai, *Probability, Random Variables, and Stochastic Processes*, 4th ed. 1em plus 0.5em minus 0.4em McGraw-Hill, 2002.

Nima Anzabi-Nezhad received the B.S. degree in electronics engineering from Ferdowsi University of Mashhad, Mashhad, Iran in 1995, and M.S. degree in philosophy of science from Sharif University of Technology, Tehran, Iran in 1999. He received the Ph.D. degree in communication engineering in 2014 from Ferdowsi University of Mashhad, Mashhad, Iran. He is an assistant professor in Department of Electrical Engineering at Quchan University of Technology, Quchan, Iran, since 2014. His research interests include information theory, communication theory and cryptography.

Ghosheh Abed Hodtani received the B.Sc. degree in electronics engineering and the M.Sc. degree in communications engineering, both from Isfahan University of Technology, Isfahan, Iran, in 1985, 1987, respectively. He joined Electrical Engineering Dept., at Ferdowsi University of Mashhad, Mashhad, Iran, in 1987. He decided to pursue his studies in 2005 and received the Ph.D. degree (with excellent grade) from Sharif University of Technology, Tehran, Iran, in 2008; and he is a full professor in electrical engineering since 2016, and has been selected as a distinguished engineering professor by Ferdowsi Academic Foundation in 2016. His research interests are in multi-user information theory, communication theory, wireless communications, information-theoretic learning and signal processing. Prof. Hodtani is the author of a textbook on electrical circuits, the winner of the best paper award at IEEE ICT -2010 and a member of technical program and steering committees of Iran workshop on Communication and Information Theory (IWCIT).

Embedding Virtual Machines in Cloud Computing Based on Big Bang–Big Crunch Algorithm

Afshin Mahdavi

Afshinmahdavi96@yahoo.com

Department of Computer Engineering, Tabriz branch, Islamic Azad University, Tabriz, Iran

Ali Ghaffari*

Department of Computer Engineering, Tabriz branch, Islamic Azad University, Tabriz, Iran

A.Ghaffari@iaut.ac.ir

Received: 27/Dec/2019

Revised: 29/Apr/2020

Accepted: 09/May/2020

Abstract

Cloud computing is becoming an important and adoptable technology for many of the organization which requires a large amount of physical tools. In this technology, services are provided and presented according to users' requests. Due to the presence of a large number of data centers in cloud computing, power consumption has recently become an important issue. However, data centers hosting Cloud applications consume huge amounts of electrical energy and contributing to high operational costs to the environment. Therefore, we need Green Cloud computing solutions that can not only minimize operational costs but also reduce the environmental impact. Live migration of virtual machines and their scheduling and embedding lead to enhanced efficiency of dynamic resources. The guarantee of service quality and service reliability is an indispensable and irrevocable requirement with respect to service level agreement. Hence, providing a method for reducing costs of power consumption, data transmission, bandwidth and, also, for enhancing quality of service (QoS) in cloud computing is critical. In this paper, a Big Bang–Big Crunch (BB-BC) based algorithm for embedding virtual machines in cloud computing was proposed. We have validated our approach by conducting a performance evaluation study using the CloudSim toolkit. Simulation results indicate that the proposed method not only enhances service quality, thanks to the reduction of agreement violation, but also reduces power consumption.

Keywords: Cloud computing; Virtual machine; Big Bang–Big Crunch algorithm; Energy; Service level agreement.

1- Introduction

Cloud can be defined as a new computing technology and paradigm that provides scalable, on-demand, and virtualized resources for users [1]. Cloud computing is a computing model based on computer networks which presents a new pattern for providing, consuming and delivering computational services (such as infrastructure, software and other computational resources) via using network [2, 3]. Information technology resources are such as wireless sensor networks (WSNs) [4-6], mobile ad hoc networks (MANETs) [7-9] and Internet of things (IoT) accessed at users' request time and based on their needs; they are delivered in a flexible and scalable manner through the internet. That is, users only pay the costs of their own consumed electricity and water. In case cloud computing is applied, users will only pay the cost of services they have used [10-12]. A user and a service provider should sign an agreement so that a service can be provided and delivered to the user. In such an agreement, the level and content of the provided services, i.e. data

quality, data management, costs, etc., are mentioned and announced by the providers of cloud computing.

Indeed, as mentioned above, cloud computing is a requested computational model which requires a large amount of physical tools. The services are provided as soon as they are requested by the users. Hence, excessive requests of cloud computing increases power consumption in data centers. These centers consume huge amounts of power which consequently leads to releasing large amounts of carbon. Resource optimization, power consumption reduction and virtual machine stabilization are considered as the major challenges in this research domain [13].

Systems can be used simply and straightforwardly through the application of cloud computing. Scalable resource management in cloud computing is made possible via virtual machines [14, 15]. As the number of requests for using virtual machines increases, cloud computing resources will be also increasingly used. As a result, scheduling resources can be considered as an effective method [16, 17]. As the number of requests increases, the number of virtual machines will be insufficient; thus, solutions are needed for enhancing scalability and

* Corresponding Author

reducing overheads in resource scheduling and enhancing service quality [18-20].

Since the application and utilization of cloud computing has increased significantly [21], approaches and solutions need to be proposed for enhancing service quality, service reliability, optimal resource use and saving power consumption. Virtualization carries out live migration of virtual machine in the network by guaranteeing the lowest drop. Durability of the fixed virtual machine can enhance the efficiency of the dynamic resources.

Service quality guarantee and reliability according to service level agreement is considered to be an irrevocable and indispensable requirement for cloud service providers [22].

Power consumption optimization in cloud data centers is composed of two steps: the first step is the effective allocation of virtual machines which is aimed at maximal utilization of available resources and saving energy [23-27]. The second step is the optimization of the allocated resources. Given the above-mentioned discussions, it is imperative that methods be presented for reducing costs of power consumption, data transmission, bandwidth and also for enhancing service quality in cloud computing.

In this paper, by capitalizing on Big Bang–Big Crunch algorithm [28] in the resource allocation stage, we made an effort to enhance QoS (quality of service) and reduce power consumption. In this paper, we consider the power consumption of different network elements that create the network topology in backbone networks as well as the power consumption in physical machines and data centers. Simulation results show that the proposed scheme is an energy efficient embedded scheme for cloud computing environment.

The rest of the paper is organized as follows: in Section 2, previous methods and works regarding power consumption in cloud computing are briefly reviewed. In Section 3, the proposed algorithm is presented. In Section 4, the required simulations for the proposed algorithm were carried out in Cloudsim environment and the obtained results were compared with those of other similar algorithms. In Section 5, the conclusions and findings are summed up and directions for further research on optimizing service quality in cloud computing are recommended.

2- Related Works

One of the critical issues which has attracted most users' attention is the optimal and maximal use of resources. Embedding virtual machine is one of the most fundamental approaches for virtualization in cloud computing [29]. Embedding or allocating resources can be done in different domains such as operating systems,

management of data centers and the selection of the best virtual machine. It was aimed at meeting users' needs.

Embedding virtual machines is a mapping process from virtual machine to physical machine. In other words, it is the selection of the best and most ideal physical machine for virtual machine. The number of virtual machines in each class may be different depending on users' demands [30]. In the architecture of embedding a virtual machine, each data center has a series of physical servers. Each physical server is virtualized through several virtual machines [31].

The current problem in embedding virtual machines is concerned with how to embed virtual machines in achieving maximal efficiency. Here, the term efficiency refers to lesser use of load threshold, cost reduction, scalability and power consumption in cloud data center. However, it should be noted that some of the objectives are in contrast with each other. Hence, all of them may not be achieved in an embedding scheme. In case an appropriate method is not used in data centers, undesirable results such as increased power consumption, reduced service quality, reduced customer satisfaction and other problems might be imposed on service providers.

By capitalizing on proper allocation of virtual machines or techniques in data resource migration among virtual machines, power consumption can be optimized in cloud computing. Here, some of the previous works on saving power consumption in cloud data centers and efficient techniques of embedding virtual machines are reviewed in short.

In [32], the authors investigated the issue of allocating virtual machines by focusing on maximal use of multi-dimensional resources and reducing power consumption. The problem of allocating virtual machine was solved by using honeybee algorithm via hierarchical clustering which was intended to reduce power consumption in servers.

In [33], the authors proposed a method for allocating dynamic resources of data centers based on functional demands. It was aimed at optimizing the number of active servers and supported green computation. An effective algorithm, namely variable item size bin packing (VISBP), was obtained which operates well in real environments by adjusting available resources in physical servers. Their proposed approach supports green computing by optimizing the number of servers used. Experimental results show that the VISBP has better performance in hot spots' migration and load balance when compared to the existing algorithm. On the other hand, the assumption in the paper that all physical machines are homogeneous with unit capacity may be the cardinal restriction of its application.

In [34], the authors proposed the idea of the pool of available physical resources which is presented like a backpack; it is solved by genetic algorithm so as to

achieve optimal allocation. In this way, a better solution can be achieved by considering several multi-dimensional parameters in the demand for virtual machine. As a result, virtual machine migration is reduced and energy is saved as much as possible [35].

In [22], virtual machine was selected and embedded for reducing power consumption and enhancing service quality in virtualization and virtual machine stabilization. It was intended for achieving optimal resource efficiency. In this work, researchers used a computational evolutionary algorithm, i.e. compatible genetic algorithm, based on virtual machine stabilization method. Virtual machine was embedded by compatible genetic algorithm via various virtual machine selection policies such as minimum migration time, highest solidarity and correlation and random selection. It was implemented by different processors; it was found that compatible genetic method is desirably responsive to large areas. In cases where energy issue is of high significance, this method can guarantee high service level and service quality for large cloud computing systems with the least amount of agreement violation.

In [36], the issue of embedding virtual machine and data was regarded as an NP-Hard problem. The meta-heuristic algorithm of ant-colony optimization was used for solving this problem. In this method, a set of neighboring physical machines was selected for embedding data and virtual machines. Data are distributed in physical storage tools of the selected physical machines. For processing the capacity of each physical machine, a set of virtual machines is embedded in these physical machines so that the data stored in them can be processed. Simulation results indicate that this method selects physical machines close to each other and tasks are carried out in the allocated physical machines. In this way, the total time for doing tasks is reduced. However, in this scheme each virtual machine is mapped onto single physical machine, which shows that the amount of network resource allocated is not less than its demand and hence the proposed embedding algorithm may not give efficient result in terms of power consumption.

In [37], the authors presented a method for allocating virtual machine by using eagle strategy from Hybrid Krill Herd (KH). It was aimed at enhancing the use of services for internet users. Cloud computing services are presented through the communications of cloud data centers and through sharing the resources of virtual machines. Efficiency parameter is reduced in allocating virtual machine and in abnormal working load of cloud centers and in obstructions. Quality of experience (QoE) is reduced in data centers. The policy of optimizing virtual machine allocation eliminates the impact of remoteness from data center and obstruction impact. In this way, uninterrupted access with appropriate throughput is provided for optimizing the quality of experience.

Furthermore, an additional optimization flag is added to the service level agreement. When the workload is explosive, the virtual machine is automatically expanded by using Hybrid Krill Herd (KH). If optimization flag is adjusted properly, optimization is done again. Change and reaction protocol and the agreement protocol were recommended for predicting optimal resources in virtual machines for avoiding abnormality and density. Initial parameters such as delay, packet delivery rate and throughput are continuously controlled and observed for activating re-optimization and accessing the quality of experience with minimal load. Experimental results indicate the enhanced capacity of hybrid KH algorithm on particle optimization algorithm, ant colony optimization algorithm and genetic algorithm.

