



Journal of
**COMPUTER AND KNOWLEDGE
ENGINEERING**

Ferdowsi University of Mashhad

ISSN: 2717-4123

General Director: S. A. Hosseini Seno

Editor-in-Chief: M. Kahani

Publisher: Ferdowsi University of Mashhad

Editorial Board:

Mahmoud Naghibzadeh	Professor	Ferdowsi University of Mashhad, Iran
Mohammad H Yaghmaee-Moghaddam	Professor	Ferdowsi University of Mashhad, Iran
Dick H Epema	Professor	Delft Technical University, the Netherlands
Rahmat Budiarto	Professor	University Utara Malaysia, Malaysia
Mohsen Kahani	Professor	Ferdowsi University of Mashhad, Iran
Mohammad R Akbarzadeh-Tootoonchi	Professor	Ferdowsi University of Mashhad, Iran
Madjid Fathi	Professor	University of Siegen, Germany
Hossein Nezamabadi-pour	Professor	Bahonar University of Kerman, Iran
Ahmad Ghafarian	Professor	University of North Georgia, USA
Hamid Reza Pourreza	Professor	Ferdowsi University of Mashhad, Iran
Hadi Sadoghi-Yazdi	Professor	Ferdowsi University of Mashhad, Iran
Seyed Amin Hosseini Seno	Associate Professor	Ferdowsi University of Mashhad, Iran
Abedin Vahedian-Mazloun	Associate Professor	Ferdowsi University of Mashhad, Iran
Ebrahim Bagheri	Associate Professor	Ryerson University, Canada
Hossein Asadi	Associate Professor	Sharif University of Technology, Iran
Mahdi Kargahi	Associate Professor	University of Tehran, Iran
Hamid Reza Ekbia	Associate Professor	Indiana University, USA
Seyed Hassan Mirian Hosseinabadi	Associate Professor	Sharif University of Technology, Iran
Abbas Ghaemi Bafghi	Associate Professor	Ferdowsi University of Mashhad, Iran
Farhad Mahdipour	Associate Professor	Kyushu University, Japan

Administrative Director: T. Hooshmand

Journal of Computer and Knowledge Engineering

Faculty of Engineering, Ferdowsi University of Mashhad

P. O. Box. 91775-1111, Mashhad, I.R. IRAN

Tel: +98 51 38806024, Fax: +98 51 38763301, Email: cke@um.ac.ir, Site: cke.um.ac.ir

CONTENTS

A Data Replication Algorithm for Improving Server Efficiency in Cloud Computing Using PSO and Fuzzy Systems	Mostafa Sabzekar - Ehsan Mansouri Arash Deldari	1
An Efficient Ramp Secret Sharing Scheme Based on Zigzag-Decodable Codes	Saeideh Kabirirad - Sorour Sheidani Ziba Eslami	15
Analysis of the Impact of Wireless Three-User Multiple Access Channel Coefficients Correlation on Outage Probability: A Copula-Based Approach	Mona Sadat Mohsenzadeh Ghosheh Abed Hodtani	25
Optimization of FlexiTP Energy-Aware Algorithm in Wireless Sensor Networks	Hamid Mirvaziri	31
Dynamic Security Risk Management Considering Systems Structural and Probabilistic Attributes	Masoud Khosravi-Farmad Abbas Ghaemi-Bafghi	41
A Lightweight Secure Scheme for Data Aggregation in Large-Scale IoT-Based Smart Grids	Mohammad J. Abdolmaleki Amanj Khorramian Mohammad Fathi	57



A Data Replication Algorithm for Improving Server Efficiency in Cloud Computing Using PSO and Fuzzy Systems *

Research Article

Mostafa Sabzekar¹ , Ehsan Mansouri², Arash Deldari³

DOI: [10.22067/cke.2023.83351.1090](https://doi.org/10.22067/cke.2023.83351.1090)

Abstract: In different scientific disciplines, large-scale data are generated with enormous storage requirements. Therefore, effective data management is a critical issue in distributed systems such as the cloud. As tasks can access a nearby site to access the required file, replicating the desired file to an appropriate location improves access time and reliability. Replicating the popular file to an appropriate site is a good choice, as tasks can get the necessary file from a nearby site. In this research, a novel data replication algorithm is proposed that is consisted of four main phases: 1- determining 20% of commonly used files, 2- computing five conflicting objectives (i.e., average service time, load variance, energy consumption, average response time and cost) 3- finding the near-optimal solution (i.e., suitable locations for new replica) by the PSO technique to acquire a trade-off among the desired objectives. 4- replica replacement considering a fuzzy system with three inputs (i.e., Number of accesses, size of replica and the last access time). The experimental results denote that the proposed replication algorithm outperforms the Profit oriented Data Replication (PDR) and Bee colony-based approach for Data Replication (BCDR) strategies in terms of energy consumption, average response time, load variance, number of connections, Hit ratio, Storage usage, and cost.

Keywords: Cloud computing, Data Replication, Meta-heuristic algorithms, Fuzzy Systems, Power consumption.

1. Introduction

In the contemporary landscape, cloud computing plays a pivotal role in driving a burgeoning array of internet services. Numerous factors have contributed to the transformation of computing systems to cater to diverse industries' requirements, encompassing scientific breakthroughs, storage technologies, escalated utilization of multiple processes, and burgeoning user demands. From an infrastructural perspective, the cloud denotes a distributed and parallel system comprising a cluster of interconnected virtual machines. Its principal objectives revolve around facilitating users to lease, rather than purchase, computing resources that are accessible from any internet-connected location. Generally, service providers offer cloud computing

infrastructure, thereby curbing user expenses through service rental [1, 2].

On one hand, the volume and size of content generated in distributed systems continue to surge, while on the other hand, the quantum of data necessitating processing escalates by the second. The acquisition of this data poses a formidable challenge for system and network designers [3–6]. Data replication emerges as a valuable solution to diminish task execution duration by making data accessible to all relevant nodes. Strategic distribution of replicas across the cloud environment not only balances the load but also augments system efficiency. Data replication involves duplicating files across distinct segments of the system, averting disruptions to the workflow in case a file is damaged or inaccessible by relying on an available copy [7, 8].

Accessing cloud computing services mandates the availability of resources. Nonetheless, scarcity of resources within the infrastructure poses challenges, necessitating allocation of distinct computational resources for different processing tasks. Environments leveraging cloud computing can yield substantial advantages for projects necessitating extensive data processing, such as those in astronomy or meteorology domains. However, the colossal data volume poses a formidable challenge, particularly the replication of data. This predicament accentuates the significance of data replication across multiple servers and locations to uphold data integrity and accessibility. Generally, replication methods are employed to bolster system efficiency, with a focus on the 'how' and 'when' of replication, as well as its elimination. A gamut of methods and strategies has been developed to address these questions, all geared toward curtailing execution time. Consequently, each method necessitates the utilization of an iterative algorithm [9–11].

The main contributions of the paper can be listed as follows:

1. Centrality-based Replica Placement: The method introduces a novel approach for selecting the best site for storing replicas based on the centrality factor and the number of accesses. By considering these factors, the proposed strategy aims to reduce access time and improve data retrieval efficiency.

* Manuscript received: 2023 July 10, Revised, 2023 September 1, Accepted, 2023 September 18.

¹ Corresponding Author: Assistant Professor, Department of Computer Engineering, Birjand University of Technology, Birjand, Iran.

Email: sabzekar@birjandut.ac.ir.

² Department of Computer and Technology, Birjand University of Medical Sciences, Birjand, Iran

³ Assistant Professor, Department of Computer Engineering, University of Torbat Heydarieh, Torbat Heydarieh, Iran

2. Improved Data Center Utilization: The method also focuses on selecting appropriate data centers for specific services. This selection process benefits both customers and service providers by optimizing data center utilization, leading to improved resource allocation and overall system performance.
3. Fuzzy-Based Replica Replacement: To address storage limitations, the paper introduces a replacement strategy that utilizes a fuzzy system to evaluate the value of each replica. Unpopular replicas that are unlikely to be accessed in the future are replaced with more valuable ones, enhancing overall data management efficiency.

2. Background

2.1. Cloud computing service layers

While the architecture of cloud computing may appear straightforward, its effective operation hinges on astute management at the Network layer, facilitating seamless interconnection of systems. As depicted in Figure 1, cloud services are categorized into three overarching types [12–14], as perceived by cloud computing providers:

1. Software as a Service (SaaS) layer: This stratum empowers users to access existing software via a web browser. Notably, users gravitate towards the latest software iterations, and service providers assume the responsibility of delivering updates. Importantly, a customer's utilization of the service remains detached from the hardware's specifications and capabilities, as all computational processes are executed on the server-side.
2. Platform as a Service (PaaS) layer: This tier entails services offered by providers for application development and deployment. A suite of software accessible as services can be integrated into other layers, thereby reducing the necessity for direct placement of numerous programs onto the virtual machine. A notable exemplar includes an operating system that operates within this context.
3. Infrastructure as a Service (IaaS) layer: Within this stratum, resources such as processors, storage, and network components are accessible to end-users through virtualization. Direct control or access to the cloud computing infrastructure is restricted in this service model. A pivotal characteristic of this layer involves the dynamic allocation of resources through virtualization, underscoring its significance.

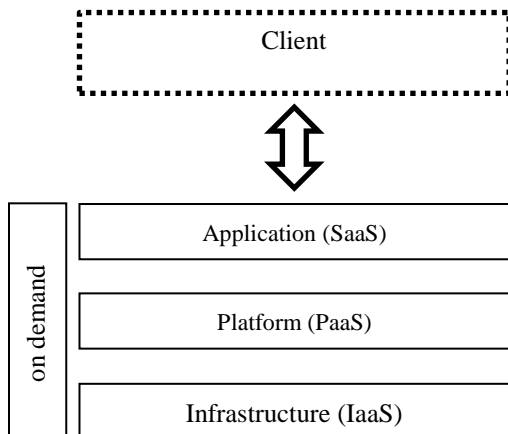


Figure 1. Cloud computing service layers

2.2. Particle Swarm Optimization (PSO) Algorithm

Based on the behaviors of birds and fishes, Particle Swarm Optimization (PSO) represents a population-based method. Groups of particles are characterized by position and velocity vectors, dictating new positions in each iteration. The new position is updated based on particle velocity, particle's best position, and overall best position.

In the PSO algorithm, particle positions are updated using these equations:

$$V_i(t) = w * V_i(t - 1) + c_1 * rand_1 * (P_{i.best} - X_i(t - 1)) + c_2 * rand_2 * (P_{g.best} - X_i(t - 1)) \quad (1)$$

$$X_i = X_i(t - 1) + V_i(t) \quad (2)$$

where, w is the weighted coefficient of inertia, c_1 and c_2 are constant training coefficients, $rand_1$ and $rand_2$ are two random numbers with a uniform distribution in the range of 0 to 1. Furthermore, X_i and V_i are the position vector and the velocity vector of the i -th particle, respectively. The best position found by particle i and the best position found by the swarm denotes by $P_{i.best}$ and $P_{g.best}$, respectively.

Evolutionary methods have both advantages and disadvantages due to their random nature. Selecting an algorithm is challenging, but evolutionary algorithms have been used successfully for various optimization problems. PSO algorithm offers several advantages [15]:

1. Memory benefit: Past information informs decisions.
2. Particle cooperation: Particles adjust positions based on group conditions, exchanging information to approach the best solution.
3. High convergence: Sharing particle information and quick decisions lead to fast convergence.
4. Implementation simplicity: All stages, from definition to decision-making, are easy to implement without complex math or stats.

2.3. Fuzzy logic System

In uncertain conditions, fuzzy logic is a suitable approach, replacing numerical variables with linguistic analysis. Fuzzy logic deals with values between 0 and 1 and allows verbs like "perhaps to be" or "to be if." Membership in a set is graded, allowing partial membership [16–18]. Notable features of this theory include:

1. Human-like thinking and decision-making simulation
2. Definition of approximations and non-deterministic answers
3. Complex function definition in linear and non-linear forms
4. Simple yet flexible implementation

Figure 2 displays the fuzzy system's architecture and data entry/exit process.

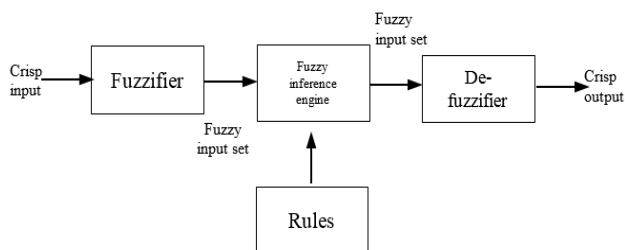


Figure 2. Architecture of a fuzzy system.

As shown in Figure 2, in the first step, all input numbers are converted into fuzzy sets. Then, a fuzzy inference engine evaluates the rules and after collecting the output rules, it is converted into an explicit or numerical value by the de-fuzzifier unit. Finally, the fuzzy results are converted into real numbers.

3. Related works

For data replication in cloud computing, numerous algorithms have been proposed [19], many employing evolutionary techniques [20]. Herein, various algorithms that have enhanced their performance using evolutionary methods at different implementation stages are discussed.

Leo et al. [21] presented a novel strategy using genetic algorithms for new copy replacement in cloud computing. Their approach factors in two main aspects: grouping highly dependent files together to reduce data migration across data centers and considering transfer cost related to file size. The fitness function utilizes total transmission time to determine data transfer amounts. Results show this algorithm reduces data displacement compared to k-means.

Chunlin [22] proposed an algorithm that improves system performance, accounting for file unavailability, data center load, and network transmission costs. It employs a quick sort genetic algorithm to solve the multi-objective copy placement problem. By considering processor capability, memory, disk space, and network bandwidth, this algorithm determines the number and location of copies. The proposed strategy, utilizing a copy transfer approach, outperforms dynamic adjustment strategies (DRAS) [23].

Huang et al. [24] introduced a cost-effective replica placement algorithm under high read/write conditions. It considers storage, update, transmission time, and processing costs. The algorithm determines the number of copies while also finding suitable locations for new versions.

For initializing the population, a heuristic rule based on data support amount and degree is proposed. A hybrid genetic algorithm (HGA) is used, with HGA solutions approaching optimal quality. Navimipour et al. [25] suggest an ant colony strategy to enhance replica selection. Ants choose a center randomly, with subsequent ants attracted to centers with the target file. This strategy significantly reduces access time compared to RTRM by utilizing pheromone information.

Azimi [26] proposed a dynamic data replication algorithm based on the Bee Colony Evolutionary Algorithm (BCDR) for cloud computing environments. The authors considered a hierarchical topology with three levels. There are two levels of connectivity: the first level involves low bandwidth areas

and the second level includes LAN (local area network) areas with higher bandwidth connections. The third level includes the sites of each LAN that are connected to each other through high bandwidth. As a result of the proposed algorithm, honey bees stay in a new location if the food area is better or has more nectar than the previous one, and one unit is added to the index. During the search phase, the worker bees determine which sites have the best probability of containing a file. Based on the number of requests, the best site is determined. Based on the evaluation results, the proposed method reduced the execution time compared to LRU, LFU, and BHR [27].

To guarantee the profitability of the cloud service provider, Mokadem et al. [28] proposed a data replication algorithm. The proposed method consists of two main steps. In the first step, the response time is estimated and then compared to a predetermined threshold. In the second step, if the predicted time exceeds the threshold, the supplier is given a new iteration that will provide the maximum profit to the supplier. Based on simulation results, the proposed algorithm reduces file transfer costs by taking into account processor, storage, network, and service costs.

According to the research conducted by Salem et al. [29], an ABC algorithm-based iteration strategy was developed. To determine the optimal copy space, the proposed algorithm first solves the shortest path problem based on the knapsack problem. To achieve a load balance in the system and save the copy from the shortest path at the lowest cost is the main goal of the project. A second step involves implementing an algorithm for finding the optimal sequence of data replication and determining the best path to data centers based on cost. As compared to the strategy (DCR2S) [30] and genetic algorithm (GA), the introduced strategy can reduce data transmission.

Tos et al. [31] introduce a method (PDR) guaranteeing customer performance and cloud service provider profitability. It estimates query response times and profitability-affecting costs to decide if the operation should be repeated.

In [32], The authors introduce a new replication method called hierarchical data replication strategy (HDRS) in this article. The HDRS algorithm involves creating replicas that can increase or decrease based on exponential growth or decay rate, placing replicas based on access load and labeling technique, and replacing replicas based on the future value of the file. The authors compare various dynamic data replication methods using CloudSim simulation and find that HDRS outperforms other algorithms by reducing response time and bandwidth usage. HDRS can efficiently identify popular files and replicate them to the most suitable site, reducing unnecessary replications and balancing site loads to decrease access latency.

The authors in [33] suggested a new replication management strategy called EIMORM, which is an improvement on the MORM algorithm. EIMORM differs from MORM in two ways: it takes into account the cost of replication when placing replicas and assigns weights to data files to determine popular files based on their last access time. The simulation results show that EIMORM can effectively reduce the total cost, particularly for a large

number of tasks. However, EIMORM does not address the important step of replica replacement when storage space is full.

The authors in [34], proposed ALO-Tabu algorithm that utilizes a hybrid of ant lion optimization and Tabu search algorithms for solving the replica management problem. The process of selecting the initial population is performed in

such way that the algorithm can find better solution.

A comparison of different data replication algorithms' parameters is presented in Table 1. While bandwidth consumption and response time are often prominent, other factors like energy consumption, system load balance, and cost are less emphasized.

Table 1. Parameter comparisons between different data replication algorithms.

Algorithm	Main Idea	Year	Decision making	Replacement	Selection	Cost	Accessibility	Load balancing	Response time	Bandwidth	Energy consumption	Test amount	Metaheuristic method
[21]	Combining similar tasks	2017	-	+	-	-	-	+	+	+	-	5-13 Nodes	GA
[22]	Modeling and using it	2019	-	+	-	+	+	+	+	+	-	1-10 Nodes	GA
[24]	Definition of data classification	2018	-	+	-	+	-	-	+	-	-	10-50 Nodes	EGA
[25]	Accessibility scheme for reading files	2016	-	-	+	-	-	-	+	+	-	5000 Nodes	Ant Colony
[26]	Definition of three-level structure	2019	-	+	+	-	-	-	+	+	-	100-1500 Jobs	ABC
[28]	Profitability of the supplier	2020	+	-	+	+	+	+	+	+	-	500-1500 Nodes	-
[29]	Using the backpack method to solve problems	2019	+	+	+	-	+	+	+	+	-	1-15 Nodes	ABC
[31]	Profitability of the supplier	2016	+	-	+	+	-	-	-	+	-	6 Nodes	-

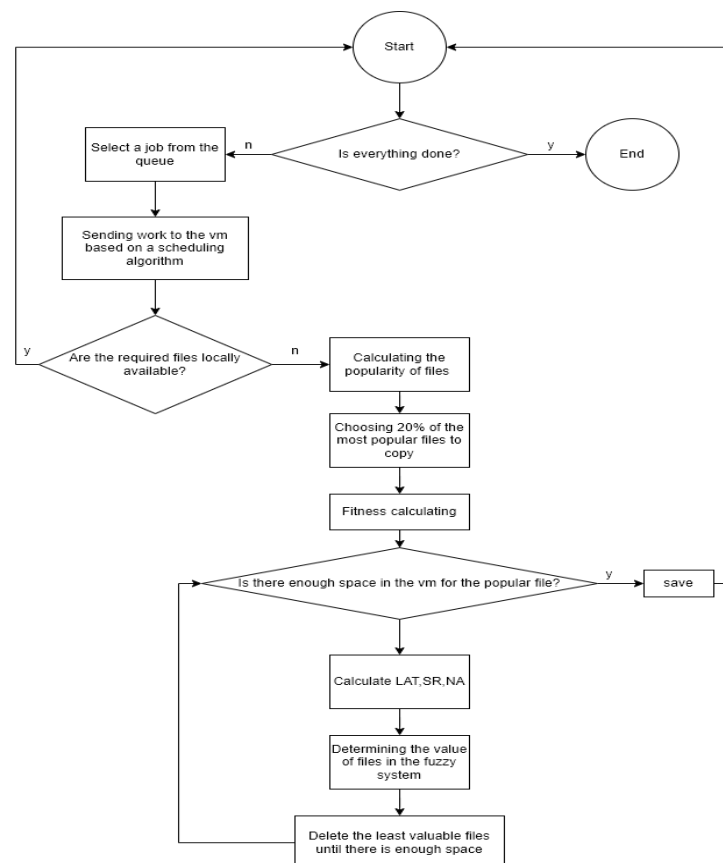


Figure 3. Flowchart of the algorithm

4. The Proposed Method

The algorithm's primary objectives involve determining which files to replicate and when to replicate them. To address this, the algorithm calculates the number of file accesses and the file's popularity. Subsequently, it creates copies for a specific percentage of files based on these factors, with the PSO algorithm guiding the placement of new copies.

To identify candidate files for replication, a value is assigned to each copy, prioritizing less valuable files. This process employs a fuzzy system that takes into account three input parameters: file accesses, copy size, and the time of last access.

The proposed algorithm employs distinct methods for each step, working toward optimal states for individual steps as well as the entire system. Refer to Figure 3 for an overview of the general steps. The algorithm's steps are detailed in the subsequent subsections.

4.1. File Selection and Replication Timing

Cloud system infrastructure constraints, including limited network bandwidth, necessitate careful duplication of files to avoid computational overhead and network congestion. Challenges such as inefficient storage usage, extended task completion times, and security concerns related to user access further underline the importance of an optimal solution. By minimizing the number of files requiring replication, these issues can be mitigated.

To identify less frequently accessed files, the popularity parameter is utilized for selecting the ideal files for replication. This selection process is based on the calculation of each file's popularity, determined using the provided equation:

$$PD_i = (ac_i \times dnc_i) \quad (3)$$

where PD_i represents the popularity of file F_i , ac_i the number of requests, dnc_i the number of data centers that requested the file F_i . Accordingly, the threshold value is determined as follows:

$$T = \frac{\frac{1}{n} \sum_{i=1}^n PD_i}{DN_{count}} \quad (4)$$

where the numerator represents the average file popularity and the denominator represents the number of data centers. When a task is assigned to a data center for execution, if the required files are not available locally, the data replication process is applied. In other words, the popularity of the files and the threshold are calculated first, and a certain percentage of the files whose popularity exceeds the replication threshold is selected. A middle limit (20%) is considered in this research based on the availability of capacity and space.

4.2. Locating a New Replica

PSO algorithm finds new replica locations. Key parameters—service time, load variance, energy use, response time, and cost—are fitness functions. PSO helps solve data replication by modeling and determining optimal solutions.

4.3. Locating the New Replication Site

The PSO algorithm is also used to locate new replications exactly. To begin with, four important parameters that are considered fitness functions and have a great impact on cloud computing efficiency. In this research, the average service time, load variance, energy consumption, average response time, and cost are discussed, and then the PSO algorithm is used to determine the best solution to the data replication problem.

4.4. Average Service Time

In order to increase the system throughput of the system, the average service time must be reduced. Reducing the average service time means increasing the processing speed. It is possible to reduce this average time by placing popular files on high performance nodes and less frequently visited files on low performance nodes. Calculating the service time of file f_i in data node j is as follows [35]:

$$st(i, j) = \frac{\Phi(i, j) \times s_i}{tp_j} \quad (5)$$

where s_i is the file size of f_i and tp_j is the transfer rate of the data node D_j . Each f_i file contains r_i copies distributed across different data nodes. We assume that requests from file f_i are modeled as a Poisson function with average access rate $A(i)$.

So, we have:

$$A(i) = \sum_{j=1}^m A(i, j) \quad (6)$$

where $A(i, j)$ is the access rate of reading requests that are made for the file f_i from the data node D_j . If the file f_i is not in D_j node, we set $A(i, j)=0$. The average service time for the f_i file is calculated as follows:

$$\overline{st}(i) = \sum_{j=1}^m \left(st(i, j) \times \frac{A(i, j)}{A(i)} \right) \quad (7)$$

The average service time is calculated as follows:

$$MST = \frac{1}{n} \times \sum_{i=1}^n \sum_{j=1}^m \left(\Phi(i, j) \times s_i / tp_j \times \frac{A(i, j)}{A(i)} \right) \quad (8)$$

4.5. Load Variance

Load variance is used as a measure to show the load balance of the system. In other words, the lower the load variance, the better the load balance in the system. Since the combination of the access rate and the service time of the file f_i gives the exact amount of its load, this load amount $l(i, j)$ of the file f_i in the data node D_i will be as follows:

$$l(i, j) = A(i, j) \times st(i, j) \quad (9)$$

The load variance (LV) is calculated as follows [33]:

$$LV = \sqrt{\frac{\sum_{j=1}^m (l(j) - \bar{l})^2}{m-1}} = \sqrt{\frac{\sum_{j=1}^m \left(\sum_{i=1}^n A(i,j) \times \frac{S_i}{tp_j} \times \Phi(i,j) - \frac{1}{m} \times \sum_{j=1}^m \sum_{i=1}^n A(i,j) \times \frac{S_i}{tp_j} \times \Phi(i,j) \right)^2}{m-1}} \quad (10)$$

4.6. Energy Consumption

The total energy consumption is mainly composed of renewable energy consumption (RE) and cooling energy consumption (CE), both of which should be kept as low as possible. In today's world, environmental concerns are one of the biggest challenges, especially in industrialized countries. Among the major factors contributing to pollution of the environment is energy consumption, which should be given a lot of attention and the algorithm should strive to reduce it. Servers' power consumption can be described by a linear relationship between energy consumption and efficiency.

To calculate the renewable energy consumption of the data node D_j which we name in the ERE(j) formula, we will use the following equation [36]:

$$E_{RE}(j) = \sum_{i=1}^n \Phi(i,j) \times l(i,j) \times (P_{max}(j) - P_{idle}(j)) + P_{idle}(j), \quad (11)$$

where $P_{max}(j)$ shows the maximum power of the data node D_j in the maximum workload and also $P_{idle}(j)$ the power consumption during idle time, so the renewable energy consumption of all nodes is calculated as follows:

$$E_{RE} = \sum_{j=1}^m E_{RE}(j) \quad (12)$$

For the same reason, if the outside temperature is 30 degrees and the inside temperature is 20 degrees, we can calculate the total cooling energy consumption as follows [36]:

$$E_{CE} = \sum_{j=1}^m E_{CE}(j) \quad (13)$$

where:

$$E_{CE}(j) = E_{RE}(j)/Q, \quad Q = \frac{1}{\frac{T_{out}-1}{T_{in}}} \quad (14)$$

$$l(i,j) = A(i,j) \times st(i,j).$$

According to the above, the total energy consumption (EC) can be calculated as follows:

$$EC = \left(1 + \frac{1}{Q}\right) \times \sum_{j=1}^m \left(\sum_{i=1}^n \Phi(i,j) \times l(i,j) \times (P_{max}(j) - P_{idle}(j)) + P_{idle}(j) \right) \quad (15)$$

4.7. Average Response Time

There is an important role to play in reducing latency in any storage system. High bandwidth reduces the amount of delay

significantly, so, in this study, we only considered reading delay. Based on the fact that each file has several copies, the average delay (\bar{L}_i) is calculated as follows [35]:

$$L_i = \frac{1}{r_i} \times \sum_{j=1}^m \Phi(i,j) \times \frac{S_i}{B(j)} \times A(i,j) \quad (16)$$

where $A(i,j)$ is the percentage of read requests sent from the data node D_j to read the file F_i . $B(j)$ is the minimum bandwidth of the data node D_j , so, we will calculate the average delay time for the whole system as follows [36]:

$$ML = \sum_{i=1}^n L_i/n = \sum_{i=1}^n \left(\frac{1}{r_i} \times \sum_{j=1}^m \Phi(i,j) \times \frac{S_i}{B(j)} \times A(i,j) \right) / n \quad (17)$$

4.8. Cost

A common way to replicate data in traditional systems is to create as many copies as possible to maximize the use of resources to increase overall system performance. In cloud systems, the implementation of this method of data replication is not cost-effective for cloud service providers and can lead to excessive and incorrect use of resources and reduced system efficiency. Maintaining an optimal number of copies saves resources and overall costs, especially for servers, so making as many copies as possible is not always the best option.