In [38], using imperialistic competitive algorithms and genetic algorithm, researchers tried to allocate virtual machines to physical hosts. Here, the intersection operation of the genetic algorithm was mixed with IC. Simulation results indicate notable improvements with regard to power consumption of the proposed algorithm. This scheme infers that the users need to send the maximum resource requirement of the virtual machines and virtual links, which may lead to the resource over-subscription problem and cost ineffectiveness.

In [39], the authors developed a new and effective evolutionary method for allocating VMs which was capable of maximizing the energy efficiency of a cloud data center via reservation of VMs. A new fitness function was presented based on energy definition which can effectively reduce power consumption and use the resources of data centers based on reserve. According to this evolutionary algorithm, a VM allocation method was put forth which was able to carry out VM to PM mapping with the best energy efficiency. Finally, an efficient simulation engine was developed for maximizing heuristic solutions of optimal VM allocation which can reduce the required time for evaluating VM allocation solutions in each iteration. The experimental results of the simulation in CloudSim and cloud environment indicate that the method proposed in this study can achieve not only an optimal allocation solution for a set of protected VMs but also can obtain further VMs with fewer physical machines for achieving higher energy efficiency.

In [40], the authors presented a novel merge-and-split-based coalitional game-theoretic scheme for VM consolidation in heterogeneous clouds. At first, they partition PMs into different groups based on their workload levels, then they employ a coalitional-game-based VM consolidation algorithm (CGMS) in selecting members from such groups to form effective coalitions, performs VM migrations among the coalition members to maximize the payoff of every coalition, and finally keeps PMs running in a high energy-efficiency state.

In [41], the authors proposed a spatial task scheduling and resource optimization (STSRO) method to minimize the total cost of their provider by cost-effectively scheduling all arriving tasks of heterogeneous applications to meet tasks' delay-bound constraints. STSRO well exploits spatial diversity in distributed green cloud data centers (DGDCs). In each time slot, the cost minimization problem for DGDCs is formulated as a constrained optimization one and solved by the proposed simulated annealing-based bat algorithm (SBA). As a result, the acceptance ratio and resource utilization is not optimized.

In [42], the authors proposed an energy aware clustered load balancing system in which, heterogeneous resources are clustered into different groups by using a partitioning-based clustering algorithm. The clustering reduces number of resources needs to be searched and hence minimizes the time required for resource discovery. An energy aware best-fit virtual machine (VM) allocation is used for reducing the power consumption. The process allocations to VMs are done based on best-fit allocation strategy for optimal space utilization.

The above-mentioned brief review of the related works shows that several methods have been developed and proposed for embedding and allocating virtual machines. Nonetheless, controlling and reducing power consumption and enhancing service quality still remain as serious challenges in cloud computing. The method proposed in this paper for optimizing efficiency and reducing power consumption in cloud computing by embedding virtual machine via Big Bang–Big Crunch algorithm are discussed in the following section.

3- The Big Bang–Big Crunch Optimization Algorithm

Big Bang–Big Crunch optimization algorithm is a meta-heuristic population based evolutionary scheme presented by Erol and Eksin [28]. It includes the following stages:

Step 1: (Big-Bang phase): An initial generation of N candidates is generated randomly in the search space, similar the other evolutionary search algorithms.

Step 2: The cost function values of all the candidate solutions are computed.

Step 3: (Big Crunch phase): This phase comes as a convergence operator. Either the best fit individual or the center of mass is chosen as the Centre point. The Centre of mass is calculated as:

$$x_c = \frac{\sum_{i=1}^N \frac{x_i}{f_i}}{\sum_{i=1}^N \frac{1}{f_i}} \quad (1)$$

Where in Eq. (1), x_c is the position of the center of mass, x_i is the position of the candidate, f_i is the cost function value of the i th candidate, and N is the population size.

Step 4: New candidates are calculated around the new point calculated in Step 3 by adding or subtracting a random number whose value decreases as the iterations elapse, which can be formalized as:

$$x^{new} = x_c + \frac{\gamma \rho (x_{max} - x_{min})}{k} \quad (2)$$

Where in Eq. (2), γ is a random number, ρ is a parameter limiting search space, x_{min} and x_{max} are the upper and lower limits, and k is the iteration step.

Step 5: Repeat step 2-4 until stopping criteria has not been achieved.

4- The proposed Method

In this paper, the technique of optimally allocating virtual machines to the requests was proposed for enhancing efficiency and service quality and reducing power consumption in cloud computing. Accordingly, Big Bang–Big Crunch meta-heuristic method was proposed for allocating virtual machines to the requests. In the next stage, since the objective of the proposed method was to reduce the utilization of physical machines for reducing power consumption and enhancing system efficiency, system should be planned in such a way that the highest number of physical hosts should be put in the sleep mode. Hence, sleep/awake technique and threshold were used for determining hosts which should be turned off.

4-1- Fitness Function

The value and fitness of a solution should be measured by a fitness function so as to find out how appropriate the response of that solution is. Based on effective parameters about the quality of a solution, this function attributes a value to the solution. In the proposed algorithm, by applying this function on all the solutions, the fitness of each one is measured; the solution with the best value which may be maximal or minimal according to the policy of placing parameters is considered as the most suitable solution. The following parameters were used in the proposed method. Table 1 gives the parameters used in the formulas.

Table 1. Parameters used in the proposed method

Parameter	Description
V	The set of virtual machines
P	The set of physical hosts
V_i	One virtual machine in the V set
V_i^{CPU}	The amount of required processor for virtual machine i
V_i^{mem}	The amount of required main memory for virtual machine i

P_j	One physical host in the set P
P_j^{CPU}	The processing capability of physical machine P_j
P_j^{wcpu}	The total workload of the processor of the physical host P_j
V_{jp}	The set of virtual machines allocated to physical hosts

In the physical host, μ_j is measured according to Eq. (3):

$$\mu_j = P_j^{wcpu} / P_j^{cpu} \quad (3)$$

Where, in Eq. (3) μ_j refers to efficiency rate of the processor of the physical host p_j . Consequently, power consumption will be computed as follow [35].

$$P(u) = k.P_{max} + (1 - k).P_{max}.u \quad (4)$$

Where in Eq. (4), P_{max} is the maximum power consumed when the server is fully utilized; k is the fraction of power consumed by the idle server; and u is the CPU (central process unit) utilization. In this way, the fitness function for measuring the value of each response is computed via Eq. (5).

$$fitness = 1 / \sum_{j=1}^N P(u_j) \quad (5)$$

In the first stage of the proposed method, after a virtual machine is requested by a user, virtual machines are randomly allocated to requests according to Big Bang–Big Crunch algorithm. Given the total allocation methods, N allocation methods are defined as the initial population. In the third stage, solutions change their directions. Then, in the fourth stage, the best virtual machine allocation is searched by orbiting around the initial response so that a more optimal allocation is obtained. After a certain number of the iteration of the algorithm, the best virtual machine allocation to the requests is determined.

Excessive use of resources, especially the processor, leads to the overheating that resource and, consequently, heat loss and power loss. Furthermore, resource reception requests which arrive at the resources with overload should stay in the service reception queues; hence, it results in the increased response time, violation of service level agreement and reduced service quality. On the other hand, resources with an amount of load less than normal which are on need to consume energy so that they can keep their physical machines in the awake mode. They use a negligible portion of their capacities and have poor efficiency; this is considered as a waste of resource and power. Hence, it is necessary that a technique be used for determining hosts with low load and hosts with excessive

load. For doing so, efficiency threshold is determined for processors of the physical machines. For finding machines which should migrate, we need to determine machines with overload or machines with less than usual load. The existence of hosts with overloads leads to reduced system efficiency and service quality and, also, increased temperature of the system. As a result, it is imperative that hosts with overload be determined and that working loads be distributed in a balanced way in the entire system. The method developed in [43] was used for determining threshold.

Figure 1 indicates an example for virtual network embedding in physical machines.

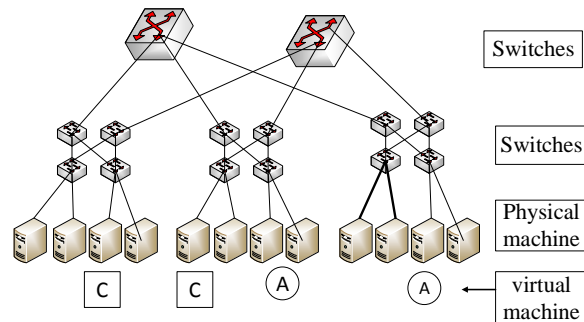


Fig. 1. An example of virtual network embedding: Physical network.

Degree of CPU use of all host machines was considered as a criterion for determining overly used machines and less used machines. If a host machine is specified as an overly used machine, some of the virtual machines of this host machine will migrate to another host machine. Also, in case the host machine is determined as a less used machine, some of the virtual machines of this host machine will migrate to another host machine. Then, according to sleep/awake technique, unused machines will go into sleep/awake modes. According to Beloglazov method, if CPU use is less than low threshold (LT) value, that host machine will be regarded as a less used machine. On the other hand, if CPU use is more than up threshold (UT) value, it will be considered as an overly used machine. In this way, LT and UT are used for specifying less used and overly used machines.

As shown in Fig. 2, by defining threshold boundaries, we will deal with three types of hosts. That is to say, hosts with values less than low threshold are regarded as hosts with low working loads. Hosts with working loads which are higher than up threshold value are known as hosts with overloads. Finally, hosts with working load between UT and LT are considered to be the ideal hosts.

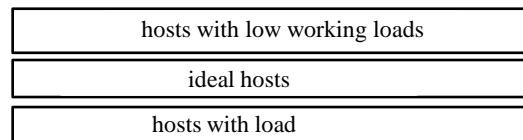


Fig. 2. Classification of hosts in terms of working load

In the proposed method, we tend to reduce the utilization of working machines in line with reducing power consumption and enhancing system efficiency. After low-load hosts are determined, we can migrate virtual machines of these hosts to other hosts according to sleep-awake algorithm and change their mode to the sleep mode. In this way, power consumption is reduced; even by turning off these hosts, their power consumption can be reduced to zero. Hence, it can be argued that turning off an idle host can prevent the loss of power consumption. Algorithm 1 shows Pseudo code of possible migration.

Algorithm 1. Pseudo code of possible migrate
1: Input: VMList: $G^V(N^V, E^V)$, Physical list: $G^P(N^P, E^P)$
2: Output: allocation of VMs, Embedded list
3: for $i=1$ to all host (PMs)
4: if (is_underload (host _i)==true)
5: VM_selection=Random Selection (between all VMs in Host _i);
6: VM_selection used as input for allocation agent
7: else
8: foreach VM in VMList do
8: If ((Possible_migrate (all Host _i)==true)
9: all Host _i VMs migrate
10: end if
11: end for
12: end if
13: end for
14: return migration_List
15: end.

As a result of identifying a host with overload, there will be two possible modes. The first mode is that there are ideal hosts and by sending one virtual machine of the host with overload, we can modify this host into an idea host. If one machine of the host is identified as an overly used machine, a number of the virtual machines of this host will be migrated to another host. Random selection (RS) policy was recommended for selecting the virtual machine which should be migrated from the host. According to this policy, the system randomly selects one of the virtual machines for migration; as a result of this migration, the host is modified into an ideal host and it is no longer considered as a host with overload. Consequently, service quality is enhanced because no queue will be established for receiving services; the system will produce less heat and power consumption will be optimized. The second mode is that there are no ideal machines; in this case, the mode of one sleeping machine will be changed into awake mode. Then, the virtual machine of the host with overload will migrate to that host.