Several nodes are used by cloud providers to manage and handle user requests. Each of these nodes requires electricity and some hardware to function properly. These items add to the computational costs (C_i). Another cost is related to the use of the network (C_b). The information required by the requests are continuously sent to different parts of the network and to different cloud destinations globally. Consumable memory (C_s) is another cost-related item that the provider pays for each node to provide an empty space for use and placing copies in it. The cost of each of the items varies from one cloud service provider to another, so it is hard to determine which costs more. As a result, the total amount of expenses (ex) is calculated as follows:

$$ex = C_i + C_b + C_s \quad (18)$$

Now, the fitness function is defined as the sum of the above functions (normalized values) and the best place to store a new replica is determined using the PSO algorithm.

Suppose there are n number of files to be copied and there are m data centers, so a position matrix of the number of particles is created which contains the values zero and one. A value of one indicates that the replication should be located in that data center. For example, in Figure 4, file copy 1 (F_1) should be stored in data center 1 (DC_1) and file 2 (F_2) should be stored in data center 3 (DC_3).

	F ₁	F ₂	F ₃	F ₄
DC ₁	1	0	0	0
DC ₂	0	1	0	1
DC ₃	0	0	1	0

Figure 4. Example of a position matrix

Now, based on the position matrix, the velocity matrix, is also constructed, and its values are placed in the range as follows:

$$V_k^{ij} \in [-V_{max}, +V_{max}],$$

$$i \in \{1, 2, \dots, n\}, j \in \{1, 2, \dots, m\}$$
(19)

where V_{max} represents the maximum speed of all particles and V_k represents the speed matrix of the k -th particle. Moreover, Figure 5 shows the velocity matrix for a particle with the position matrix of Figure 4 and the range $[-1, 1]$.

	F ₁	F ₂	F ₃	F ₄
DC ₁	0.7	-0.3	0.3	-0.8
DC ₂	0.2	0.5	0.1	0.43
DC ₃	-0.2	0.4	-0.4	0.134

Figure 5. Velocity matrix

It should be noted that p_{best} and g_{best} are two special particles. p_{best} shows the local best particle and g_{best} shows the global best particle among all particles. Note that p_{best} and g_{best} particles are updated in each iteration.

The proposed algorithm evaluates the particles based on the fitness function described before. Therefore, for each particle we have a new fitness function. If a particle's new fitness value is better than that particle's p_{best} fitness value, then p_{best} should be replaced by that particle. In addition, the proposed algorithm uses the p_{best} of all particles and replaces the p_{best} with the current g_{best} , which is better than the current g_{best} .

4.9. Determining the Replica to Be Deleted

If there is not enough free space in the selected data center, the system must delete one or more files to free up space, and after each deletion, update the general copy manager and the local copy manager. The general copy manager includes information about all files in all clusters and the local copy manager includes information about all files in its cluster. In the proposed algorithm, three parameters, number of accesses (NA), replica size (SR) and last copy access time (LAT) are considered.

Considering that the fuzzy system is more accurate and simpler during the decision-making process. Therefore, it is very suitable for decision-making problems with several parameters due to considering the interaction between parameters. For this reason, the use of the fuzzy system for the problem of replacing the copy file has received much attention due to the existing complexities [37, 38].

In the proposed method, a value (RV) is assigned to each copy, and this value is based on the input parameters of the fuzzy system, i.e. copy size, number of accesses, and the last access time of the copy. In the next step, the copies are placed in the list based on their value and in ascending order, so the files at the beginning of the list are candidates to be deleted and free up enough space. Table II contains fuzzy rules in which the different states that exist for determining a file replica. For example, when the number of accesses to the file and the size of the copy and the last access time are high, the overall value of that file is high. On the other hand, if the number of accesses and the file size are low and the last access time is average, the overall value of the file is low and

it is placed at the beginning of the list to be deleted.

The fuzzy system used in the proposed method is based on Mamdani system with triangular membership function. There are many membership functions, including Gaussian, etc., but considering that in this problem, many changes are made dynamically in a short period of time, the most suitable function is the triangular function [39].

Considering that linear functions are much easier to design and understand, and also, triangular functions are in the same category. Moreover, due to the dynamic behavior and the need for high speed, the use of these functions will be very useful [40–42].

In general, there are two methods to design a fuzzy system. In the first method, the knowledge of an expert in that field is used, and in the second method, if there is no access to an expert, the learning method can be used. According To the available history of file accesses and the need for only three input parameters, the membership function for the output parameter includes low, medium and high values.

As mentioned earlier, in the fuzzification phase, all input numbers (number of accesses (NA), copy size (SR) and last copy access time (LAT)) after creating input membership functions to the fuzzy set are considered. These steps are shown in Figure 6. De-fuzzification is also shown in Figure 7, where fuzzy rules (if-then rules) are used to process the file, and membership functions are used to convert the fuzzy results to real numbers so that decision can be made.

Table 2. Fuzzy rules

Number of accesses	Copy size	last time the copy accessed	Value
High	High	High	High
High	Middle	High	High
High	High	Middle	High
Low	High	High	Middle
High	Middle	Middle	Middle
Low	Low	Middle	Low
Low	Middle	Low	Low

5. Experimental Results

An 8 GB laptop with an 8th generation Core i5 Intel processor was used to simulate and implement the algorithm. MATLAB 2019 was used to implement the simulation. A description of all the settings and values used to simulate the algorithm can be found in Table III. The number of instructions to execute is between 500 and 4500 Million instructions per second (MIPS) and 100 virtual machines are used to simulate the algorithm.

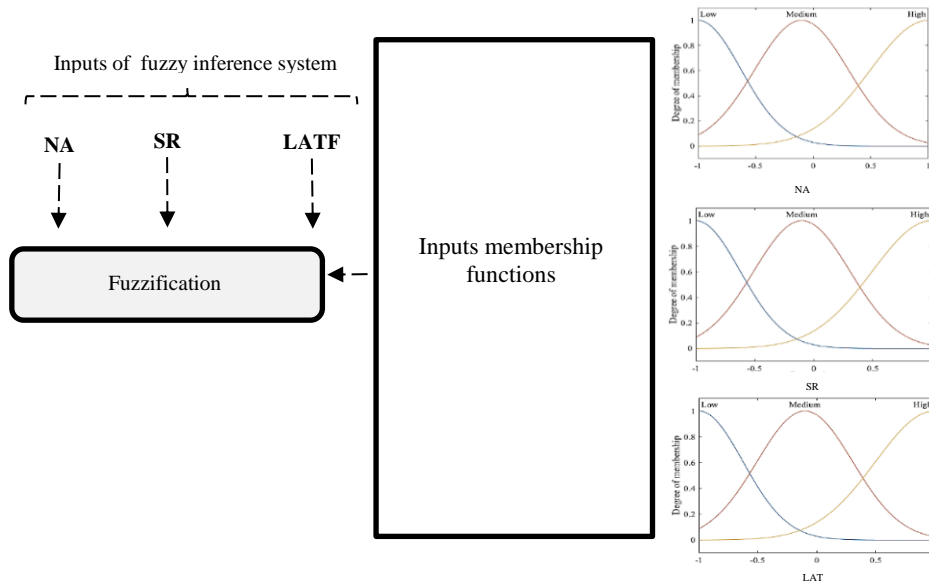


Figure 6. fuzzification steps

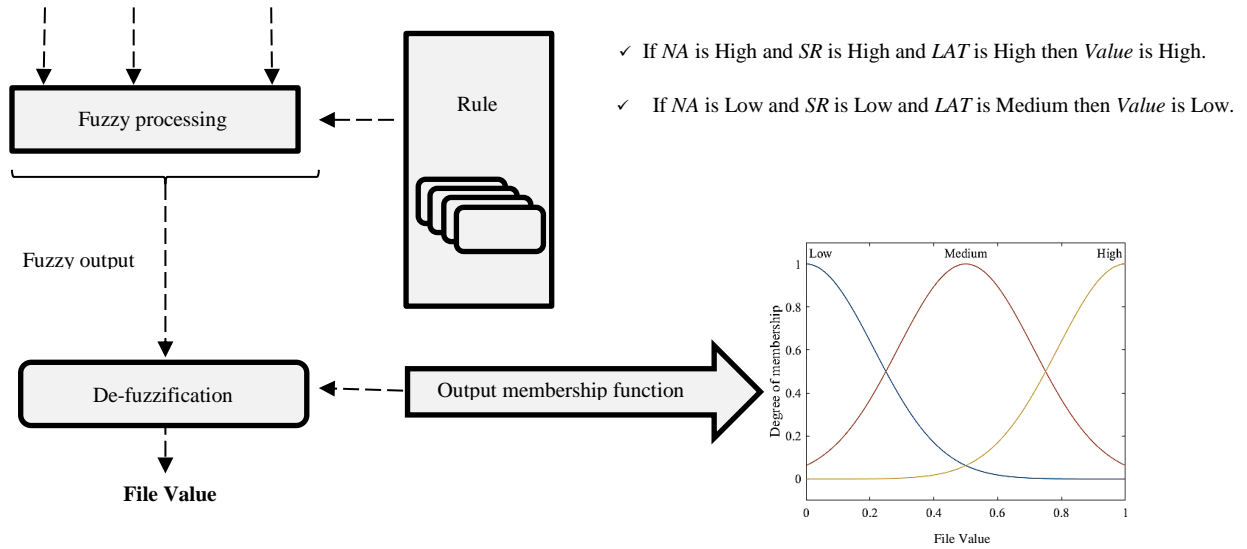


Figure 7. De-fuzzification steps

Table 4. Parameters used for assessment of the method

Parameter	value
Number of data centers	10-50
Number of virtual machines	100
Number of instructions per second (MIPS)	500-4500
number of processing elements per virtual machine	1-4
Each virtual machine's RAM memory	15-35 GB
Total number of tasks	100-500
Length of each task	100-2000 (MI)
Cost of file transfer	\$0.05 per GB
Storage cost	\$0.1 per GB
Processing cost	\$1 per 106 MI

In the following of this section, we will report and discuss the obtained results with respect to different criteria.

5.1. Average Response Time

A response time is defined as the amount of time it takes from the time the job is sent to receiving the response. The average response time is calculated as follows [43]:

$$\text{AverageResponseTime} = \frac{\sum_{j=1}^m \sum_{k=1}^{m_j} (ts_{jk}(rt) - ts_{jk}(st))}{\sum_{j=1}^m m_j} \quad (20)$$

where $ts_{jk}(st)$ and $ts_{jk}(rt)$ indicate the time of sending and receiving task k from/to user j . Moreover, m_j shows the number of jobs of user j .

Three different algorithms with different numbers of tasks are shown in Figure 8. Generally, the average response time is considered to be an important efficiency measure for data replication. As can be seen, the average response time naturally increases as the number of tasks increases, but it

should be noted that the shorter this time, the more efficient the algorithm is. The experimental results in Fig. 8 indicate that the proposed algorithm has the fastest response time compared to BCDR and PDR algorithms for the number of tasks, reducing the response time by approximately 25% and 17%, respectively, indicating a significant improvement in efficiency and performance. One of the most effective reasons for this improvement is to store the most accessed file in the best data centers.

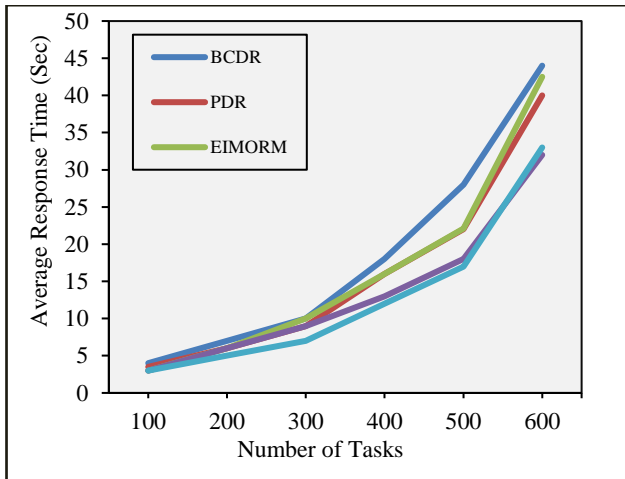


Figure 8 Average response time comparison Energy Consumption

Nowadays, one of the most important issues that has received much attention is the management of energy consumption and trying to reduce the amount of consumption. Considering important issues such as the environment, fuel limitations, high costs, etc. Many experts try to provide various solutions to reduce energy consumption. Most of the existing methods for achieving minimum energy consumption disregard various parameters, which leads to a decrease in system efficiency to some extent.

In the proposed method, as can be seen in Figure 9, the energy consumption for different number of tasks and jobs is always lower than the PDR and BCDR methods. For example, to execute 500 tasks, the proposed algorithm consumes 298 kilojoules of energy, while the results of PDR and BCDR algorithms was 330 and 360 kJ, respectively, and this indicates a 10% and 17% reduction in consumption. One of the reasons for this reduction is the consideration of the energy consumption parameter in the fitness function of the PSO algorithm. Also, other reasons such as increasing workload balance and reducing the number of connections, which will be discussed further, are also effective in reducing energy consumption.

5.2. Load Variance

Usually, load balance in the network is described by a parameter called load variance. Load variance means the standard deviation of data nodes in cloud storage. One of the most important factors influencing the performance of a distributed system is the correct distribution of the load so

that the system is in an optimal state. This parameter increases as the number of files increases. In an optimal method, this value should be minimized as much as possible. Figure 10 shows the load variance for different number of files. As can be seen in this diagram, the variance of the load in the HDR algorithm is much lower compared to the PDR and BCDR algorithms, especially for a high number of files. For example, the load variance for the proposed algorithm for the number of 600 files compared to PDR and BCDR has been reduced by 17 and 40%, respectively. The reason for this decrease is that the files are optimally deployed in the data centers, so, the difference in workload on the data centers is also reduced. Considering that in real environments and in the application space, the number of files is generally very large, which clearly shows the proposed algorithm outperforms compared to other methods.