The time complexity of the proposed scheme with m number of physical machines (physical servers) and n number of virtual machines (VMs) is $O(m \log m + n^2)$. According to algorithm 1, the running time of calculation of bandwidth availability of each physical machine is $O(m$

$\log m$). On the other hand, embedding n virtual machine to m physical machine using algorithm 1 and Big Bang–Big Crunch algorithm is $O(n^2)$. Hence, the time complexity of the proposed scheme can be written as: $O(m \log m + n^2)$ Figure 3 indicates the flowchart of the proposed scheme. The set of v nodes are grouped such that those that are not connected to each other in the topology graph are put into one group. Separate groups are created for all the other nodes that are connected to each other to avoid embedding nodes that are connected into the same substrate node.

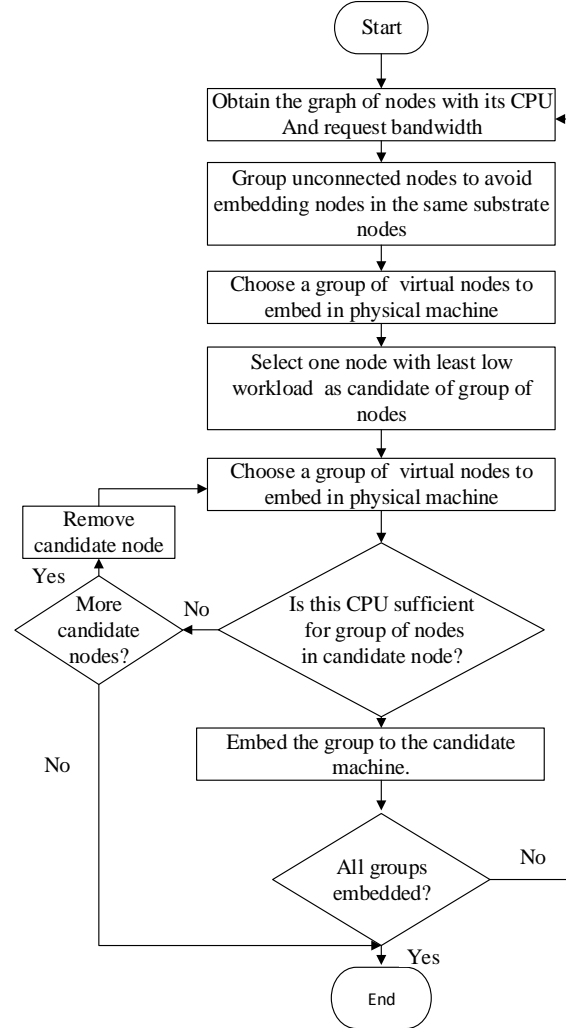


Figure 3. Flow chart of the proposed scheme

5- Performance Evaluation

For evaluating the proposed method, we simulated, investigated and compared it with genetic algorithm and PABFD (Power Aware Best Fit Decreasing) algorithm [43]. CloudSim was used for simulating the proposed method which is a well-known Toolkit in Java language for simulating cloud computing. In the respective scenario, a data center was simulated by 100 heterogeneous physical

nodes. Each node was considered with the following specification: a CPU core with identical performance of 1000, 2000 or 3000 millions of commands at each second (MIPS), 8 gigabyte of RAM memory and one terabyte storage space. The degree of power consumption of a physical host ranges from 175 watt with 0% CPU use to 250 watt with 100% CPU use. Each virtual machine needs 250, 500, 750 or 1000 MIPS, 128 megabyte RAM and one gigabyte storage space. The user records requests for supplying 290 homogeneous virtual machines which simulates the capacity of the entire data center. Each virtual machine executes a web application program or any other application with a variable working load. For the sake of usefulness and efficiency, CPU was modeled according to a randomly distributed variable. The application program for 15000 MIPS is equal to a 10-minute execution on 250 CPU units with 100% usefulness. At the beginning, VMs were considered according to the requested specifications and 100% usage. Each experiment was executed for 10 times and the results were obtained according to average values.

One of the significant parameters of the proposed method is the number of iterations of the algorithm which is regarded as the condition for the termination of the algorithm. The number of iterations affects the results and it has been experimentally demonstrated and determined which is shown in Fig. 4. In these experiments, the proposed scenario was executed with differing number of iterations. The vertical axis of Fig. 3 indicates power consumption based on kilowatt per hour and the horizontal axis indicates the number of iterations of the algorithm. It was observed that the best power consumption was related to 100 iterations of the algorithm.

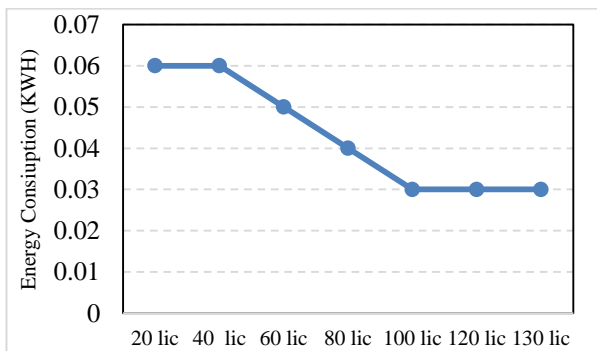


Fig. 4. The impact of the iterations in Big Bang–Big Crunch algorithm

In one experiment, the degree of power consumption was investigated by assuming IQR as host overload detection algorithm and RS (Random Selection) and MMT (Minimum Migration Time) algorithms were considered as VM selection algorithms. The proposed algorithm, genetic algorithm and PABFD algorithm were implemented in 10 random executions. As shown in Fig. 5, with regard to using RS algorithm, the proposed method had the average

power consumption of 47 watts per hour which was the lowest power consumption in comparison with GA (Genetic Algorithm) and PABFD algorithms. Furthermore, with respect to using MMT algorithm, the proposed algorithm had the lowest power consumption of 45 watts per hour in comparison with GA and PABFD.

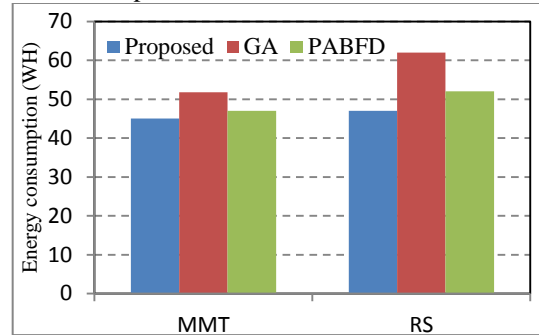


Fig. 5. Average power consumption by considering IQR as host overload detection algorithm in 10 executions

In another experiment, the degree of violated SLA (service level agreement) was investigated by assuming IQR as host overload detection and MMT and RS were used as VM selection algorithms. As shown in Fig. 6, by using MMT algorithm as VM selection algorithm, the degree of violated SLA for the proposed method was 1.321% which was better than those of GA and PABFD.

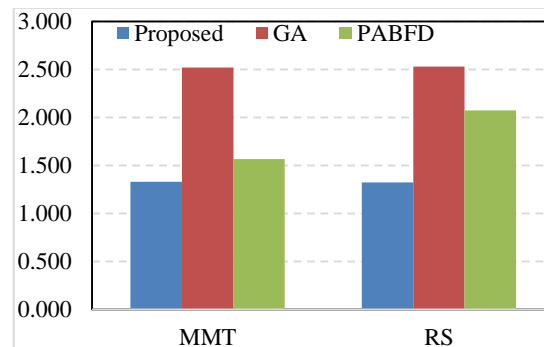


Fig. 6. Average rate of violated SLA by assuming IQR as host overload detection algorithm in 10 executions

In another experiment, power consumption was examined by assuming MAD as host overload detection and RS and MT were used as VM selection algorithms. The proposed method, GA and PABFD were implemented in 10 random executions. The results indicate that the proposed method had the lowest power consumption of 47 watts per hour in comparison with GA and PABFD. Moreover, regarding the use of MMT, the proposed method had the lowest power consumption (45 watts per hour) in comparison with the other two algorithms.

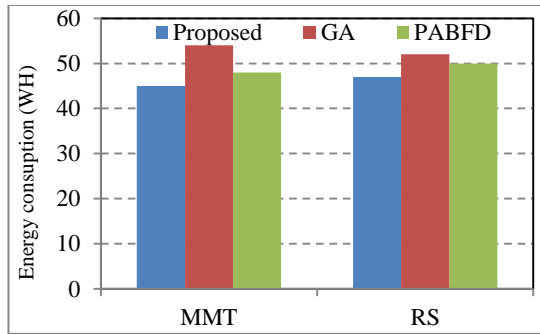


Fig. 7. Average power consumption by assuming MAD as host overload detection algorithm in 10 executions

As depicted in Fig. 7, it is observed that the proposed method had the lowest power consumption in comparison with GA and PABFD. Using RS and MAD, PABFD had better performance than GA. In this figure, the vertical axis denotes power consumption based on Watt per hour.

In another experiment, the degree of violated SLA was examined by considering MAD as host overload detection algorithm and MMT and RS were used as VM selection algorithm. The degree of violated SLA was randomly investigated in 10 executions. The average of ten executions is shown in Fig. 8. Given MMT as VM selection algorithm, the average degree of violated SLA for the proposed method was 1.038%. Also, given RS as VM selection algorithm, the degree of violated SLA was 0.435%. it can be maintained that the proposed method outperformed GA and PABFD with respect to the degree of violated SLA in using RS and MMT algorithms.

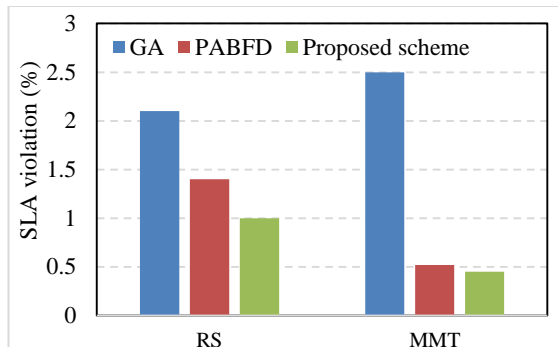


Fig. 8. Degree of violated SLA by using MAD as host overload detection algorithm in 10 executions

In on more experiment, power consumption was investigated by considering LR as host overload detection algorithm and RS and MMT as VM selection algorithms. GA and PABFD as well as the proposed method were implemented in 10 random executions. As depicted in Fig. 9, the proposed method had the lowest average power consumption of 56 watts per hour in comparison with GA and PABFD. Also, regarding the use of MMT algorithm, the proposed method consumed 44 watts per hour which was less than the power consumptions of GA and PABFD.

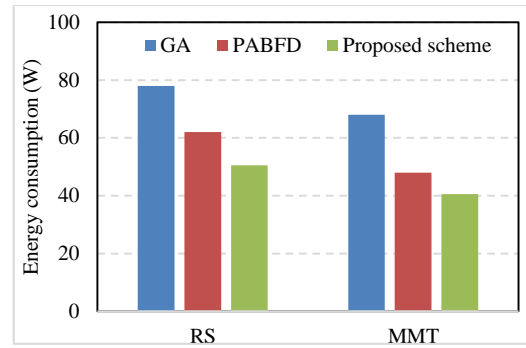


Fig. 9. Average power consumption by considering LR as host overload detection algorithm in 10 executions

In another experiment, LR, RS and MMT were considered as host overload detection algorithm and VM selection algorithms for investigating the degree of violated SLA. As depicted in Fig. 10, the degree of violated SLA for the proposed method regarding the use of MMT and RS algorithms were 1.087% and 0.672%, respectively. Hence, it can be highlighted that, given RS and MMT algorithms, the degree of violated SLA in the proposed method was better than those of GA and PABFD.

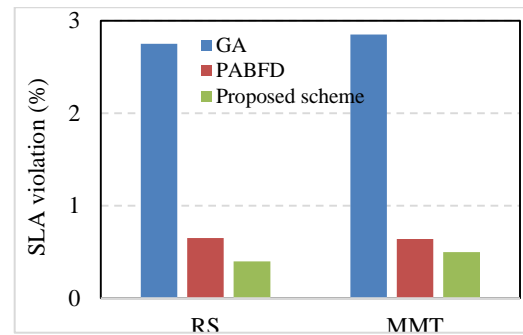


Fig. 10. Average degree of violated SLA by considering LR as host overload detection algorithm in 10 executions

Moreover, LRR, MMT and RS were, respectively, used as host overload detection algorithm and VM selection algorithm for checking power consumption. As Fig. 11 shows, the average power consumption of the proposed method was 46 Watts per hour which was less than those of GA and PABFD. Also, with respect to using MMT algorithm, power consumption of the proposed algorithm was shown to be 42 Watts per hour which was the least amount among the three algorithms.

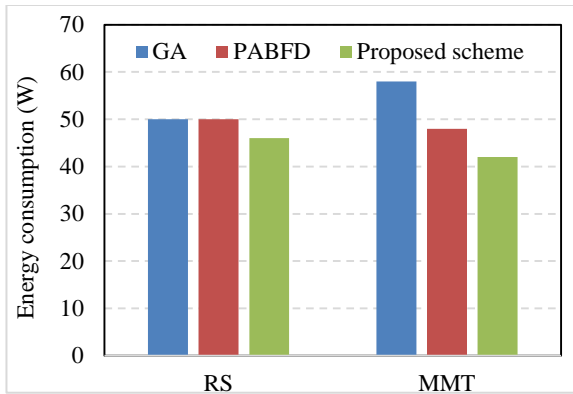


Fig. 11. Average power consumption with respect to LRR as host overload detection algorithm in 10 executions

Moreover, LRR, MMT and RS were respectively used as host detection and VM selection algorithms for examining the degree of violated SLA. As shown in Fig. 12, the degree of violated SLA for the proposed method regarding the use of MMT algorithm as VM selection algorithm was 1.076%. The degree of violated SLA for the proposed method, using RS algorithm as VM selection algorithm, was 0.418% which was better than those of GA and PABFD.

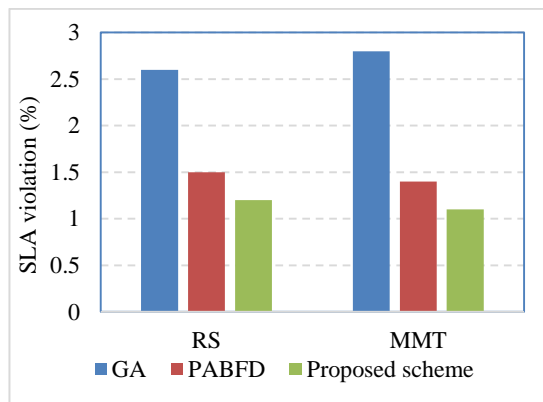


Fig. 12. Average degree of violated SLA with respect to using LRR as host overload detection in 10 executions

In sum, the results of simulations indicate that the proposed method along with LRR and MMT algorithms were desirable for detecting overload of physical host and selecting virtual machine. The average power consumption of the proposed method was 42 Watts per hour. The proposed method along with LRR and MMT was able to optimize power consumption for 11.90%, ESV for 5%, and EST for 49%.

6- Conclusion and Directions for Further Research

Cloud computing centers consume vast amounts of power which, consequently, lead to the release of significant

amounts of carbon. Resource optimization, reduction of power consumption and stabilization of virtual machines are considered as critical challenges in this domain of study which has attracted many researchers' attention. Indeed, reduction of power consumption, reduction of SLA agreement violation and service level and optimal use of resources were the primary objectives of the present study. Also, heat production reduction, cost reduction and having more green pastures were regarded as the secondary objectives of the study. By capitalizing on Big Bang–Big Crunch algorithm in allocating virtual machines, the proposed method was able to optimize power consumption and service quality in cloud computing. Furthermore, the proposed method optimized agreement violation and power consumption. Simulation results indicate that the proposed method along with LRR algorithm was suitable for detecting overload of physical host and MMT algorithm was appropriate for selecting virtual machine. The average power consumption of the proposed method was 42 Watts per hour which is regarded as optimal power consumption. Also, the proposed method as well as LRR and MMT algorithms was able to optimize power consumption for 11.90%, ESV for 5% and EST for 49%.

As directions for further research, for covering the difference regarding the degree of violated SLA in service level, cloud service providers can provide more service level than that requested by the users. Hence, it is recommended that a two-objective fitness function be used for reducing power consumption and service level agreement violation.

References

- [1] M. H. Ghahramani, M. Zhou, and C. T. Hon, "Toward cloud computing QoS architecture: Analysis of cloud systems and cloud services," *IEEE/CAA Journal of Automatica Sinica*, vol. 4, no. 1, pp. 6-18, 2017.
- [2] B. Varghese and R. Buyya, "Next generation cloud computing: New trends and research directions," *Future Generation Computer Systems*, vol. 79, pp. 849-861, 2018.
- [3] M. Noshay, A. Ibrahim, and H. A. Ali, "Optimization of live virtual machine migration in cloud computing: A survey and future directions," *Journal of Network and Computer Applications*, vol. 110, pp. 1-10, 2018.
- [4] A. Ghaffari, "Designing a wireless sensor network for ocean status notification system," *Indian Journal of Science and Technology*, vol. 7, no. 6, p. 809, 2014.
- [5] A. Ghaffari and A. Rahmani, "Fault tolerant model for data dissemination in wireless sensor networks," in *2008 International Symposium on Information Technology*, 2008, vol. 4: IEEE, pp. 1-8.
- [6] D. KeyKhosravi, A. Ghaffari, A. Hosseinalipour, and B. A. Khasragi, "New Clustering Protocol to Decrease Probability Failure Nodes and Increasing the Lifetime in WSNs," *Int. J. Adv. Comp. Techn.*, vol. 2, no. 2, pp. 117-121, 2010.
- [7] A. Ghaffari, "Vulnerability and security of mobile ad hoc networks," in *Proceedings of the 6th WSEAS international conference on simulation, modelling and optimization*, 2006:

- World Scientific and Engineering Academy and Society (WSEAS), pp. 124-129.
- [8] A. Ghaffari, "Real-time routing algorithm for mobile ad hoc networks using reinforcement learning and heuristic algorithms," *Wireless Networks*, vol. 23, no. 3, pp. 703-714, 2017.
- [9] R. Mohammadi and A. Ghaffari, "Optimizing reliability through network coding in wireless multimedia sensor networks," *Indian Journal of Science and Technology*, vol. 8, no. 9, p. 834, 2015.
- [10] W. Shu, W. Wang, and Y. Wang, "A novel energy-efficient resource allocation algorithm based on immune clonal optimization for green cloud computing," *EURASIP Journal on Wireless Communications and Networking*, vol. 2014, no. 1, p. 64, 2014.
- [11] M. Gahlawat and P. Sharma, "Survey of virtual machine placement in federated clouds," in *2014 IEEE International Advance Computing Conference (IACC)*, 2014: IEEE, pp. 735-738.
- [12] Z. Xiao, W. Song, and Q. Chen, "Dynamic resource allocation using virtual machines for cloud computing environment," *IEEE transactions on parallel and distributed systems*, vol. 24, no. 6, pp. 1107-1117, 2012.
- [13] M. Masdari, S. S. Nabavi, and V. Ahmadi, "An overview of virtual machine placement schemes in cloud computing," *Journal of Network and Computer Applications*, vol. 66, pp. 106-127, 2016.
- [14] F. López-Pires, B. Barán, L. Benítez, S. Zalimben, and A. Amarilla, "Virtual machine placement for elastic infrastructures in overbooked cloud computing datacenters under uncertainty," *Future Generation Computer Systems*, vol. 79, pp. 830-848, 2018.
- [15] S. Sotiriadis, N. Bessis, and R. Buyya, "Self managed virtual machine scheduling in Cloud systems," *Information Sciences*, vol. 433, pp. 381-400, 2018.
- [16] A. Kamalinia and A. Ghaffari, "Hybrid task scheduling method for cloud computing by genetic and PSO algorithms," *J. Inf. Syst. Telecommun*, vol. 4, pp. 271-281, 2016.
- [17] A. Kamalinia and A. Ghaffari, "Hybrid task scheduling method for cloud computing by genetic and DE algorithms," *Wireless Personal Communications*, vol. 97, no. 4, pp. 6301-6323, 2017.
- [18] V. Priya and C. N. K. Babu, "Moving average fuzzy resource scheduling for virtualized cloud data services," *Computer Standards & Interfaces*, vol. 50, pp. 251-257, 2017.
- [19] M. Elhoseny, A. Abdelaziz, A. S. Salama, A. M. Riad, K. Muhammad, and A. K. Sangaiah, "A hybrid model of internet of things and cloud computing to manage big data in health services applications," *Future generation computer systems*, vol. 86, pp. 1383-1394, 2018.
- [20] A. Satpathy, S. K. Addya, A. K. Turuk, B. Majhi, and G. Sahoo, "Crow search based virtual machine placement strategy in cloud data centers with live migration," *Computers & Electrical Engineering*, vol. 69, pp. 334-350, 2018.
- [21] K. R. Remesh Babu and P. Samuel, "Service-level agreement-aware scheduling and load balancing of tasks in cloud," *Software: Practice and Experience*, vol. 49, no. 6, pp. 995-1012, 2019.
- [22] P. R. Theja and S. K. Babu, "Evolutionary computing based on QoS oriented energy efficient VM consolidation scheme for large scale cloud data centers," *Cybernetics and Information Technologies*, vol. 16, no. 2, pp. 97-112, 2016.
- [23] M. Abdel-Basset, L. Abdle-Fatah, and A. K. Sangaiah, "An improved Lévy based whale optimization algorithm for bandwidth-efficient virtual machine placement in cloud computing environment," *Cluster Computing*, pp. 1-16, 2018.
- [24] X. Fu, J. Chen, S. Deng, J. Wang, and L. Zhang, "Layered virtual machine migration algorithm for network resource balancing in cloud computing," *Frontiers of Computer Science*, vol. 12, no. 1, pp. 75-85, 2018.
- [25] Z. Ning, X. Kong, F. Xia, W. Hou, and X. Wang, "Green and sustainable cloud of things: Enabling collaborative edge computing," *IEEE Communications Magazine*, vol. 57, no. 1, pp. 72-78, 2018.
- [26] H. Wang and H. Tianfield, "Energy-aware dynamic virtual machine consolidation for cloud datacenters," *IEEE Access*, vol. 6, pp. 15259-15273, 2018.
- [27] M. S. Mekala and P. Viswanathan, "Energy-efficient virtual machine selection based on resource ranking and utilization factor approach in cloud computing for IoT," *Computers & Electrical Engineering*, vol. 73, pp. 227-244, 2019.
- [28] O. K. Erol and I. Eksin, "A new optimization method: big bang–big crunch," *Advances in Engineering Software*, vol. 37, no. 2, pp. 106-111, 2006.
- [29] P. Zhang and M. Zhou, "Dynamic cloud task scheduling based on a two-stage strategy," *IEEE Transactions on Automation Science and Engineering*, vol. 15, no. 2, pp. 772-783, 2017.
- [30] S. Chaisiri, B.-S. Lee, and D. Niyato, "Optimal virtual machine placement across multiple cloud providers," in *2009 IEEE Asia-Pacific Services Computing Conference (APSCC)*, 2009: IEEE, pp. 103-110.
- [31] J. Gao and G. Tang, "Virtual Machine Placement Strategy Research," in *2013 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery*, 2013: IEEE, pp. 294-297.
- [32] M. Hemalatha, "Cluster based BEE algorithm for virtual machine placement in cloud data center," *Journal of Theoretical & Applied Information Technology*, vol. 57, no. 3, 2013.
- [33] W. Song, Z. Xiao, Q. Chen, and H. Luo, "Adaptive resource provisioning for the cloud using online bin packing," *IEEE Transactions on Computers*, vol. 63, no. 11, pp. 2647-2660, 2013.
- [34] N. Janani, R. S. Jegan, and P. Prakash, "Optimization of virtual machine placement in cloud environment using genetic algorithm," *Research Journal of Applied Sciences, Engineering and Technology*, vol. 10, no. 3, pp. 274-287, 2015.
- [35] A. Beloglazov, J. Abawajy, and R. Buyya, "Energy-aware resource allocation heuristics for efficient management of data centers for cloud computing," *Future generation computer systems*, vol. 28, no. 5, pp. 755-768, 2012.
- [36] T. Shabeera, S. M. Kumar, S. M. Salam, and K. M. Krishnan, "Optimizing VM allocation and data placement for data-intensive applications in cloud using ACO metaheuristic algorithm," *Engineering Science and Technology, an International Journal*, vol. 20, no. 2, pp. 616-628, 2017.