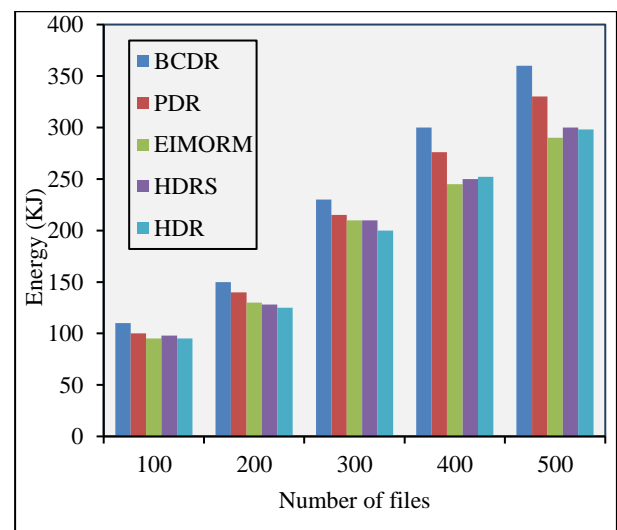


Figure 9. Comparison of energy consumption

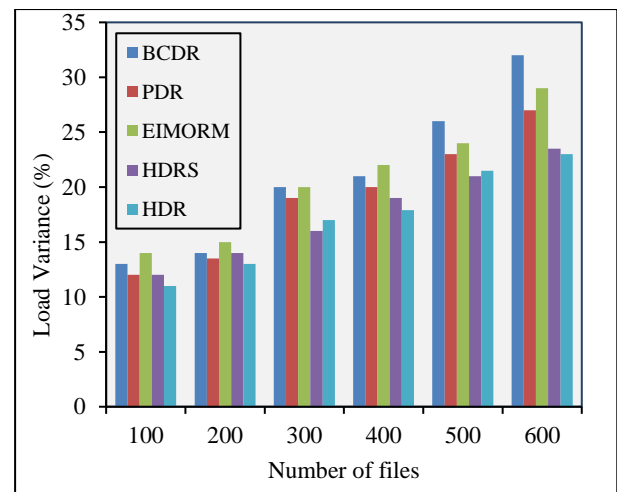


Figure 10. Comparison of load variance

5.3. Number of Connections

Figure 11 shows the number of connections. The number of connections plays an important role in the response time and System efficiency. In other words, reducing the total number of connections, even if it is a small amount, is necessary to reduce data access delay and prevent bandwidth congestion. The HDR algorithm stores the copy file in the best place in

terms of time and space, which will reduce the number of connections. The simulations for similar tasks showed 1459, 1530, and 1605 connections for HDR, PDR, and BCDR algorithms, respectively, representing a 5% and 10% reduction in the number of connections.

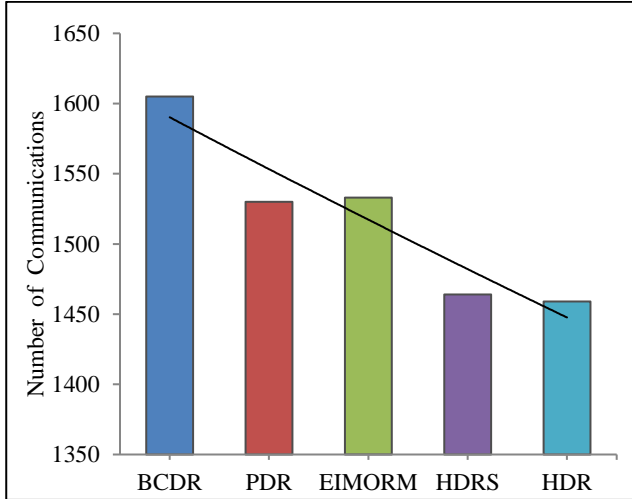


Figure 11. Comparison of load variance

5.4. Hit Ratio

The hit rate is equal to the ratio of the number of local file accesses to the total number of accesses. Total accesses include local file accesses, total number of copies, and total number of remote file accesses. Figure 12 shows the hit rate for 1000 jobs in the compared algorithms. It is easy to understand that the HDR algorithm has the highest rate compared to the PDR and BCDR algorithms. Therefore, in the proposed method, the hit rate has increased by 37% and 57%, respectively, compared to PDR and BCDR algorithms. Because of this increase the number of local accesses to files is increased due to storing copies in the right places and based on the number of accesses to files and creating unnecessary copies, so, the total number of copies and the number of remote accesses to files is reduced ambiguous. The hit rate increases by decreasing the denominator of the fraction.

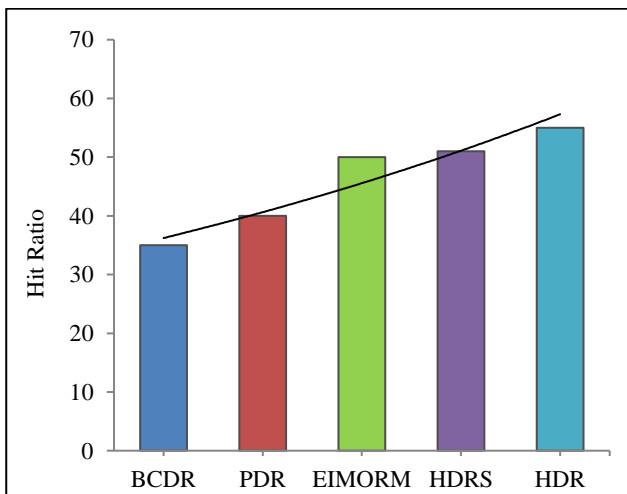


Figure 12. Comparison of hit ratio

5.5. Storage Usage

Storage space is undoubtedly one of the main elements in the cloud, so monitoring the usage of storage resources can provide useful information. This issue can be considered in proposing an efficient replication method from two important perspectives: on the one hand, the goal can be to minimize the storage space consumption, because the cost of resources is proportional to the amount used. On the other hand, the cost may be fixed and the main goal is to maximize the use of storage space.

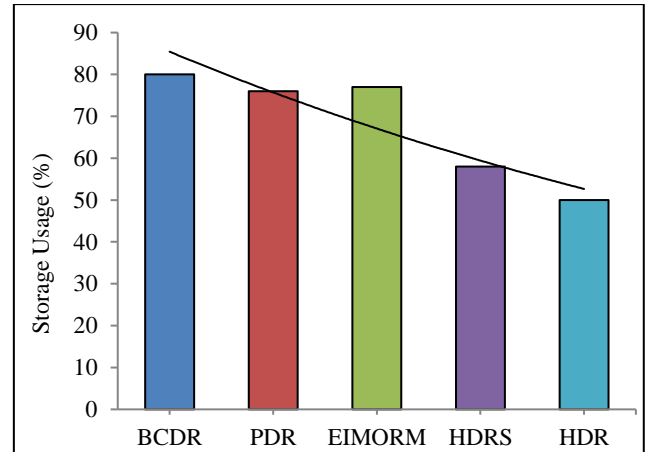


Figure 13. Comparison of storage usage

Figure 13 shows the amount of storage space for the discussed algorithms. As it is clear, in the proposed algorithm, a reduction of 38 and 34 percent of storage space consumption has occurred, respectively, relative to the BCDR and PDR algorithms, which is a significant reduction. The reason for this decrease is two basic things: firstly, considering the file popularity parameter, only popular and frequently used files are copied, so many additional copies cannot be done, and secondly, the file replacement method is used, that is, files that have less value are deleted. Considering the hardware limitations as well as the high costs of preparing, setting up and maintaining storage spaces and considering the high importance of economic justification in the world of information technology, this reduction in storage space consumption is one of the most important features of the proposed method in this research.

5.6. Cost

As mentioned in the previous parts, one of the important parameters in the use of cloud systems is financial issues. Most of today's researches and solutions try to minimize costs for end users, while one of the important issues in this environment is the costs of cloud providers and improving profitability by reducing their costs. These costs include various things such as space and bandwidth consumption, processing and transfer costs of files, as well as maintenance costs. Therefore, providing a method to guarantee the profitability of cloud service providers along with other benefits for users and creating a balance in this environment will be very useful. As seen in Figure 14, for a certain number of tasks, the cost for PDR and BCDR algorithms is \$40 and \$46, respectively, while it is \$32 for the proposed method, and this represents a 20% and 30% reduction in cost.

It is general. One of the important reasons for this reduction is that in the PDR and BCDR algorithms, the files are not copied in the right place, and when requesting a job, the file does not exist locally, and it is necessary to move the file, which incurs various costs, including It involves transfer and processing, but in the proposed method, due to considering the popularity of the file, the copy files have a suitable distribution, and due to the local presence of the files, this problem is avoided to a large extent.

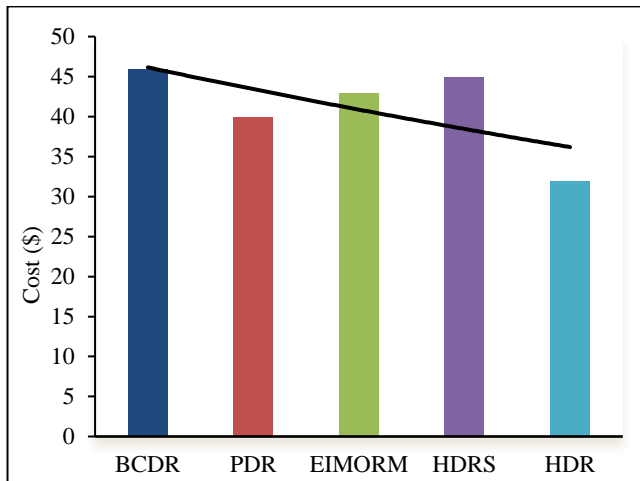


Figure 14. Comparison of Cost

5.7. Average Response Time (ART)

The response time for a datafile is the interval between the submission time of the task and return time of the result. The average response time of a system is the mean value of the response time for all data request tasks of the users. Therefore, for the last experiment, we compare the average response time of the proposed method with others. Figure 15 shows the obtained results.

As shown in Figure 15, the proposed method reports lower ART.

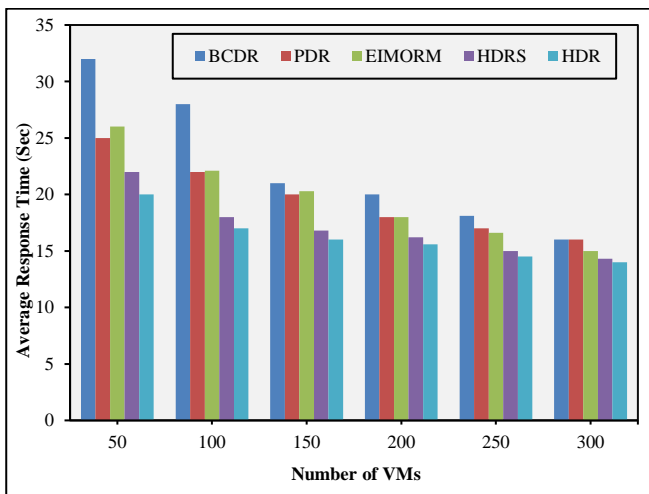


Figure 15. Comparison of average response time

6. Conclusion

The study introduces a novel data replication approach, referred to as HDR, aimed at enhancing data recovery within

the cloud environment. The HDR method aims to decrease program response times, lower connection count, elevate hit rates, establish balanced system loads, and minimize storage consumption. By strategically placing copies near desired data centers and minimizing interconnections, the approach achieves load balancing across the entire system, leading to a remarkable 25% reduction in average response time—a pivotal efficiency metric.

In cloud environments, an optimal system should ensure swift response times, a goal accomplished by simultaneously considering parameters such as access frequency, copy size, and last access time in a fuzzy framework. Moreover, system cost-effectiveness is crucial, given the substantial expenses related to hardware and infrastructure. The research places significant emphasis on this aspect, evident in graphical outcomes demonstrating an average storage space reduction of approximately 35%. As storage expenses significantly impact cloud system costs alongside hardware constraints, the adoption of the proposed HDR method emerges as a potent strategy to substantially curtail overall expenditure.

7. References

- [1] Chauhan, S., Pilli, E., Joshi, R., and Singh, G., "Govil, M.C. Brokering in Interconnected Cloud Computing Environments: A Survey", *J Parallel Distrib Comput*, Vol .133, pp. 193–209, doi:10.1016/j.jpdc.2018.08.001, 2019.
- [2] Qiu, X., Sun, P., and Dai, Y., "Optimal Task Replication Considering Reliability, Performance, and Energy Consumption for Parallel Computing in Cloud Systems", *Reliab Eng Syst Saf*, 215, 107834, doi:10.1016/J.RESS.2021.107834, 2021.
- [3] Shojaiemehr, B., Rahmani, A., Qader, N., "Cloud Computing Service Negotiation: A Systematic Review", *Comput Stand Interfaces*, Vol. 55, pp. 196–206, doi:10.1016/j.csi.2017.08.006, 2018.
- [4] Huang, K., and Li, D., "MRMS: A MOEA-Based Replication Management Scheme for Cloud Storage System", *International Conference on Communications in China*, 2016.
- [5] Moura, D., and Hutchison, D., "Review and Analysis of Networking Challenges in Cloud Computing", *Journal of Network and Computer Applications*, Vol. 60, pp. 113–129, doi:10.1016/j.jnca.2015.11.015, 2016.
- [6] Aznoli, F., and Navimipour, N., "Cloud Services Recommendation: Reviewing the Recent Advances and Suggesting the Future Research Directions", *Journal of Network and Computer Applications*, 2017.
- [7] Slimani, S., Hamrouni, T., Charrada, F., "Service-Oriented Replication Strategies for Improving Quality-of-Service in Cloud Computing: A Survey", *Cluster Comput*, Vol. 24, pp. 361–392, doi:10.1007/S10586-020-03108-Z/METRICS, 2021.
- [8] Bello, S., Oyedele, L., Akinade, O., Bilal, M., Davila Delgado, J., Akanbi, L., Ajayi, A., and Owolabi, H., "Cloud Computing in Construction Industry: Use Cases, Benefits and Challenges", *Autom Constr*, 122, 103441, doi:10.1016/J.AUTCON.2020.103441, 2021.
- [9] Mansouri, N., and Javidi, M., "A New Prefetching-Aware Data Replication to Decrease Access Latency in Cloud Environment", *Journal of Systems and Software*,

- Vol. 144, pp. 197–215, doi:10.1016/J.JSS.2018.05.027, 2018.
- [10] Mansouri, N., "QDR: A QoS-Aware Data Replication Algorithm for Data Grids Considering Security Factors", *Cluster Comput*, Vol. 19, pp. 1071–1087, doi:10.1007/S10586-016-0576-7/METRICS, 2016.
- [11] Sun, S., Yao, W., Li, X., "DARS: A Dynamic Adaptive Replica Strategy under High Load Cloud-P2P", *Future Generation Computer Systems*, Vol. 78, pp. 31–40, doi:10.1016/J.FUTURE.2017.07.046, 2018.
- [12] Madhubala, R. P., "Survey on Security Concerns in Cloud Computing", *Proceedings of the 2015 International Conference on Green Computing and Internet of Things, ICGCIOT 2015* 2016, pp. 1458–1462, doi:10.1109/ICGCIOT.2015.7380697.
- [13] Tao, M., Ota, K., Dong, M., DSARP: Dependable Scheduling with Active Replica Placement for Workflow Applications in Cloud Computing. *IEEE Transactions on Cloud Computing*, Vol. 8, pp. 1069–1078, doi:10.1109/TCC.2016.2628374, 2020.
- [14] Xie, F., Yan, J., Shen, J., "Towards Cost Reduction in Cloud-Based Workflow Management through Data Replication", *Proceedings - 5th International Conference on Advanced Cloud and Big Data, CBD 2017*, pp. 94–99, doi:10.1109/CBD.2017.24, 2017.
- [15] Marini, F., Walczak, B., "Particle Swarm Optimization (PSO)", *A Tutorial. Chemometrics and Intelligent Laboratory Systems*, Vol. 149, pp. 153–165, doi:10.1016/J.CHEMOLAB.2015.08.020, 2015.
- [16] Ojha, V., Abraham, A., Snášel, V., "Heuristic Design of Fuzzy Inference Systems: A Review of Three Decades of Research", *Eng Appl Artif Intell*, Vol. 85, pp. 845–864, doi:10.1016/J.ENGAPPAI.2019.08.010, 2019.
- [17] Aubry, P., Marrez, J., Valibouze, A., "Computing Real Solutions of Fuzzy Polynomial Systems", *Fuzzy Sets Syst*, Vol. 399, pp. 55–76, doi:10.1016/J.FSS.2020.01.004, 2020.
- [18] Guillaume, S., "Designing Fuzzy Inference Systems from Data: An Interpretability-Oriented Review", *IEEE Transactions on Fuzzy Systems*, Vol. 9, pp. 426–443, doi:10.1109/91.928739, 2001.
- [19] Shingne, H., Shriram, R., "Heuristic Deep Learning Scheduling in Cloud for Resource-Intensive Internet of Things Systems", *Computers and Electrical Engineering*, doi:10.1016/j.compeleceng.2023.108652, 2023.
- [20] Zhou, G., Tian, W.H., Buyya, R., Wu, K., "Growable Genetic Algorithm with Heuristic-Based Local Search for Multi-Dimensional Resources Scheduling of Cloud Computing", *Appl Soft Comput*, doi:10.1016/j.asoc.2023.110027, 2023.
- [21] Liu, L., Yang, Y., Wang, H., Tan, Z., Li, C., "A Group Based Genetic Algorithm Data Replica Placement Strategy for Scientific Workflow", *Proceedings - 16th IEEE/ACIS International Conference on Computer and Information Science, ICIS 2017*, pp. 459–464, doi:10.1109/ICIS.2017.7960036, 2017.
- [22] Li, C., Wang, Y.P., Tang, H., Luo, Y., "Dynamic Multi-Objective Optimized Replica Placement and Migration Strategies for SaaS Applications in Edge Cloud", *Future Generation Computer Systems*, Vol. 100, pp. 921–937, doi:10.1016/J.FUTURE.2019.05.003, 2019.
- [23] Nousias, I., Khawam, S., Milward, M., Muir, M., Arslan, T. A Multi-Objective GA Based Physical "Placement Algorithm for Heterogeneous Dynamically Reconfigurable Arrays", *Proceedings - 2007 NASA/ESA Conference on Adaptive Hardware and Systems, AHS-2007*, pp. 504–510, doi:10.1109/AHS.2007.8, 2007.
- [24] Huang, X., Wu, F., "A Cost-Effective Data Replica Placement Strategy Based on Hybrid Genetic Algorithm for Cloud Services", *Lecture Notes in Business Information Processing*, Vol. 327, pp. 43–56, doi:10.1007/978-3-319-99040-8_4/FIGURES/6, 2018.
- [25] Navimipour, N. J., Milani, B. A., "Replica Selection in the Cloud Environments Using an Ant Colony Algorithm", *2016 3rd International Conference on Digital Information Processing, Data Mining, and Wireless Communications, DIPDMWC 2016*, pp. 105–110, doi:10.1109/DIPDMWC.2016.7529372, 2016.
- [26] khalili azimi, S., "A Bee Colony (Beehive) Based Approach for Data Replication in Cloud Environments", *Lecture Notes in Electrical Engineering*, Vol. 480, pp. 1039–1052, doi:10.1007/978-981-10-8672-4_80/COVER, 2019.
- [27] Park, S. M., Kim, J. H., Ko, Y. B., Yoon, W. S., "Dynamic Data Grid Replication Strategy Based on Internet Hierarchy", *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 3033, pp. 838–846, doi:10.1007/978-3-540-24680-0_133/COVER, 2004.
- [28] Mokadem, R., Hameurlain, A., "A Data Replication Strategy with Tenant Performance and Provider Economic Profit Guarantees in Cloud Data Centers", *Journal of Systems and Software*, Vol. 159, 110447, doi:10.1016/J.JSS.2019.110447, 2020.
- [29] Salem, R., Salam, M.A., Abdelkader, H., Awad Mohamed, A., "An Artificial Bee Colony Algorithm for Data Replication Optimization in Cloud Environments", *IEEE Access*, Vol. 8, pp. 51841–51852, doi:10.1109/ACCESS.2019.2957436, 2020.
- [30] Gill, N. K., Singh, S., "A Dynamic, Cost-Aware, Optimized Data Replication Strategy for Heterogeneous Cloud Data Centers", *Future Generation Computer Systems*, Vol. 65, pp. 10–32, doi:10.1016/J.FUTURE.2016.05.016, 2016.
- [31] Tos, U., Mokadem, R., Hameurlain, A., Ayav, T., Bora, S., "A Performance and Profit Oriented Data Replication Strategy for Cloud Systems", *In Proceedings of the 2016 Intl IEEE Conferences on Ubiquitous Intelligence & Computing, Advanced and Trusted Computing, Scalable Computing and Communications, Cloud and Big Data Computing, Internet of People, and Smart World Congress (UIC/ATC/ScalCom/CBDCoM/IoP/SmartWorld)*, pp. 780–787, 2016.
- [32] Mansouri, N., Javidi, M., Zade, B.M.H., "Hierarchical data replication strategy to improve performance in cloud computing", *Front. Comput. Sci*, Vol. 15, 152501. <https://doi.org/10.1007/s11704-019-9099-8>, 2021.
- [33] Edwin, E. B., Umamaheswari, P., Thanka, M. R., "An efficient and improved multi-objective optimized

- replication management with dynamic and cost aware strategies in cloud computing data center", *Cluster Comput* 22 (Suppl 5), pp. 11119–11128. <https://doi.org/10.1007/s10586-017-1313-6>, 2019.
- [34] Zade, B., Mansouri, N., Javidi, M. M., "A new hyper-heuristic based on ant lion optimizer and Tabu search algorithm for replica management in cloud environment", *Artif Intell Rev.*, Vol. 56, pp. 9837–9947. <https://doi.org/10.1007/s10462-022-10309-y>, 2023.
- [35] Sun, D. W., Chang, G. R., Gao, S., Jin, L. Z., Wang, X. W., "Modeling a Dynamic Data Replication Strategy to Increase System Availability in Cloud Computing Environments", *J Comput Sci Technol*, Vol. 27, pp. 256–272, doi:10.1007/S11390-012-1221-4/METRICS, 2012.
- [36] Long, S. Q., Zhao, Y. L., Chen, W., "MORM: A Multi-Objective Optimized Replication Management Strategy for Cloud Storage Cluster", *Journal of Systems Architecture*, Vol. 60, pp. 234–244, doi:10.1016/J.SYSARC.2013.11.012, 2014.
- [37] Singh, H., Gupta, M. M., Meitzler, T., Hou, Z. G., Garg, K.K., Solo, A. M. G., Zadeh, L. A., "Real-Life Applications of Fuzzy Logic", *Advances in Fuzzy Systems 2013*, doi:10.1155/2013/581879, 2013.
- [38] Hu, J., Chen, P., Chen, X., "Intuitionistic Random Multi-Criteria Decision-Making Approach Based on Prospect Theory with Multiple Reference Intervals", *Scientia Iranica*, Vol. 21, pp. 2347–2359, 2014.
- [39] Bai, Y., Wang, D., "Fundamentals of Fuzzy Logic Control — Fuzzy Sets, Fuzzy Rules and Defuzzifications", *Advances in Industrial Control*, pp. 17–36, doi:10.1007/978-1-84628-469-4_2/COVER, 2006.
- [40] Pedrycz, W., "Why Triangular Membership Functions?" *Fuzzy Sets Syst*, Vol. 64, pp. 21–30, doi:10.1016/0165-0114(94)90003-5, 1994.
- [41] Herva, M., Franco-Uría, A., Carrasco, E. F., Roca, E., "Application of Fuzzy Logic for the Integration of Environmental Criteria in Ecodesign", *Expert Syst Appl*, Vol. 39, pp. 4427–4431, doi:10.1016/J.ESWA.2011.09.148, 2012.
- [42] Gulati, S., Pal, A., "Tuning Fuzzy Logic Controller with SGWO for River Water Quality Modelling", *Mater Today Proc*, Vol. 54, pp. 733–737, doi:10.1016/J.MATPR.2021.10.467, 2022.
- [43] López-Pires, F., Barán, B., "Many-Objective Virtual Machine Placement", *J Grid Comput*, Vol. 15, pp. 161–176, doi:10.1007/S10723-017-9399-X/METRICS, 2017.



An Efficient Ramp Secret Sharing Scheme Based on Zigzag-Decodable Codes*

Research Article

Saeideh Kabirirad¹, Sorour Sheidani², Ziba Eslami³

DOI: [10.22067/cke.2023.83418.1094](https://doi.org/10.22067/cke.2023.83418.1094)

Abstract: Secret sharing schemes are ideally suited to save highly sensitive information in distributed systems. On the other hand, Zigzag-Decodable (ZD) codes are employed in wireless distributed platforms for encoding data using only bit-wise shift and XOR operations. Recently, Vandermonde-based ZD codes have been utilized in secret sharing schemes to achieve high computational efficiency such that sharing and recovering of secrets can be realized by lightweight operations. However, the storage overhead of using these ZD codes remains a problem which is addressed in the present paper. Here, a ramp secret sharing scheme is proposed based on an efficient ZD code with less storage overhead in comparison with existing literature. The novelty of the proposed scheme lies in the careful selection of the number of positions to shift the bits of the secret such that security and zigzag decodability are guaranteed simultaneously. In addition to prove gaining these features, we show that the scheme improves speed of recovery.

Keywords: Boolean operation, Efficiency, Ramp secret sharing scheme, Zigzag decodable codes.

1. Introduction

Security is a vital necessity in distributed systems and cloud environments. With the rise in cloud computing environments and Internet of things systems, secret sharing (SS) schemes have become increasingly important cryptographic primitives. In an SS scheme, a secret is distributed to some shares such that any qualified subset of shareholders can reconstruct the secret while no unqualified subset can gain any information about it. SS schemes are used as the key element of numerous security protocols, most notably in distributed storage systems, threshold cryptography and secure multi-party computation. However, in many of these

applications, lightweight schemes are a requirement. In this context, there have been efforts to present the schemes that can be implemented using only Boolean operations, namely Shift and XOR.

(t, n) -threshold SS schemes constitute an important type of secret sharing schemes. During these schemes, a secret is

distributed among n participants in a way that any t or more participants be able to reconstruct the secret [1]. A large number of methods have been applied to improve the efficiency of threshold schemes in which using the boolean operations in the schemes establish a considerable part of it [2].

In [3], Shiina et al. presented a (t, n) -threshold SS scheme to improve Shamir's threshold SS scheme. In spite of the improvement in terms of computational time, their method imposes a large storage overhead for the shares. Kurihara et al. [4] presented a $(3, n)$ -threshold SS scheme using XOR that was *ideal* (i.e., the domain of the shares and the secret are equal). Then, they generalized their method in [5] to any arbitrary threshold value by presenting an ideal (t, n) -threshold SS scheme using boolean operations. They also extended the (t, n) -threshold SS and provided the concept of a (t, w, n) -Ramp Secret Sharing (RSS) scheme [6] where w determines a boundary for the minimum number of participants who have to form a coalition to achieve some information about the secret [7].

In Kurihara et al.'s schemes, the shares are computed by applying XOR operation to the secret pieces and the sequences of random bits. In their schemes, after collecting the required number of shares, they are saved in a vector. Then, the secret can be reconstructed through multiplying this vector and a matrix calculated according to Gaussian elimination. Although Gaussian elimination imposes high computational complexity in practice, for some parameters, it is experimentally shown that this elimination is faster than Shamir's scheme. [8] discusses the subject of the scheme of Kurihara et al.'s using the properties of circular matrices. The authors achieve a new security analysis for these secret sharing schemes. Afterward, Wang and Desmedt [9] presented a (t, n) -threshold scheme which needed just XOR and cyclic shift operations. Additionally, Chen et al. [10] who recommended a boolean-based (t, n) -threshold scheme, claimed that it is more efficient than Kurihara et. al's and also Wang and Desmedt's schemes. However, their scheme has a limitation and cannot be used in general case, i.e. it works

* Manuscript received: 2023 July 14, Revised, 2023 August 20 , Accepted, 2023 September 30.

¹ Corresponding author. Assistant professor, Department of Computer Science, Birjand University of Technology, Birjand, Iran.

Email: kabiri@birjandut.ac.ir.

² Ph.D., Department of Data and Computer Sciences, Shahid Beheshti University, G.C., Tehran, Iran.

³ Associate professor, Department of Data and Computer Sciences, Shahid Beheshti University, G.C., Tehran, Iran.

only for $n = t, t + 1$. Shima et al. [11] suggested a way to fix this problem and then extended the improved method to a hierarchical scheme. In another work [12], a (n, n) -threshold secret sharing scheme based on binary trees and XOR operation has been proposed. Some studies applied boolean operations to optimize common SS schemes such as Shamir's method [13].

Boolean-based methods have also been considered for large secrets such as images [14]–[16]. The threshold schemes employed by lightweight operations such as shift and XOR benefit from high computational efficiency, lossless image recovery, ability of multiple images sharing, supporting any image formats, fast recovery, etc [17]–[20].