- [37] D. Kesavaraja and A. Shenbagavalli, "QoE enhancement in cloud virtual machine allocation using Eagle strategy of hybrid krill herd optimization," *Journal of Parallel and Distributed Computing*, vol. 118, pp. 267-279, 2018.
- [38] F. Farhadian, M. M. R. Kashani, J. Rezazadeh, R. Farahbakhsh, and K. Sandrasegaran, "An efficient IoT cloud energy consumption based on genetic algorithm," *Digital Communications and Networks*, 2019.
- [39] X. Zhang *et al.*, "Energy-aware virtual machine allocation for cloud with resource reservation," *Journal of Systems and Software*, vol. 147, pp. 147-161, 2019.
- [40] X. Xiao, W. Zheng, Y. Xia, X. Sun, Q. Peng, and Y. Guo, "A workload-aware VM consolidation method based on coalitional game for energy-saving in cloud," *IEEE Access*, vol. 7, pp. 80421-80430, 2019.
- [41] H. Yuan, J. Bi, and M. Zhou, "Spatial Task Scheduling for Cost Minimization in Distributed Green Cloud Data Centers," *IEEE Transactions on Automation Science and Engineering*, vol. 16, no. 2, pp. 729-740, 2018.
- [42] K. R. Babu and P. Samuel, "Energy aware clustered load balancing in cloud computing environment," *International Journal of Networking and Virtual Organisations*, vol. 19, no. 2-4, pp. 305-320, 2018.
- [43] A. Beloglazov and R. Buyya, "Optimal online deterministic algorithms and adaptive heuristics for energy and performance efficient dynamic consolidation of virtual machines in cloud data centers," *Concurrency and Computation: Practice and Experience*, vol. 24, no. 13, pp. 1397-1420, 2012.

Afshin Mahdavi received the B.S. degree in Computer Engineering from Azad University, Ardebil Branch, Iran in 2014, and M.Sc. degree in Computer engineering from Azad University, Tabriz Branch, Iran, in 2019. His research interests include Cloud computing and Computer networks.

Ali Ghaffari received his BSc, MSc. and Ph.D. degrees in computer engineering from the University of Tehran and IAU (Islamic Azad University), TEHRAN, IRAN in 1994, 2002 and 2011 respectively. As an associate professor of computer engineering at Islamic Azad University, Tabriz branch, IRAN, his research interests are mainly in the field of software defined network (SDN), Wireless Sensor Networks (WSNs), Mobile Ad Hoc Networks (MANETs), Vehicular Ad Hoc Networks (VANETs), networks security and Quality of Service (QoS). He has published more than 60 international conference and reviewed journal papers.

Reallocation of Virtual Machines to Cloud Data Centers to Reduce Service Level Agreement Violation and Energy Consumption Using the FMT Method

Hojjat Farrahi Farimani

Department of Computer Engineering, Neyshabur Branch Islamic Azad University Neyshabur, Iran
farrahi_hojjat@yahoo.com

Seyed Reza Kamel Tabbakh*

Department of Computer Engineering, Mashhad branch Islamic Azad University, Mashhad, Iran
rezakamel@computer.org

Davoud Bahrepour

Department of Computer Engineering, Mashhad branch Islamic Azad University, Mashhad, Iran
bahrepor@gmail.com

Reza Ghaemi

Department of Computer Engineering, Quchan Branch Islamic Azad University, Quchan, Iran
r.ghaemi@iauu.ac.ir

Received: 04/Jan/2020

Revised: 24/Mar/2020

Accepted: 08/May/2020

Abstract

Due to the increased use of cloud computing services, cloud data centers are in search of solutions in order to better provide the services demanded by their users. Virtual machine consolidation is an appropriate solution to the trade-off between power consumption and service level agreement violation. The present study aimed to identify low, medium, and high load identification techniques, as well as the energy consumption and SLAv to minimize. In addition to the reduced costs of cloud providers, these techniques enhance the quality of the services demanded by the users. To this end, reallocation of resources to physical hosts was performed at the medium load level using a centralized method to classify the physical hosts. In addition, quartile was applied in each medium to reduce the energy consumption parameters and violation level. The three introduced SMT - NMT and FMT methods for reallocation of resources were tested and the best results were compared with previous methods. The proposed method was evaluated using the Cloudsim software with real Planet Lab data and five times run, the simulation results confirmed the efficiency of the proposed algorithm, which tradeoff between decreased the energy consumption and service level of agreement violation (SLAv) properly.

Keywords: Cloud Computing, Energy Consumption, Service Level Agreement Violation, Virtual Machine Consolidation

1- Introduction

Use of cloud computing is on a rising trend, and cloud providers in cloud data centers constantly attempt to reduce energy consumption, while maintaining an acceptable service level agreement (SLA). Cloud services are offered to users with a wide variety (e.g., IaaS-PaaS-SaaS), and each provided service could offer various services depending on the needs of the cloud providers and users [1, 17]. Due to the growing demand for cloud platforms by numerous users across different networks, the main challenge faced by cloud providers is to provide services that are in proportion to the needs of the users, while also delivering the desired level of services and minimizing the power consumption to save costs through

effective measures such as keeping the servers cool and reducing the levels of environmental pollutants [1, 15].

Cloud providers use proper solutions to maximize the capacity of their servers, which in turn reduces the number of the active servers and minimizes the power consumption in these centers. On the other hand, shutting down several servers causes the number of active servers to become overloaded, which increases the probability of a service level agreement violation (SLAV) and ultimately discourages users. For this, cloud providers apply techniques such as virtual machine (VM) consolidation, which simultaneously strives to maintain a proper level of energy consumption and reduce agreement violations. Furthermore, the VM consolidation processes that use both heuristics and meta-heuristic algorithms employ multi-objective functions due to the NP-hard model of cloud computing. Such problems are aimed at finding the optimal solution from among the available solutions, and

*Corresponding Author

the application of the algorithm may vary depending on the type of the problem [8, 21].

VM consolidation, a balance must trade of between energy consumption and service level agreement by identifying high-load, low-load, and medium-load physical hosts in order to minimize the mentioned parameters.

The present study aimed to minimize the violation of user-demand service contracts by proposing a dynamic algorithm and decrease energy consumption, so that the proper physical host could be determined to reallocate resources based on the usage for minimal changes in the high-load and low-load of the physical host in the future. The consumed power in each cloud data center for each physical host was considered based on the amount of CPU usage in the physical host. Although other parameters are also important in this regard (e.g., main memory usage and network bandwidth), the most consumed power in a physical host is the CPU utilization rate due to the maximum CPU power utilization [8,20,21] , while other parameters (e.g., main memory and network bandwidth) consume small amounts of energy. As a result, high-load and low-load hosts could be distinguished, thereby moving (migration) VMs from the high-load physical host to the appropriate (low-load) host. Due to the migration of these VMs, the intended physical host exits the high-load state, reducing the possibility of agreement violation. However, the energy consumption is likely to increase due to the higher number of the VMs migrating to appropriate physical hosts (low-load).

The proposed method by introducing three policies to resource reallocation for this goal, we used physical hosts and by IQR Method (1) classification. In quartiles then find median of each quartile by median method to find best policies to resource reallocation. It discusses in section 3.

The proposed algorithm attempts to identify the medium-sized hosts that reduce their energy consumption through their reallocation and minimization of the service level agreement violation and to implement and compare the proposed algorithm with other algorithms, we will compare each algorithm during the five stages of program execution, testing, and results, which we will explain in details in Section 4.

2- Literature Review

Allocation of VMs to proper physical hosts in the cloud environment is a substantial challenge to decrease the power consumption and agreement violations (SLAV). Extensive research has been focused on the calculation and reduction of energy consumption, the most important of which is VM consolidation. One of the primary aims of VM consolidation is to find the proper solution for multi-purpose, predefined allocation [1]. In this regard, the

algorithms of VM consolidation have been presented based on the approach of reducing energy consumption using heuristic and meta-heuristic methods [2].

The main methods that have been proposed for the further reduction of energy consumption and service level agreement in the cloud based on VM consolidation are implemented in several steps, including the detection of high-load hosts, selection of the proper VM for migration from high-load hosts or migrating all VMs from low-load physical hosts, and reallocation to other physical hosts [1, 3].

In [4], a new VM consolidation algorithm has been proposed based on the VM resource usage history. According to the obtained results, the service quality and energy consumption could be improved through the balancing of the energy consumption and service quality. This method consists of two steps, including the detection algorithm of the low-load physical host, which prioritizes the expansion of the number of the VMs on each host to select the optimal solution and turn off the low-load host to reduce the total system energy consumption, and identification of the high-load hosts to prevent agreement violations. The high-load host is unable to respond to VMs, which necessitates the migration process. For one thing, the energy consumption is kept down in an attempt to decrease the service level agreement [1, 3, 4] as it could increase the number of the migrants, as well as the energy consumption.

In [5], the minimum migration time (MMT) method was employed to migrate VMs from physical hosts and minimize service level agreement by reducing the number of the migrations, and the energy consumption observed a descending trend. In the same study, the number of the migrants had to be minimized in order to establish the efficiency and energy. In the migration process, the migration points are determined, and the hosts with lower efficiency are suspended through the migration of the VM hosts. If a host is over the threshold productivity, it is checked before the migration of its VM [3]. This method has also proposed as a load prediction algorithm to decide whether to induce the migration of a VM and determine which host is to be allocated depending on its workload in the future. According to the findings of the mentioned research, the number of the migrations and amount of consumed energy decreased due to the quality of the service.

The interquartile range (IQR) method [1] is used to identify THE high-load hosts that are above the threshold, assuming a threshold of 25% and minimum of 75% for the CPU utilization. Moreover, the median absolute deviation (MAD) method [1, 12] uses the median absolute deviation. In [7], four classes of VMs were considered based on the amount of the used CPU resources by classifying various physical hosts into different categories, with each category selected to reallocate resources, and only one category was

evaluated. Use of categorical selection and selection of low-load physical hosts for the reallocation of resources [4, 6] often increased the probability of physical hosts to become high-load again.