Lately, zigzag-decodable (ZD) erasure codes have been employed in conjunction with ramp secret sharing schemes. ZD codes are the first XOR-based Maximum Distance Separable (MDS) codes proposed for distributed storage systems [21] that can correct node failures. In ZD codes, both encoding and decoding processes are carried out exclusively by the operators of bitwise shift and XOR, without the need for finite field multiplication. Also, it is possible to perform decoding process easily via physical layer network coding [22]. These advantages imply that in a wide variety of applications where efficiency is important (such as big-data storage, wireless distributed storage systems and resource-constrained devices), ZD-based schemes can be conveniently employed [23], [24].

However, this efficiency is at the expense of some storage overhead. In [25], Gong et al. proposed a (t, w, n) -RSS scheme which has been adapted from ZD codes based on Vandermonde matrix, abbreviated as VZD-RSS. Their scheme inherited the features of ZD codes, i.e., it has less computational complexity compared to the schemes in the literature and also has some storage overhead.

In [21], the authors presented a sufficient condition, so-called "increasing difference" property, for enabling zigzag decodability. Based on this condition, two constructions based on the Vandermonde and the Hankel matrices were proposed [26]. Afterwards, Dai et al. [27] presented another ZD code using circular matrices with less storage overhead. Their method does not provide the sufficient condition and they demonstrate a new proof for constructing feasible ZD code.

1.1. Our Contributions

The main contribution of the present paper is proposing a secure (t, n) -RSS scheme based on ZD codes with less storage overhead than VZD-RSS. That is, we introduce a new generator matrix that reduces size of shares compared to the Vandermonde-based schemes while maintaining security. Here, our contributions in comparison with the existing literature are briefly listed:

1. Our proposed scheme supports arbitrary values of n and t , has low computational complexity, and provides zigzag decodability.
2. Our proposed scheme reduces storage overhead by

almost half compared to Vandermonde-based scheme. The overhead can be neglected for large secrets or cases where n and t are close.

3. Our proposed scheme is proven to achieve zigzag decodability property according to new conditions.
4. Our proposed scheme satisfies security requirements.

1.2. Paper Organization

The organization of the rest of the paper is as follows: Section II reviews some definitions of RSS schemes. Section III introduces ZD codes. The proposed scheme is presented in Section IV. Section V describes security analysis and conditions of zigzag decodability. In Section VI, the efficiency of the proposed scheme is analyzed and compared with boolean-based methods in the literature. Conclusions are presented in Section VII and finally, we provide the details of parts of our proofs in the Appendix section.

2. C Preliminaries

In this section, we review required definitions and provide necessary notations for ramp secret sharing (RSS) schemes.

Let X and Y be two jointly distributed random variables. Let $H(X)$ denote the *Shannon entropy* of X and let $H(X|Y)$ be the *conditional entropy* of X given Y .

According to Shannon entropy, $H(X|Y) = 0$ indicates that X is a deterministic function of Y . However, $H(X|Y = y) = H(X)$ indicates that in case $\{Y = y\}$, no information about X is leaked.

Definition: Assume t, w and n are integers where $0 < w \leq t \leq n$. A (t, w, n) -RSS scheme distributes a secret message K among n participants such that two conditions hold:

1. *Decodability.* Any subset A of t or more participants, can uniquely recover K , i.e., $H(K|A) = 0$.
2. *Secrecy.* Any set A' of at most $(t - w)$ participants, obtains no information about K , i.e., $H(K|A') = H(K)$.

By definition, a $(t, 1, n)$ -RSS scheme is a (t, n) -threshold secret sharing scheme. In fact, RSS schemes are solutions to reduce the size of shares while losing secrecy to some extent.

Definition: A (t, w, n) -ramp secret sharing scheme is *linear* if for any subset of participants A that $|A| = r$ and $t - w < r < t$, we have $H(K|A) = \frac{w-r}{w} H(K)$.

It means that, after pooling $(t - w)$ shares, every further share reveals $\frac{1}{w}$ bits of information about the secret K [27].

3. Review of ZD Codes

The encoding and decoding processes of ZD codes are based solely on boolean operations, including shift and XOR.

Zigzag decodability is the ability of recovering the original data by zigzag decoding [28]. In this section, the coding and decoding processes of ZD codes are reviewed in general.

3.1. Coding

Given a message K with length $\lambda = tL$ bits, we split it into t

pieces K_1, K_2, \dots, K_t .

The bit-length of each piece of message K_i is L bits. Furthermore, polynomial representation of K_i is:

$$K_i(z) = K_{i,0} + K_{i,1}z + \dots + K_{i,(L-1)}z^{L-1} \quad (1)$$

where $K_{i,j}$ is an element in $GF(2)$.

By linear combination of the t pieces of the message, n encoded packets $C_1(z), C_2(z), \dots, C_n(z)$ are generated.

Each $C_i(z), i = 1, 2, \dots, n$ is $L' = L + l$ bits long, where l denotes the storage overhead, i.e., the encoded packets are l bits longer than the pieces of message. Then, the polynomial representation of $C_i(z)$ is given by:

$$C_i(z) = C_{i,0} + C_{i,1}z + \dots + C_{i,(L'-1)}z^{L'-1} \quad (2)$$

Each $C_i(z)$ is generated in two phases: 1) shifting pieces of message and 2) adding them. Hence, the i -th encoded packet is obtained as follows:

$$C_i(z) = z^{e_{i,1}} K_1 + z^{e_{i,2}} K_2 + \dots + z^{e_{i,t}} K_t \quad (3)$$

where $e_{i,j} \in \mathbb{Z}$. Note that multiplying by z^j means shifting by j positions while add operation (performed in $GF(2)$) means XOR.

According to (3), storage overhead of each encoded packet will be $l = \max_{i,j} \{e_{i,j}\}$. Considering the source and encoded data, the corresponding matrix form is:

$$C(z) = G(z) \times K(z) \quad (4)$$

where $C(z)$ is a vector of length n and its i -th element is $C_i(z)$. Additionally, $K(z)$ is a t -dimensional vector containing pieces of message.

$G(z)$ is called the generator matrix and is an $n \times t$ matrix with $z^{e_{i,j}}$ as (i,j) -th element. Note that matrix $G(z)$ is t -reliable, this implies that any $t \times t$ submatrix of $G(z)$ can be used to recover the pieces of message.

So far, there are some suggestions for matrix $G(z)$ in the literature, such as Vandermonde, Hankel, etc. In VZD-RSS [25], choosing Vandermonde matrix has fulfilled the security requirements and the storage overhead equals $(n-1)(t-1)$. In Section IV, we propose a generator matrix such that the storage overhead is reduced by half compared to VZD-RSS.

3.2. Zigzag Decoding

Suppose that t arbitrary coded packets are available. We now describe how the source packets are recovered by zigzag decoding algorithm.

First, a $t \times t$ submatrix $M(z) = [z^{g_{i,j}}]$ of $G(z)$ is constructed using the corresponding indices of the available encoded packets. Consider a $t \times t$ integer matrix $E = [e_{i,j}]$ in which its elements are exponents of the corresponding elements in $M(z)$. The main idea of zigzag decoding algorithm is to find an encoded packet that has a bit which

can be directly extracted. Such a bit is called an "exposed" bit.

Afterwards, the bit is deduced from other encoded packets. This process is done repeatedly until recovering all source bits. In Figure 1, an example of zigzag decoding procedure with two encoded packets is shown. The computational complexity of zigzag decoding is $O(t^2L)$.

In the following, we review the details of zigzag decoding algorithm as stated in [21].

Let i be the index of an encoded packet and similarly, j be the index of a source packet. Also, let m and m' be the set of indexes of the encoded packets and set of indexes of unrecovered source packets. The polynomials $\hat{x}_j(z)$ and $y_i(z)$ are the decoded portion of j -th source packet and also the not decoded part of i -th packet.

Furthermore, for a polynomial $f(z)$, consider $\Omega(f(z))$ and $\omega(f(z))$ as the term with the smallest order and the order of that term, respectively.

Zigzag Decoding Algorithm:

Step 1: (Initialization) Let $m' := m$ and $\hat{x}(z) := 0$. Let

$$\eta_j(z) := 1 + z + \dots + z^{L+l-1}, \text{ for all } j \in m' \quad (5)$$

Step 2: (Searching for an exposed bit) Find an $i^* \in m$ and some $j^* \in m$ such that

$$\omega(z^{t^*j^*} \eta_{j^*}(z)) < \omega(z^{i^*j} \eta_j(z)) \text{ for all } j \in m' \setminus \{j^*\} \quad (6)$$

Step 3: (Updating variables)

1. Let $\hat{x}_{j^*}(z) := \hat{x}_{j^*}(z) + \Omega(y_{i^*}(z))$.
 2. Let $y_i(z) := y_i(z) - z^{i^*j^*} \Omega(x_{j^*}(z))$ for all $i \in m$.
 3. Remove the term of $\eta_{j^*}(z)$ which has the smallest order.
- If there is no more term in $\eta_{j^*}(z)$, delete j^* from m' .

Step 4: If $m' \neq \emptyset$ go to Step 2, else exit and output $\hat{x}_j(z)$ for all $j \in m$.

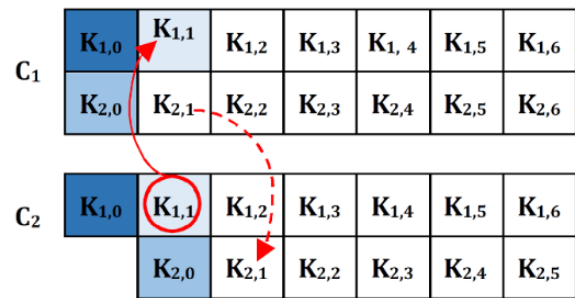


Figure 1. Illustration of zigzag decoding for two coded packets.

4. The Proposed (t,w,n)-RSS Scheme

In this section, we first explain our ZD code that decreases storage overhead in the recovery phase. This is achieved by substituting the ZD code's Vandermonde matrix by another modified matrix. Then, we propose a (t,w,n) -RSS scheme using this ZD code. Our proposed generator matrix G is defined as (7).

$$G = \left(\begin{array}{ccc|ccc} 1 & \dots & z^{t-w-1} & z^{t-w} & z^{t-w+1} & \dots & z^{t-1} \\ 1 & \dots & z^{2(t-w-1)} & z^{2(t-w)} & z^{2(t-w+1)} & \dots & z^{2(t-1)} \\ \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & \dots & z^{\lfloor \frac{n}{2} \rfloor (t-w-1)} & z^{\lfloor \frac{n}{2} \rfloor (t-w)} & z^{\lfloor \frac{n}{2} \rfloor (t-w+1)} & \dots & z^{\lfloor \frac{n}{2} \rfloor (t-1)} \\ \hline z^{t-w-1} & \dots & 1 & z^{t-1+1} & z^{t-2} & \dots & z^{t-w} \\ z^{2(t-w-1)} & \dots & 1 & z^{2(t-1)+1} & z^{2(t-2)} & \dots & z^{2(t-w)} \\ \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ z^{\lfloor \frac{n}{2} \rfloor (t-w-1)} & \dots & 1 & z^{\lfloor \frac{n}{2} \rfloor (t-1)+1} & z^{\lfloor \frac{n}{2} \rfloor (t-2)} & \dots & z^{\lfloor \frac{n}{2} \rfloor (t-w)} \end{array} \right) \quad (7)$$

In this matrix, the first $\lfloor \frac{n}{2} \rfloor$ rows constitute a Vandermonde matrix and the last $\lfloor \frac{n}{2} \rfloor$ rows are obtained by using the upper half of G .

In another representation, we can define G as four submatrices as (4).

$$G = \begin{pmatrix} V([z, \dots, z^{\lfloor \frac{n}{2} \rfloor}], [0, \dots, t-w-1]) & V([z, \dots, z^{\lfloor \frac{n}{2} \rfloor}], [t-w, \dots, t-1]) \\ V([z, \dots, z^{\lfloor \frac{n}{2} \rfloor}], [t-w-1, \dots, 0]) & V([z, \dots, z^{\lfloor \frac{n}{2} \rfloor}], [t-2, \dots, t-w]) \end{pmatrix} \quad (8)$$

Where $V(x, b)$ is an $r \times c$ matrix, $x = (x_1, x_2, \dots, x_r)$ is an r -tuple of indeterminates and $b = (b_1, b_2, \dots, b_c)$ is an c -tuple of non-negative integers. The elements of matrix are obtained as $V(x, b) = [x_i^{b_j}]$.

As can be seen $\max_{i,j} \{g_{i,j}\} = z^{\lfloor \frac{n}{2} \rfloor (t-1)+1}$ which is approximately 1/2 of that of Vandermonde matrix, this reduces the storage overhead by half.

4.1. Sharing Phase:

Consider a secret K with wL bits. First, K is divided into w segments $K_i, i = 1, 2, \dots, w$, in which K_i is L bits long. Then, each segment is considered as the coefficients of a polynomial of order at most $L - 1$. Next, the following steps are performed to generate n shares:

1. Generator matrix G with dimension $n \times t$ is produced as in (7).
2. $(t - w)$ random strings, $R_i, i = 1, 2, \dots, t - w$ with L' bits are generated. Each string R_i is represented by the coefficients of a polynomial:

$$R_i(z) = R_{i,0} + R_{i,1}(z) + \dots + R_{i,L'-1} z^{L'-1} \quad (9)$$

3. Each share $S_i, i = 1, 2, \dots, n$ is calculated as follows:

$$S_i(z) = \begin{cases} \sum_{r=1}^{t-w} (R_r(z) z^{(i-1)(r-1)}) \\ + \sum_{m=1}^w (K_m(z) z^{(i-1)(t-w+m-1)}) \bmod z^{L'} & \text{if } i < \lfloor \frac{n}{2} \rfloor \\ \sum_{r=1}^{t-w} (R_r(z) z^{(i-1)(t-w-r)}) + K_1(z) z^{i(t-1)+1} \\ + \sum_{m=2}^w (K_m(z) z^{(i-1)(m-1)}) \bmod z^{L'} & \text{otherwise} \end{cases} \quad (10)$$

where $\bmod z^{L'}$ denotes the truncation at degree L' .

Alternatively, the encoding can be illustrated by the following notation. At first, multiplication of a matrix-vector is calculated by:

$$S'(z) = G \times \begin{pmatrix} R_1 \\ R_2 \\ \vdots \\ R_{t-w} \\ K_1 \\ K_2 \\ \vdots \\ K_w \end{pmatrix} \quad (11)$$

Then, each share S_i is obtained by truncating $S'_i(z)$ at degree L' .

4.2. Recovery Phase:

With having t shares $S_{i_1}, S_{i_2}, \dots, S_{i_t}$ the secret can accurately be recovered. Let set $I = \{i_1, i_2, \dots, i_t\}$ contain indices of the available shares. The recovery steps can be performed by zigzag decoding algorithm explained in Section III-B.

5. Security Analysis

In this section we prove that the decodability and secrecy requirements are guaranteed in our proposed scheme. Theorem 2 shows that we can recover the secret by t shares, and additionally, by $(t - w)$ or fewer shares no information of the secret is gained. Theorem 4 provides proof for zigzag decodability. For the sake of simplicity, we make use of some lemmas as well.

Lemma 1. Any $t \times t$ square submatrix of the matrix G (defined in Section IV) is invertible, or equivalently, any t rows of G are linearly independent.

Proof: The proof of this lemma can be found in the Appendix section of the paper.

Theorem 2. Let V_A denotes the set of shares corresponding to a random subset of participants A . So, we have:

$$H(K|V_A) = \begin{cases} H(K) & m \leq t - w \\ 0 & m \geq t \end{cases} \quad (12)$$

Where $m = |A|$, i.e. number of participants of A .

Proof: Let $A = \{P_{t_0}, P_{t_1}, \dots, P_{t_{m-1}}\}$. The generator matrix $G = [U \ V]$ is defined such that we have:

$$S' = G \times \begin{bmatrix} R \\ K \end{bmatrix} = (U \times R) \oplus (V \times K) = [S'_1, S'_2, \dots, S'_n]^T \quad (13)$$

where $R = [R_1, R_2, \dots, R_{t-w}]^T$ and $K = [K_1, K_2, \dots, K_w]^T$. According to Step 3 of Sharing phase, each share S_i is directly obtained by truncation of S'_i .

Suppose that R is selected uniformly and that K and R are mutually independent. According to Lemma 1, any t rows of G are linearly independent. Also, any $(t - w)$ rows of U and any w rows of V are linearly independent, i.e. for $m \leq (t - w)$: $\text{rank}(G) = \text{rank}(U) = m$. Hence all elements obtained by $U \times R$ are random and mutually independent. Then, we suppose that S' is a certain subset of the shares that can be obtained with uniform probability from any chosen $V \times K$. Therefore, K is independent of S' and $H(K|V_A) = H(K)$ is satisfied if $m \leq (t - w)$. This means that no information about K can be extracted.

If $m \geq t$, then $\text{rank}(G) = t$. Therefore, solving the system of linear equation 13 specifies the elements of R and K uniquely. This means that any t shares are able to recover the secret.

In the following, we prove that the zigzag decoding algorithm can be applied on shares produced by the proposed scheme. Since the proposed generator matrix does not satisfy the sufficient condition for zigzag decodability (i.e. increasing difference property), a new proof is required.

In [22], new conditions are presented for ZD codes which can reduce storage overhead provided that the matrix is selected correctly. The following results give necessary conditions for zigzag decodability.

First for $p \in \{1, 2, \dots, n\}$ and $m, r \in \{1, 2, \dots, t\}$, define $\Delta_{m,r}^p = g_{pm} - g_{pr}$.

Lemma 3. Assume that G is a generator matrix of a ZD code such that for any row indices i and j , and for any column indices m and r of G , where $i \neq j$ and $m \neq r$, we have:

- 1) $\Delta_{m,r}^i \neq 0$;
- 2) $\Delta_{m,r}^i \neq \Delta_{m,r}^j$;
- 3) If $g_{im} > g_{jm}$ and $\Delta_{m,r}^i > 0$, then $\Delta_{m,r}^i > \Delta_{m,r}^j$.

Then, the original message can be reconstructed by the zigzag decoding algorithm.

Proof: This lemma is proved in Theorems 1 and 2 of [22].

Theorem 4. Assume that $t > 2$ zigzag decoding algorithm can be applied in our (t, w, n) -RSS scheme, i.e. the existence of a share (encoded packet) containing an exposed bit is always guaranteed.

Proof: For better observation, we write the difference value between consecutive components of matrix G (defined by (7)) as follows:

$$d = \left(\begin{array}{ccc|ccc} 1 & \dots & 1 & 1 & & 1 & \dots & 1 \\ 2 & \dots & 2 & 2 & & 2 & \dots & 2 \\ \vdots & \ddots & \vdots & \vdots & & \vdots & \ddots & \vdots \\ \lceil \frac{n}{2} \rceil & \dots & \lceil \frac{n}{2} \rceil & \lceil \frac{n}{2} \rceil & & \lceil \frac{n}{2} \rceil & \dots & \lceil \frac{n}{2} \rceil \\ -1 & \dots & -1 & t-1+1 & -2 & -1 & \dots & -1 \\ -2 & \dots & -2 & 2(t-1)+1 & -3 & -2 & \dots & -2 \\ \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ -\lceil \frac{n}{2} \rceil & \dots & -\lceil \frac{n}{2} \rceil & \lceil \frac{n}{2} \rceil(t-1)+1 & -\lceil \frac{n}{2} \rceil - 1 & -\lceil \frac{n}{2} \rceil & \dots & -\lceil \frac{n}{2} \rceil \end{array} \right) \quad (14)$$

Note that $d_{i,m} = \Delta_{m,(m+1)}^i$.

It can be easily seen that the first condition of Lemma 3 holds.

Now we check condition 3. There are four cases for indices of rows i and j . In the two first cases, rows i and j both are in the upper half or the lower half G . In these cases, condition 3 is satisfied.

For the third case, consider i -th row in the upper half of matrix G and j -th row in the lower half. Without loss of generality consider $m = t - w - 1$ and $r < m$ (i.e. the left part of the matrix). For these values of m, r , we have $g_{im} > g_{jm}$, $\Delta_{m,r}^i > 0$ and $\Delta_{m,r}^j < 0$ and therefore $\Delta_{m,r}^i > \Delta_{m,r}^j$. Similarly, for $m = t - 1$ and $r < m$ (i.e. the right part of the matrix), we have $g_{im} > g_{jm}$, $\Delta_{m,r}^i > 0$ and $\Delta_{m,r}^j < 0$ and therefore $\Delta_{m,r}^i > \Delta_{m,r}^j$.

For the last case, i.e. i -th row in the lower half of matrix G and j -th row in the upper half, it can be easily seen that the condition holds. As for condition 2, in matrix d , rows i, j ($i \neq j$) in upper half of G have different differences $\Delta_{m,r}^i = i(r - m)$ and $\Delta_{m,r}^j = j(r - m)$.

Similarly, we see this for rows i, j in lower half of G . Another case is that i, j are in different parts, for example i in upper half of matrix G and j in the lower half. If m, r both are in the left part or the right part, $\Delta_{m,r}^i > 0$ and $\Delta_{m,r}^j < 0$ and therefore the condition holds. However, condition 2 may not hold when i, j are in different parts (the top and down part of G) and m, r are in different parts (the left and right part of G). But, we show that it cannot prevent the progress of ZD algorithm.

Consider we have two rows i, j and two columns m, r with $\Delta_{m,r}^i = \Delta_{m,r}^j$. Without loss of generality, consider i, j in the upper and lower half of G and $m = 1, r = t - w + 1$ (in the left and the right part of G). According to the sharing algorithm, R_1 and K_1 are multiplied by 1-th and $(t - w + 1)$ -th column of G , respectively. Now, suppose that the zigzag decoding algorithm runs on t shares including S_i, S_j . Since $\Delta_{m,r}^i = \Delta_{m,r}^j$, both i -th and j -th shares include $R_1 \oplus K_1$. It means that bits of R_1 and K_1 can not decode only by S_i and S_j and should use another row p where $\Delta_{m,r}^p \neq \Delta_{m,r}^i$ or $\Delta_{m,r}^j$. In the following, we show that all other rows have different differences from i, j .

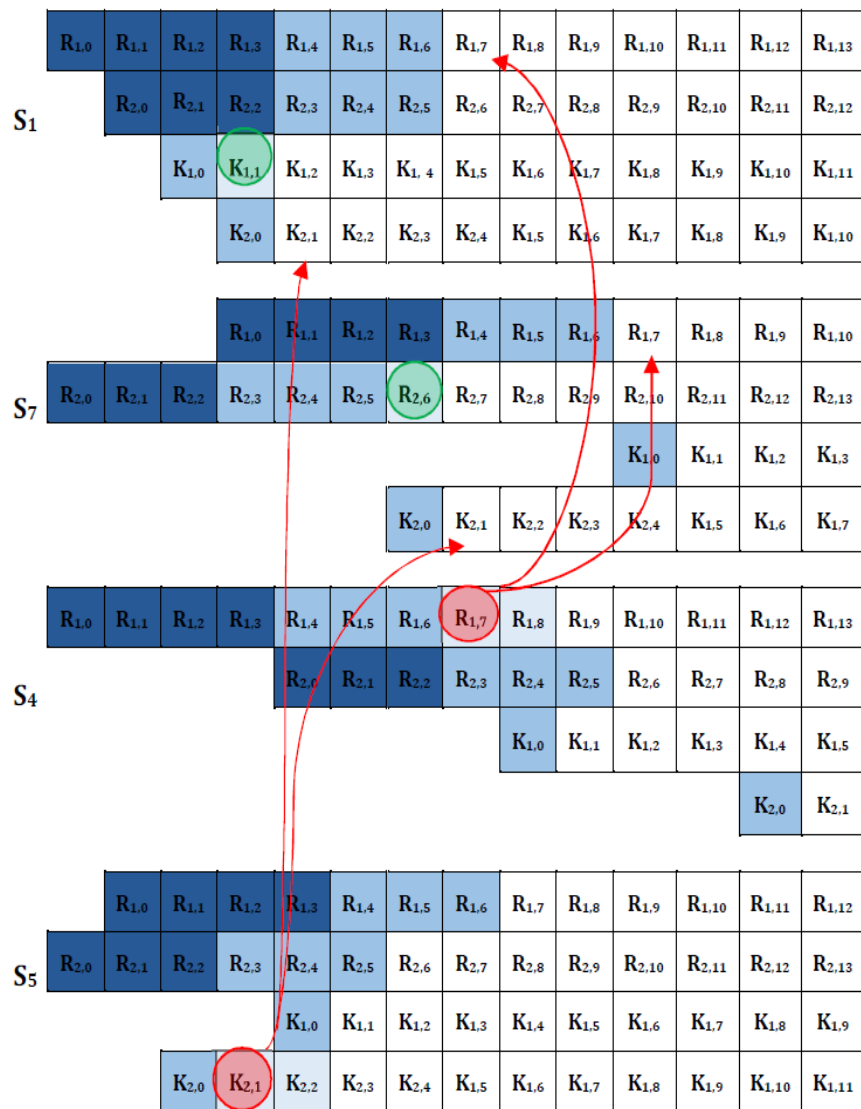


Figure 2. Recovering the secret from the proposed (4, 2, 6)-RSS scheme via zigzag decoding algorithm

Table 1. Comparison of some Boolean-based SS and RSS schemes.

Schemes	No. of computations		Support for all n, t	Only Boolean operation	Storage overhead
	Sharing phase	Recovery phase			
Kurihara et al.	$tn_p L$	$tn_p L + t^3 n_p^3$	Yes	Yes	$l < w(n_p - 1)$
Wang-Desmedt	tnL	nL^3	Yes	No	$l \leq (n - 1)$
Chen et al.	tnL	tnL	No	No	No
Deshmukh et al.	$(2^h - 2)L$	$(2^{h-1})L$	No	Yes	No
Chattopadhyay et al.	$(n)hash + 2nL$	$(n)hash + 2nL$	No	No	No
Shima-Doi	$tn_p L$	$tn_p L + t^3 n_p^3$	Yes	Yes	$l < w(n_p - 1)$
VZD-RSS	$tn(L + tn)$	$t^2 n(L + tn)$	Yes	Yes	$(n - 1)(t - 1)$
Our scheme	$tn(L + \frac{tn}{2})$	$t^2 n(L + \frac{tn}{2})$	Yes	Yes	$\frac{n(t - 1)}{2}$

All rows of upper half of G have distinct differences, then no $p(\neq i)$ has the same difference as $\Delta_{m,r}^i$. Similarly, all rows of lower half of G have distinct differences, then no $p(\neq j)$ has the same difference as $\Delta_{m,r}^j$. Therefore, for any row $p(\neq i, j)$, we have $\Delta_{m,r}^p \neq \Delta_{m,r}^i \text{ or } \Delta_{m,r}^j$.

Example. Consider (4,2,6)-RSS scheme. The generator matrix is:

$$G = \begin{pmatrix} 1 & z & z^2 & z^3 \\ 1 & z^2 & z^4 & z^6 \\ 1 & z^3 & z^6 & z^9 \\ 1 & z^4 & z^8 & z^{12} \\ z & 1 & z^4 & z^2 \\ z^2 & 1 & z^7 & z^4 \\ z^3 & 1 & z^{10} & z^6 \\ z^4 & 1 & z^{13} & z^8 \end{pmatrix} \begin{matrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \\ 8 \end{matrix} \quad (15)$$

Suppose that the shares corresponding to rows 1,4,5,7, are provided. In Figure 2, we show how the zigzag decoding algorithm runs.

As can be seen in (15), $\Delta_{1,4}^1 = \Delta_{1,4}^7$, i.e. the difference between elements of columns 1 and 4 in rows 1 and 7 are equal. According to the proof of Theorem 4, bits of R_1 and K_2 , multiplied by columns 1 and 4, should be decoded by other available shares, i.e. S_4 and S_5 .

Also, $\Delta_{2,3}^4 = \Delta_{2,3}^5$ and bits of R_2 and K_1 should be encoded by shares S_1 and S_7 . In Figure 2, we draw a circle around the exposed bits in the current round.

6. Comparison

In this section, we compare our scheme with some similar schemes proposed by Kurihara et al. [6], Wang-Desmedt [9], Chen et al. [10], Deshmukh et al. [12], Chattopadhyay et al. [20], Shima-Doi [8], and Gong et al. [25] denoted as VZD-RSS and summarize the results in Table 1. The comparison is based on computational efficiency as well as whether there exist some limitations on the values of n and t that the schemes support. The results indicate that our scheme is the only scheme that has improved computational complexity in both sharing and recovery phases, while it has no limitation on n, t . Also, its storage overhead is almost half of the VZD-RSS.