The single threshold (ST) method uses a single threshold [7], and only the high threshold is used to find the high-load hosts and reallocate the resources to the low-load machines only. On the other hand, multi-threshold methods [3, 8, 9] yield better outcomes in terms of compromising power consumption and agreement violation, while they also increase the rate of high-load physical host in the future. In addition, the middle method uses multiple thresholds to calculate the median of the physical hosts, attempting to determine the proper place for the allocation of the VM [3]. The compromise between two or more parameters (e.g., power consumption, agreement violations, and number of migrations) with direct correlations to service quality could be classified as an NP-hard problem in VM consolidation [1, 11]. Resource allocation is an inherent element in VM consolidation [8, 11].

In [10], GDR and MCP algorithms have been used based on the stable regression model in order to identify high-load hosts, as well as the dynamic BW policy for the selection of the VMs from the productive host to migrate. Although this method could significantly reduce energy consumption, it is still highly likely that the hosts reap the benefits in the future. Due to the use of a linear regression method as a more efficient technique than other methods [1, 10, 13], attempts have been made to predict the amount of the consumed energy by the physical hosts. The mentioned study provided a live migration program by examining the simulation results, while using the MAE algorithm by presenting five methods based on robust SLR [13] high-load and low-load hosts, virtual machine selection for migration, and resource reallocation.

In [13], MAE (10)-SLAV was selected as the optimal outcome based on the simulation results. After the examination and comparison of the introduced algorithms, it was observed that the algorithms that have been introduced so far provide no significant success in the identification of the average hosts that achieve desirable levels of energy consumption by resources reallocation and reduction of agreement violations. Therefore, we attempted to identify the medium-load hosts with the most significant impact on the energy consumption and agreement violation, as well as resource reallocation to the hosts after migration.

3- Methodology

In the present study, the appropriate algorithm was proposed based on the following steps:

1) Identification of the overloaded hosts;

- 2) Identification of the intermediate (medium) load hosts;
- 3) Identification of the low-load hosts;
- 4) Calculation of the probability of high-load/low-load hosts for resource reallocation to the medium-load hosts;
- 5) Selection of the proper VM from the high-load host to migrate to the medium-load host in accordance with the mentioned parameters (i.e., reduction of power consumption and agreement violation)

After selecting the proper VM to migrate from the high-load host to the medium-load host, the entire VM migrated from the low-load host to the medium-load host in order to turn off the low-load physical host. The manner of resource reallocation to the medium-load physical host with the potential of becoming high-load/low-load hosts may change in the future. In other words, if the medium-load host is reallocated in the future or the probability of being reallocated based on the current reallocation is high, the medium-load host will be reallocated again. In the following section, the proposed algorithm using IQR method (1) to classification physical host into Quartile, as shown in figure 1. Then using the median method (12) to calculating the median of each quartile, as shown in figure 2.

For physical hosts we need processed energy and SLAV that is discussed in section 3.1 and 3.2. In section 3.3 identification (Low-Medium-High) Load physical hosts and present three policies to resource re-allocation. We compare each policy to find the best classification to resource re-allocation.

3-1- Energy Consumption Model

The energy consumption model depends on various parameters, including CPU utilization, main memory utilization, and network bandwidth consumption. Since the maximum power consumption is based on CPU utilization [1, 3, 13], decreasing the number of the active processors (i.e., physical hosts) leads to the reduction of the total energy consumption of the system and proper distribution of the workload to different hosts based on the CPU utilization required to reduce energy consumption. Considering the high CPU utilization rate as energy consumption, the energy consumption model is interpreted based on the CPU utilization rate [1, 13], as follows:

$$P(u) = k p_{max} + (1 - k) p_{max} . U \quad (1)$$

In the equation above, p_{max} is the maximum consumed power when the physical machine is fully operational, K represents a fraction of the consumed power by an idle physical machine, and u shows the processor efficiency, which may change over time due to workload variability. As such, CPU efficiency is a function of time expressed as $u(t)$, while the total energy consumption of the physical

host is defined as the integral of the energy consumption function over time.

$$E = \int_{t=0}^{t=i} P(u(t))dt \quad (2)$$

Based on the proposed method and Equations 1 and 2, the consumption of the physical hosts and then their energy consumption could be calculated individually and as the sum of cloud energy consumed at a given moment, as follows:

$$E = Tt \sum_{i=0}^n E \quad (3)$$

In the equation above, n is the total number of the applied physical hosts, E_i represents the energy consumed by the host i to time t, and E shows the sum of the total cloud energy at time t.

3-2- Criteria of the Service Level Agreement Violation

Since service quality characteristics may vary in different applications, a specific criterion has been defined to evaluate SLA. SLAV violation encompasses several factors, most notably the repeated allocation of VMs to the physical host [18, 20]. For this reason, we examined this factor directly.

In other words, service quality is met if the physical host responds to the resources required by the VM for various applications in the required time. In the present study, two main criteria were considered based on [3] in order to measure the SLAV. These criteria are as follows:

1) Percentage of time; when physical machines are active and experience 100% efficiency, it is referred to as the time when each host or SLA threshold approach (SLATHA) violates the service quality.

(2, 19) Reduced overall performance with a large number of VM migrations; the reduction in performance-based migration is referred to as PDM. SLATHA is mainly used because if a physical host experiences 100% efficiency with its programs, the performance of the programs is limited by the capacity of the physical host. Therefore, VMs with the required level of service quality are not satisfied.

$$SLATAH = \frac{1}{N} \sum_{i=1}^n \frac{T_{si}}{T_{aj}} \quad (4)$$

$$PDM = \frac{1}{M} \sum_{j=1}^m \frac{C_{dj}}{C_{rj}} \quad (5)$$

In the equation above, N is the number of the physical hosts, T_{si} shows the total time that incurs while ith (physical host) has 100% efficiency and is subjected to the agreement violation, T_{ai} is the total time of ith active physical host, M estimates the number of the VMs, C_{dj} shows the violation of the ith VM that has been created by migration, and C_{rj} is the total processor capacity required by the VMj for the entire duration of the same VM. Since the SLATAH and PDM criteria independently and significantly determine the level of SLAV [1, 10, 13, 19], a composite criterion encompassing the performance violations regarding the high-load physical host was considered for the migration of the VMs. In this paper, a combined criterion for SLAV was used, as follows:

$$SLAV = SLATAH . PDM \quad (6)$$

3-3- Identification of the High-load and Low-load Hosts

In the current research, a threshold was used to detect the high-load and low-load hosts and also identify the medium-load hosts. The proposed idea was to find and reallocate the medium-load hosts to reduce the probability of other high-load and load-load hosts, while decreasing the energy consumption. To this end, the IQR algorithm was employed. Initially, the IQR threshold was calculated in order to detect the high-load or low-load physical hosts.

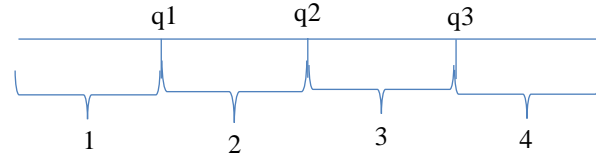


Fig 1. Classification of Hosts Based on Threshold

As is depicted in Figure 1, the IQR threshold was used to divide the total number of the physical hosts into four groups after ascending sorting, while the three considered categories were also investigated.

Table 1. Introduction of Each Category Based on Workload of Each Host

Category	Workload of Each Host
$< q1$	low-load
$q1 \leq x < q2$	medium-load
$q2 \leq x < q3$	high-load

In [13], the CPU utilizes the $q1$ - $q3$ bandwidth set as the low threshold to reallocate resources and use linear regression for its allocation. In this method, four categories are determined using the IQR threshold, and the median method is applied to calculate the median within the range of $q1$ - $q2$ or the second category (Figure 1) and $q2$ - $q3$. The

following steps are taken for the calculations in the median method:

- A. The physical hosts are arranged in an ascending order of workload (CPU usage).
- B. If the number of the physical hosts is odd, the middle of the set is selected, and if the number is even, two middle hosts are found, and the average value is considered as the median.
- C. In each step, a and c are obtained as the two sets that are repeated for each a and c set until all the physical hosts are in one set.

The name of each member of the physical host sets was determined, and we attempted to find the optimal destination for resource reallocation by calculating the median IQR threshold.

$$\text{If } \frac{A_i}{2} = 2x \quad (x \in 1,2,3,4,5, \dots, \infty) \quad (7)$$

$$\begin{cases} T_i = \text{median}(x_i) \\ T_u = \text{median}(x_j) \end{cases}$$

$$\text{If } \frac{A_i}{2} = 2x + 1 \quad (x \in 1,2,3,4,5, \dots, \infty) \quad (8)$$

$$\begin{cases} T_i = \text{median}(y_i) \\ T_u = \text{median}(y_j) \end{cases}$$

Finally, the high-load and low-load hosts were determined by establishing the following conditions:

$$\text{if } CA_i > T_u \quad (A_i = Oh) \quad (9)$$

$$\text{if } CA_i > T_i \quad (A_i = Uh) \quad (10)$$

The method applied in [3] could only be effective in the detection of high-load and low-load hosts and resource reallocation to lower-load machines in an attempt to distribute the workload and reduce SLAV. However, this reallocation will increase the number of the migrants and energy consumption, as well as the probability of high-load or low-load hosts in the near future. Therefore, we applied the IQR algorithm to provide four sets using the quartile and median for the detection of high-load and low-load hosts and identify the medium-load hosts.

3-4- Implementation Process of the Proposed Algorithm

As is depicted in Figure 2, the energy consumption list shows the same amount of energy consumption per physical host and is arranged in an ascending order. At the next stage, the IQR threshold algorithm was used to sort the entire list into four main categories. According to the information in Table 1, the hosts that were smaller than q1 were defined as the low-load hosts, and the hosts that were larger than q1 and smaller than q3, as well as those larger than q3, were defined as the high-load hosts. In addition, the median method was used in this regard to calculate the medians of q1-q2 and q2-q3, which

were defined as m1 and m2, respectively, with m considered as the median between the two intervals. Moreover, Ci was considered as the physical host. Table 2 shows the classification scheme of the proposed method.

[Arranged hosts in ascending order]

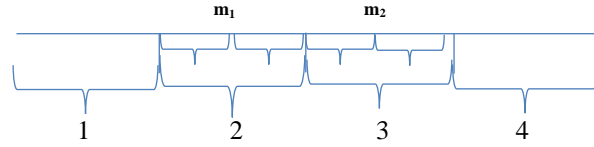


Fig 2. Quartile Sorting by Calculating Median of Two Intermediate

as well as those larger than q3, were defined as the high-load hosts. In addition, the median method was used in this regard to calculate the medians of q1-q2 and q2-q3, which were defined as m1 and m2, respectively, with m considered as the median between the two intervals. Moreover, Ci was considered as the physical host. Table 2 shows the classification scheme of the proposed method.

Table 2. Introduction of Each Category Based on Workload of Each Host in Proposed Method

Category	Workload of Each Host
$c_i < q_1$	low-load
$q_1 < c_i < q_3$	medium-load
$q_3 < c_i$	high-load

After using the proposed method to calculate the median of the second and third quartile (m1 and m2), the low- and medium-load hosts were defined, as follows:

- Policy 1: First Median Threshold (FMT)

Table 3. Medium-load Host between First Median and q2

$c_i < q_1$	low-load
$m_1 < c_i < q_2$	medium-load
$q_3 < c_i$	high-load

- Policy 2: None-median Threshold (NMT)

Table 4. Medium-load Host between q2 and q3

$c_i < q_1$	low-load
$q_2 < c_i < q_3$	medium-load
$q_3 < c_i$	high-load

- Policy 3: Second Median Threshold (SMT) and (median q1)

Table 5. Medium-load Host between Second Median and q1 (SMT)

$c_i < q_1$	low-load
$q_1 < c_i < m_2$	medium-load
$q_3 < c_i$	high-load

In the present study, three policies were considered as the possible solutions for the categorization and selection of the high-load and low-load hosts, while attempting to find the medium-load hosts as well. By reallocating resources, we were able to turn off more low-load hosts or transform the high-load hosts to medium-load hosts. Furthermore, the SLAV could be lowered to achieve better energy consumption and minimize the probability of the future hosts to become high-load and low-load through the accurate identification of the resources. All the proposed policies were used to reallocate resources to the physical hosts. Considering the optimal SLAV level and energy consumption and by reducing the probability of the filling of the physical hosts in the future, we attempted to select the optimal policy for the response to the resource reallocation

Figure 3 shows the sequence and process of implementing the. The first step involved the selection and classification of the FMT, NMT, and SMT policies between the medium-load hosts for resource reallocation.

4- Simulation and Evaluation of the Result

To simulate and evaluate the proposed method, all the introduced policies in the previous section were implemented using a simulator (CloudSim) [7, 16]. The measurable parameters included the total energy consumption of the system, agreement violation, and number of the shutdown machines, which were determined based on the low- and medium-load host identification models. The energy consumption levels have been discussed in Section 3.1 and are the metrics used in the implementation of the proposed method.

At this stage, the main objective of the research was to reduce the total energy consumption of the system.

The agreement violation rates have been described in Section 3.2 in terms of calculation (SLATAH and PDM). Since agreement violation is directly correlated with customer satisfaction rates [6, 10], cloud providers are more likely to attempt the provision of favorable levels of user demand and reduction of the SLAV; even in the cases where the energy consumption increases, customer satisfaction must be prioritized.

We review and compare three proposed host modes (low load - medium load - high load) and using median method to select suitable physical hosts from medium area and reallocate virtual machines to that hosts and the comparison algorithms are presented, considering the use of fixed datasets, each algorithm is run five times for each dataset and finally its average is compared with the other methods.

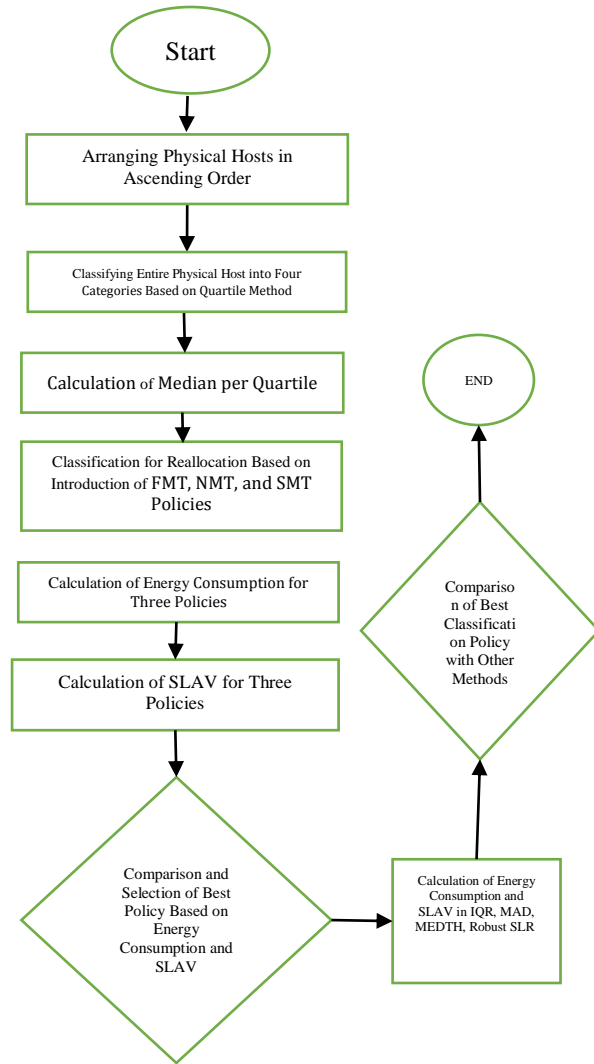


Fig 3. Flowchart of the proposed method.

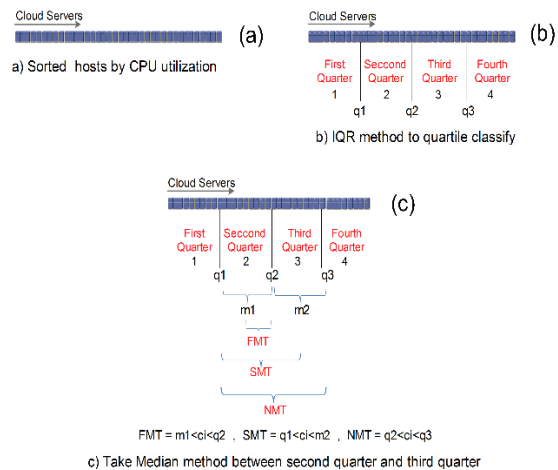


Fig 4. Show implementation of the proposed algorithm

4-1-Tested Dataset

All the introduced policies were used to implement the proposed method on the tested dataset shown in Table 6. The dataset contained actual data, and the results were simulated based on these data. The data were categorized by the number of the VMs, which were derived from the exact results of the experiments. Table 6 shows the data collected in the actual environment based on the number of the physical hosts and virtual hosts tested on specified dates [14].

4-2- Evaluation Criteria

By implementing the proposed method for all the policies and its comparison with the previous methods, as well as the IQR and MAD in similar conditions, the actual data in Table 6 were used. These data were a collection of more than thousands of VMs and physical hosts on various dates in the form of Planet Lab data in the CoMon project [14], which had been obtained on thousands of servers in 500 regions within five minutes. The rates of energy consumption and service level agreement were compared using the proposed method, and the compared criteria were as follows:

- 1) Calculation of the energy consumption based on the correlations (2, 15);
- 2) Calculation of the contract breach (6);
- 3) Calculation of the total agreement violation in the simulation based on the following equation (n=number of VMs) (11, 18, 19);

Overall SLAV

$$= \frac{\sum_{k=1}^n (requested MIPS) - \sum_{k=1}^n (allocated MIPS)}{\sum_{k=1}^n (requested MIPS)} \quad (11)$$

- 4) Number of THE shutdown hosts using the following equation:

Host Shutdowns (H) =
$$\frac{1}{n} \sum_{i=1}^n h(i) \quad (12)$$

(h(i): number of the active hosts at the time i, H: number of the shutdown hosts at time 1-n)

Table 6. Test Dataset

Row	Name of data	number of the VMs	number of the physical hosts
1	3 march 2011	1052	800
2	22 march 2011	1516	800
3	20 april 2011	1033	800

Table 7 shows the three proposed policies for the new location of resource reallocation based on the target

dataset for the energy consumption and service level agreement in each of the three datasets. By examining the simulation results of the policies, the FMT policy was considered to be the optimal policy for the reallocation the VMs. Considering that this policy accommodates a smaller range of medium-load physical hosts compared to The other proposed policies, the simulation results indicated that the selection of this policy to reallocate resources in terms of energy consumption and service level agreement violation could be further reduced by the other policies.

Table 7. Percentages of Performance Criteria for Policies 1-3

Policy review	Contract breach	Energy consumption(w)	Data	Row
Policy 1 (FMT)	0.1%	155.32	3 march 2011	1
Policy 2 (NMT)	0.106%	169.28	3 march 2011	2
Policy3 (SMT)	0.102%	162.65	3 march 2011	3
Policy 1 (FMT)	0.117%	178.52	22 march 2011	4
Policy 2 (NMT)	0.120%	184.36	22 march 2011	5
Policy 3 (SMT)	0.119%	179.12	22 march 2011	6
Policy 1 (FMT)	0.114%	130.21	20 april 2011	7
Policy 2 (NMT)	0.119%	137.59	20 april 2011	8
Policy 3 (SMT)	0.117%	133.23	20 april 2011	9

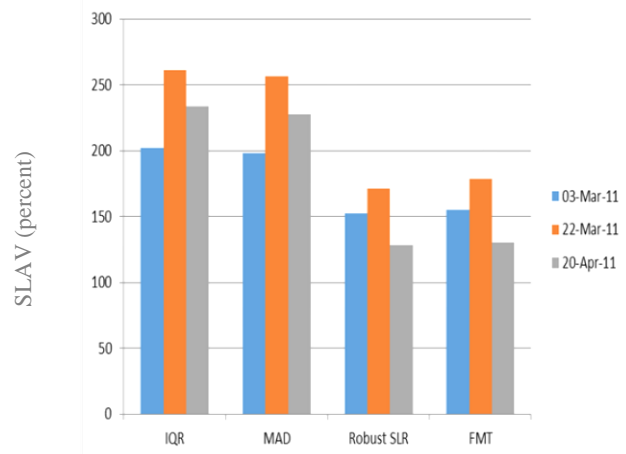


Fig 5. Percent SLA Violation Scenario 1-3

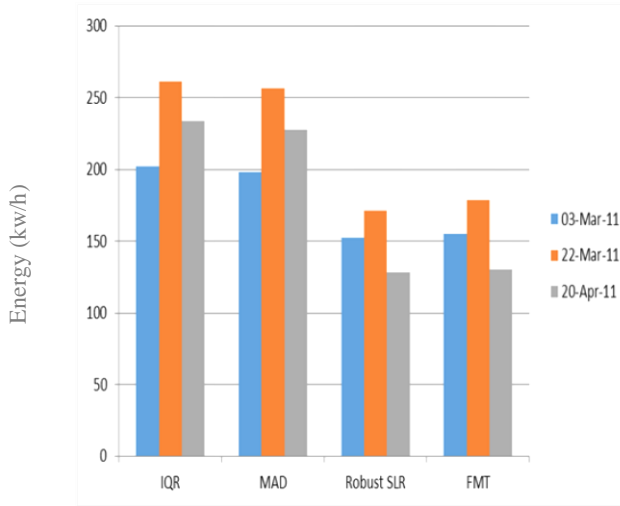


Fig 6. Power Consumption Scenario 1-3

According to the information in Table 7, the number of the VMs per data was compared between the policies in terms of energy consumption and agreement violations, while the response of the previous methods, IQR threshold, MAD threshold [12], median MEDTH method [3], and the robust SLR linear regression method were also compared [13]. Considering the implementation of the introduced policies and proposed method and the measured energy consumption (Figure 6), as well as the rate of agreement violation (Figure 5), Policy 1 (FMT) clearly yielded better results in terms of resource reallocation. Furthermore, the increased the number of the shutdown physical hosts resulted in the reduced energy consumption of the entire system. Therefore, policy 1 (FMT) was selected and compared with the other algorithms based on this policy. After selecting the FMT categorization policy, which enhanced the energy consumption and agreement violation rates, the mentioned policy and previous methods were compared based on the simulation results (Table 9). Accordingly, the proposed FMT classification policy achieved better results using the median method and quartile method in the reduction of both these parameters. For each algorithms in table 8 and table 9, we used dataset as shown in table 6. However, the results for each algorithm in table 8 and table 9 run for five times and it's average compare together.

Table 8. Comparison of Energy Consumption

Energy Consumption	Dataset	The policy of finding the threshold of low loads
201.92	3 march 2011	IQR [1]
261.37	22 march 2011	IQR
233.64	20 april 2011	IQR
198.16	3 march 2011	MAD [12]
256.18	22 march 2011	MAD
227.67	20 april 2011	MAD
152.66	3 march 2011	Robust SLR [13] MAE (10)-MME 2.5)
171.28	22 march 2011	Robust SLR MAE (10)-MME2.5)
128.52	20 april 2011	Robust SLR MAE (10)-MME 2.5)
155.32	3 march 2011	FMT
178.52	22 march 2011	FMT
130.21	20 april 2011	FMT

Table 9. Comparison of Service Level Agreement

contract breach	Dataset	The policy of finding the threshold of low loads	Row
0.350%	3 march 2011	IQR [1]	1
0.295%	22 march 2011	IQR	2
0.329%	20 april 2011	IQR	3
0.366%	3 march 2011	MAD [12]	4
0.318%	22 march 2011	MAD	5
0.351%	20 april 2011	MAD	6
0.123%	3 march 2011	MEDTH [7]	7
0.1195%	22 march 2011	MEDTH	8
0.122%	20 april 2011	MEDTH	9
0.118%	3 march 2011	Robust SLR [13] MAE (10)-MME 2.5-SLAV)	10
0.122%	22 march 2011	Robust SLR MAE (10)-MME2.5-SLAV)	11
0.121%	20 april 2011	Robust SLR MAE (10)-MME 2.5-SLAV)	12
0.100%	3 march 2011	FMT	13
0.117%	22 march 2011	FMT	14
0.114%	20 april 2011	FMT	15

As is depicted in Figure 7, the proposed FMT method yielded better results in terms of the reduction of the energy consumption and agreement violations rates compared to other methods based on

The tested datasets; Figure 8 also shows the further variations in this regard. A slight reduction was observed in the FMT with the robust SLR method in terms of energy consumption (same amount of energy consumption), which would significantly reduce the agreement violation rate as well.

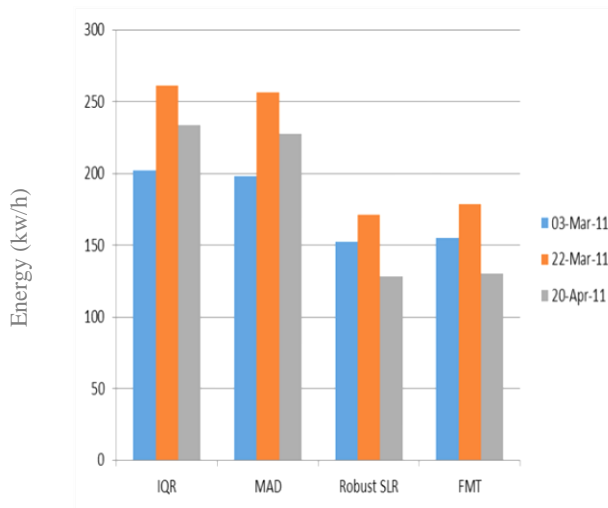


Fig 7. Comparison the energy consumption of the proposed algorithm with other methods

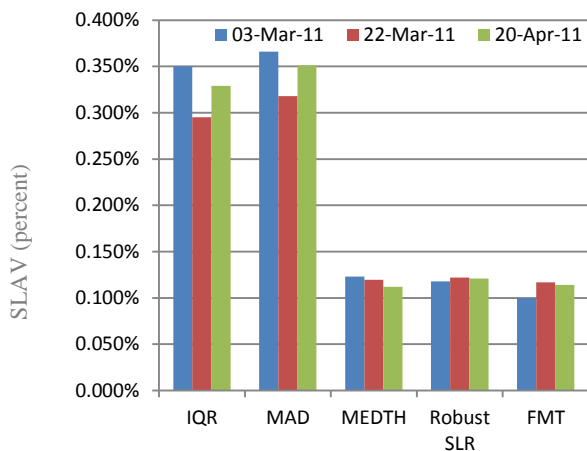


Fig 8. Comparison of contract SLA Violation percentage of proposed algorithm with other methods

5- Conclusions

In the current research, the median and quartile methods were utilized to propose a novel method for the detection of low-, medium-, and high-load hosts and resource reallocation to the medium-load hosts while keeping the energy consumption low and violating the user contract, maintaining it at the lowest level. To this end, three policies were considered based on the detection of the high-load, low-load, and medium-load hosts, and resources were reallocated to the host exhibiting the least consumption over the agreement violation.

According to the findings, the first policy (FMT) was the most viable option for resource reallocation considering the threshold and simulation results in Figures 1 and 4. As is depicted in Figures 6 and 7, the results of the adopted policy yielded better results compared to the other policies although the energy consumption of the FMT policy was approximately equivalent to the robust SLR policy; with slight variations in the energy consumption, the policy showed a significant reduction in the SLAV. Also, due to the shortage of resource reallocation intervals, this method causes overhead for physical hosts. The proposed algorithm could be used in the future in order to measure the number of migrations, consumed bandwidths, and shutdown hosts, as well as the total execution time and improve them.

References

- [1]. Beloglazov and R. Buyya, "Optimal Online Deterministic Algorithms and Adaptive Heuristics for Energy and Performance Efficient Dynamic Consolidation of Virtual Machines in Cloud Data Centers", *Concurrency and Computation: Practice and Experience*, vol. 24, pp, 2012,
- [2]. A.Varasteh and M. Goudarzi , "Server Consolidation Techniques in Virtualization Data Centers : A Survey", *IEEE System Journal* ,June 2017.
- [3]. O.Sharma , H.Saini," VM Consolidation for Cloud Data Center using Median based Threshold Approach" *Twelfth International Multi-Conference on Information Processing-2016 IMCIP*,2016.
- [4]. A.Horri, M.S.Mozafari and G.Dastghaibiyfard," Novel Resource Allocation Algorithms to Performance and Energy Efficiency in Cloud Computing", *The Journal of Super Computing*, vol. 69(3), pp. 1445–1461, 2014.
- [5]. S.Shaw and A.Singh, "Use of proactive and reactive hotspot detection technique to reduce the number of virtual machine migration and energy consumption in cloud data center", *Computers & Electrical Engineering*, 2015.
- [6]. Z.Zhou , Z.Hu and K.Li , " Virtual Machines Placement for Both Energy-Awareness and SLA Violation Reduction in cloud Data Centers " , *Hindawi Publishing Corporation Scientific Programming* , vol . 2016 , ID 5612039 , March 2016.
- [7]. R.Buyya, R.Ranjan and R.N Calleiros, "Modeling and simulation of scalable cloud computing environment and the CloudSim Toolkit: challenges on opportunities ", in

- proceedings and simulation (HPCS; 09), pp. 1-11, Leipzig, Germany, June 2009.
- [8]. A. Beloglazov, J. Abawajy and R. Buyya, "Energy-aware resource allocation heuristics for efficient management of datacenters for Cloud computing," *Future Generation Computer Systems*, vol. 28, no. 5, pp. 755–768, 2012.
- [9]. Z. Zhou, Z.G. Hu, T. Song, and J.-Y. Yu, "A novel virtual machine deployment algorithm with energy efficiency in cloud computing," *Journal of Central South University*, vol. 22, no. 3, pp. 974–983, 2015.
- [10]. R. Yadav, W. Zhang, O. Kaiwartya, P.R. Singh, I.A. Elgendy and YU. Tain, "Adaptive Energy-Aware Algorithms for Minimizing Energy Consumption and SLA Violation in Cloud Computing", *Special Selection on smart Caching, Communications, Computing and Cybersecurity for Information-Centric Internet Of Things*, DOI 10.1109/Access.2018.2872750, Vol 6, October 2018.
- [11]. E. Feller, C. Morin, and A. Esnault, "A case for fully decentralized dynamic VM consolidation in clouds", in *Proc. IEEE 4th Int. Conf. Cloud Compute. Technol. Sci.*, Dec. 2012, pp. 26–33, 2012.
- [12]. J. Xue, F. Yan, R. Birke, L. Y. Chen, T. Scherer and E. Smirni, "PRACTISE: Robust prediction of data center time series", in *Proc. 11th Int. Conf. Netw. Service Management (CNSM)*, Nov. 2015, pp. 126–134, 2015.
- [13]. L. Lianpeng, et al, "SLA-Aware and Energy-Efficient VM Consolidation in Cloud Data Centers Using Robust Linear Regression Prediction Mode", *IEEE Access*, 2019, 7: 9490–9500, 2019.
- [14]. K. S. Park and V. S. Pai, CoMon: "A Mostly-Scalable Monitoring System for PlanetLab", *ACM SIGOPS Operating Systems Review*, pp. 65–47, 2006.
- [15]. J. Shuja, S. A. Madani, K. Bilal, Kh. Hayat, S. Ullah Khan, Sh. Sarwar, "Energy-efficient data centers", *Computing* 94(12): pp. 973- 994, 2012.
- [16]. C. Rodrigo, R. Rnjan, C.A.F. De Rose, R. Buyya, Cloudsim: "A novel Framework for modeling and simulation of cloud computing infrastructure and services". *arXiv preprint arXiv*, 2009:0903.2525, 2009.
- [17]. C. Cardosa, M. Korupolu, and A. Singh, "Shares and utilities based power consolidation in virtualized server environments", In *Proceedings of IFIP/IEEE Integrated Network Management (IM)*, 2009.
- [18]. G.L. Stavrinides, H.D. Karatza, "The effect of workload computational demand variability on the performance of a SaaS cloud with a multi-tier SLA" in: *Proceedings of the IEEE 5th International Conference on Future Internet of Things and Cloud (FiCloud'17)*, pp. 10–17, 2017.
- [19]. Ferretti, Stefano, V. Ghini, F. Panzieri, M. Pellegrini, and E. Turrini, "QoS-Aware Clouds", in *Proc. IEEE 3rd Intern. Conf. on Cloud Computing (CLOUD'10)*, pp. 321-328, 2010.
- [20]. Z. Chi, W. Yuxin, Chi, Yuxin, Lv Y, Wu. H, Guo. "An Energy and SLA-Aware Resource Management Strategy in Cloud Data Centers". *Scientific Programming*. 2019, 2019.
- [21]. Xu, Heyang, Y. Liu, W. Wei, and Y. Xue. "Migration Cost and Energy-Aware Virtual Machine Consolidation under Cloud Environments Considering Remaining Runtime". *International Journal of Parallel Programming*. 2019 Jun 15; 47(3):481-501, 2019.

Hojjat Farrahi Farimani is a student of Computer Engineering at Azad University, Neyshabur branch, Iran. His research is focused on meta heuristic algorithm, Cloud Data Centers in software Engineering. Ph.D. Candidate in Azad University, Neyshabur Branch, Iran.

Seyed Reza Kamel Tabbakh is Assistant Professor in Department of Computer Engineering, Faculty of Engineering, Islamic Azad University, Mashhad branch, Iran. He received his B.Sc. degree in Software Engineering from Islamic Azad University, Mashhad branch, Iran (1999), his M.Sc. degree in Software Engineering from Islamic Azad University, South Tehran branch, Iran (2001), and his PhD in Communication and Network Engineering from Universiti Putra Malaysia (UPM), in 2011. He has several publications in national and international journals and conferences. His research interests include Internet of Things and IPv6 networks. He is an IEEE member.

Davoud Bahrepour received the M.S. and Ph.D. degrees in Computer Engineering from Azad University, Science & Research Branch, Tehran, Iran, in 2007 and 2012 respectively. Currently he is faculty member in Azad University, Mashhad Branch, Iran. His research interests include Computer architecture, Cloud computing and IoT.

Reza Ghaemi received the B.S. & Msc. degree in Computer Engineering in 1997 & 2001. He received his Ph.D. degree in Artificial Intelligence from UPM University of Malaysia in 2011. Now, he works as Assistant-professor in the Faculty of Computer Engineering at Islamic Azad University of Quchan, Iran. His area research interests include Artificial Intelligence, Machine Learning, Data Mining and Soft Computing.