Based on Table 1, in the sharing phase, computational complexity of our scheme is superior to VZD-RSS. In Kurihara et al.'s scheme, number of computations is tn_pL , where $n_p \geq n$ is a prime number. In the best case, $n_p = n$, but there are cases that n_p is much larger than n and therefore our scheme is more efficient than Kurihara et al.'s scheme. Chen et al.'s scheme has low computational complexity but it has a limitation and can be used only for $n = t, t + 1$. Shima-Doi also achieves the same computational complexity as Kurihara et al. Deshmukh et al.'s scheme has efficient computation complexity, but it is limited to the special case of (n, n) where $n = 2^{h-1}$. In Chattopadhyay et al.'s scheme, in addition to XOR operation, it is necessary to calculate the hash function n times which imposes high computational overhead on the system. For large secret, i.e. when L grows

faster than tn , the complexity of ZD-based schemes reduces to $O(tnL)$.

In the recovery phase, computational complexity of the proposed scheme is $t^2(L + tn/2)$ and therefore has higher efficiency than other schemes. For large secret, i.e. with increasing L , complexity of Wang and Desmedt's scheme (nL^3) has the worst efficiency compared to all other schemes. Also, computational complexity of the ZD-based schemes are $O(t^2L)$ and outperform the remaining three schemes with complexity $O(tnL)$.

We now discuss storage overhead of the compared schemes. Kurihara et al., Deshmukh et al., Chattopadhyay et al. and Wang-Desmedt's schemes do not increase the share size during sharing phase, however they may pad some bits to the secret before running this phase. That is, ultimately, their generated shares have some overhead compared to the original secret. In Kurihara et al., the secret is padded if its length is not a multiple of $w(n_p - 1)$. Wang-Desmedt's scheme pads the secret to a string of length n , which it is negligible.

The proposed scheme and VZD-RSS have also storage overhead. In VZD-RSS, the size of each share is $L + (n - 1)(t - 1)$. This means that the size of the overhead is $(n - 1)(t - 1)$. While, the proposed scheme generates the shares with the length of $L + (n)(t - 1)/2$ which reduces the overhead almost by half.

Another advantage of our method is that sometimes more than one bit is exposed in each iteration of the zigzag decoding algorithm. But in the VZD-RSS method, exactly one bit is exposed in each round. This can increase speed of decoding.

7. Conclusion

This paper presents a (t, w, n) -ramp secret sharing scheme based on ZD codes. The secret recovery phase as well as sharing phase are done using only Boolean operations. Storage overhead of the proposed scheme is almost half of the overhead in existing literature. We prove that while the overhead is decreased, the scheme preserves its security. We further prove that the proposed algorithm achieves zigzag decodability, i.e. the recovery phase involves only the shift and XOR operations.

8. Appendix

This section is devoted to the proof of Lemma 1. We prove the lemma in three steps:

1. We show that it is possible to split G into two submatrices whose rows are linearly independent.
2. We show that if we replace any rows of one of these submatrices with a row of the {other submatrix}, then the rows of the resulting matrix are still linearly independent.
3. We show that the claim stated in step II is valid for any number of rows.

Step I. For simplicity, we consider $n/2 = t$. First, we partition G into two non-overlapping submatrices A and B as follows.

$$G = \left(\begin{array}{cccccc} 1 & z & \dots & z^{t-w-1} & \dots & z^{t-1} \\ 1 & z^2 & \dots & z^{2(t-w-1)} & \dots & z^{2(t-1)} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ 1 & z^t & \dots & z^{t(t-w-1)} & \dots & z^{t(t-1)} \\ \hline z^{t-w-1} & z^{t-w-2} & \dots & 1 & \dots & z^{t-w} \\ z^{2(t-w-1)} & z^{2(t-w-2)} & \dots & 1 & \dots & z^{2(t-w)} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ z^{t(t-w-1)} & z^{t(t-w-2)} & \dots & 1 & \dots & z^{t(t-w)} \end{array} \right) = \left(\begin{array}{c} A \\ B \end{array} \right)$$

To prove that the rows of these submatrices are linearly independent, we show that A and B are invertible.

1. A is a Vandermonde matrix which is known to be invertible iff the values of its second column are non-repetitive [29].
2. The submatrix B is a special columnar permuted form of Vandermonde matrix. We know that the columnar elementary operations do not change the rank of a matrix, accordingly, B is invertible [30].

Step II. We first show that if we replace an arbitrary row of A with any row of B , the resulting matrix has t independent rows. To do so, we show that the new row is independent of other remaining rows of A .

To prove this, we make A upper triangular and call it A' as (16).

$$A' = \left(\begin{array}{cccccc} 1 & z & \dots & z^{i-2} & z^{i-1} & \dots & z^{t-1} \\ 0 & z^2 - z & \dots & z^{2(i-2)} - z^{i-2} & z^{2(i-1)} - z^{i-1} & \dots & z^{2(t-1)} - z^{t-1} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 0 & z^{i(i-1)} - z^{i-1} & \dots & z^{i(t-1)} - z^{t-1} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 0 & 0 & \dots & z^{t(t-1)} \end{array} \right) \quad (16)$$

Afterwards, consider the matrix A'' obtained by replacing the i -th row of A' with the j -th row of B as (17).

$$A'' = \left(\begin{array}{cccccc} 1 & z & \dots & z^{i-2} & z^{i-1} & \dots & z^{t-1} \\ 0 & z^2 - z & \dots & z^{2(i-2)} - z^{i-2} & z^{2(i-1)} - z^{i-1} & \dots & z^{2(t-1)} - z^{t-1} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \hline z^{j(t-w-1)} & z^{j(t-w-2)} & \dots & \dots & \dots & \dots & z^{j(t-w)} \\ \hline \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 0 & 0 & \dots & z^{t(t-1)} \end{array} \right) \quad (17)$$

Note that since we choose the coefficients based on the value of $(i-1)$ -th cell of i -th row to make it 0, $A''[i][i]$ will never be zero.

We can convert the first $(i-1)$ values of the i -th row to 0 by the leading coefficient of its previous rows. But we cannot make the i -th value 0 at the same time, since there is no row in which its i -th value is the leading coefficient, and the sum of previous rows cannot make $(i-1)$ -th and i -th value 0, simultaneously, in the light of the fact that there is no repetitive value in none of the G 's columns. So, at least

we will have $z^{j(t-i-1)+1}$ left in the new i -th term.

Step III. Finally, giving an incremental construction algorithm, we explain that we can generalize the result of previous step for any number ($\leq t$) of rows. Instead of replacing all new rows simultaneously, we are able to do that one by one. Using the same argument as step II, after altering one of the rows we have again a submatrix in its previous state, meaning that it has still t linearly independent rows and it is still upper-triangularizable such that all of its diagonal elements are non-zero. So, for the second row, we can continue as in the first one and so on.

This completes the proof, by virtue of the fact that the same happens when replacing a row of B with a row of A .

9. References


- [1] Shamir, A., "How to share a secret", *Commun. ACM*, Vol. 22, No. 11, pp. 612–613, 1979.
- [2] Hineman, A., and Mario, B., "A modified Shamir secret sharing scheme with efficient encoding", *IEEE Communications Letters*, Vol. 26.4, pp. 758–762, 2022.
- [3] Shiina, N., "How to convert 1-out-of- n proof into k -out-of- n proof", *Proc SCIS2004*, pp. 1435–1440, 2004.
- [4] Kurihara, J., Kiyomoto, S., Fukushima, K., and Tanaka, T., "A fast $(3, n)$ -threshold secret sharing scheme using exclusive-or operations", *IEICE Trans. Fundam. Electron. Commun. Comput. Sci.*, Vol. 91, No. 1, pp. 127–138, 2008.
- [5] Kurihara, J., Kiyomoto, S., Fukushima, K., and Tanaka, T., "A new (k, n) -threshold secret sharing scheme and its extension", in *International Conference on Information Security*, pp. 455–470, Springer, 2008.
- [6] Kurihara, J., Kiyomoto, S., Fukushima, K., and Tanaka, T., "A fast (k, L, n) -threshold ramp secret sharing scheme", *IEICE Trans. Fundam. Electron. Commun. Comput. Sci.*, Vol. 92, No. 8, pp. 1808–1821, 2009.
- [7] Beimel, A., and Othman, H., "Evolving ramp secret sharing with a small gap", *EUROCRYPT 2020*, 2020.
- [8] Shima, K., and Doi, H., "New Proof Techniques Using the Properties of Circulant Matrices for XOR-based (k, n) Threshold Secret Sharing Schemes", *J. Inf. Process.*, Vol. 29, pp. 266–274, 2021.
- [9] Wang, Y., and Desmedt, Y., "Efficient secret sharing schemes achieving optimal information rate", in *2014 IEEE Information Theory Workshop (ITW 2014)*, IEEE, pp. 516–520, 2014.
- [10] Chen, L., Laing, T. M., and Martin, K. M., "Efficient, XOR-Based, Ideal (t, n) - threshold Schemes", in *International Conference on Cryptology and Network Security*, pp. 467–483, Springer, 2016.
- [11] Shima, K., and Doi, H., "A hierarchical secret sharing scheme over finite fields of characteristic 2", *J. Inf. Process.*, Vol. 25, pp. 875–883, 2017.
- [12] Deshmukh, M., Maroti, Neeta, N., and Mushtaq, A., "Secret sharing scheme based on binary trees and Boolean operation", *Knowledge and Information Systems*, Vol. 60, No. 3, pp. 1377–1396, 2019.
- [13] Pande, D., Rawat, A. S., Deshmukh, M., and Singh, M., "Single secret sharing scheme using chinese remainder theorem, modified Shamir's scheme and XOR operation", *Wireless Personal Communications*, Vol. 130, No. 2, pp. 957–985, 2023.

- [14] Kabirirad, S., and Eslami, Z., "Improvement of (n, n) -multi-secret image sharing schemes based on Boolean operations", *J. Inf. Secur. Appl.*, Vol. 47, pp. 16–27, 2019.
- [15] Bisht, K., and Deshmukh, M., "A novel approach for multilevel multi-secret image sharing scheme", *J. Supercomput.*, pp. 1–35, 2021.
- [16] Paul, A., Kandar, S., and Dhara, B. C., "Boolean operation based lossless threshold secret image sharing", *Multimedia Tools and Applications*, Vol. 81 No. 24, pp. 35293–35316, 2022.
- [17] Huang, P.-C., Chang, C.-C., Li, Y.-H., and Liu, Y., "Enhanced (n, n) -threshold QR code secret sharing scheme based on error correction mechanism", *J. Inf. Secur. Appl.*, Vol. 58, pp. 102719, 2021.
- [18] Kabirirad, S., and Eslami, Z., "A (t, n) -multi secret image sharing scheme based on Boolean operations", *J. Vis. Commun. Image Represent.*, Vol. 57, pp. 39–47, 2018.
- [19] Nag, A., Singh, J. P., and Singh, A. K., "An efficient Boolean based multi-secret image sharing scheme", *Multimed. Tools Appl.*, pp. 1–25, 2019.
- [20] Chattopadhyay, A. K., Nag, A., Singh, J.P., and Singh, A. K., "A verifiable multi-secret image sharing scheme using XOR operation and hash function", *Multimedia Tools and Applications*, Vol 80, pp. 35051–35080, 2021.
- [21] Sung, C. W., and Gong, X., "A ZigZag-decodable code with the MDS property for distributed storage systems", in *2013 IEEE International Symposium on Information Theory, IEEE*, pp. 341–345, 2013.
- [22] Dai, M., Sung, C. W., Wang, H., Gong, X., and Lu, Z., "A new zigzag-decodable code with efficient repair in wireless distributed storage", *IEEE Trans. Mob. Comput.*, Vol. 16, No. 5, pp. 1218–1230, 2016.
- [23] Hou, H., Lee, P. P., and Han, Y. S., "ZigZag-decodable reconstruction codes with asymptotically optimal repair for all nodes", *IEEE Trans. Commun.*, Vol. 68, No. 10, pp. 5999–6011, 2020.
- [24] Lu, S., Zhang, C., and Dai, M., "CP-BZD Repair Codes Design for Distributed Edge Computing", in *2020 IEEE 26th International Conference on Parallel and Distributed Systems (ICPADS)*, IEEE, pp. 722–727, 2020.
- [25] Gong, X., Hu, P., Shum, K. W., and Sung, C. W., "A zigzag-decodable ramp secret sharing scheme", *IEEE Trans. Inf. Forensics Secur.*, Vol. 13, No. 8, pp. 1906–1916, 2018.
- [26] Gong, X., and Sung, C. W., "Zigzag decodable codes: Linear-time erasure codes with applications to data storage", *J. Comput. Syst. Sci.*, Vol. 89, pp. 190–208, 2017.
- [27] Iwamoto, M., and Yamamoto, H., "Strongly secure ramp secret sharing schemes for general access structures", *Inf. Process. Lett.*, Vol. 97, No. 2, pp. 52–57, 2006.
- [28] Gollakota, S., and Katabi, D., "Zigzag decoding: Combating hidden terminals in wireless networks", in *Proceedings of the ACM SIGCOMM 2008 conference on Data communication*, pp. 159–170, 2008.
- [29] Trappe, W., "Introduction to cryptography with coding theory", Pearson Education India, 2006.
- [30] Hoffman, K., and Kunze, R., "Linear Algebra, Prentice-Hall", Inc Englewood Cliffs N. J., pp. 122–125, 1971.



Analysis of the Impact of Wireless Three-User Multiple Access Channel Coefficients Correlation on Outage Probability: A Copula-Based Approach^{*}

Research Note

Mona Sadat Mohsenzadeh¹, Ghosheh Abed Hodtani² 

DOI: 10.22067/cke.2023.83498.1098

Abstract: In wireless communication channels, the channel coefficients are dependent on each other. In this paper, we investigate the constructive or destructive effects of fading coefficients correlation of wireless multiple access channels (MAC) on the outage probability (OP) (one of the wireless communications performances). Our desired channel is a wireless Rayleigh fading MAC with three-users, independent sources, and interdependent fading coefficients; we use the Copula theory to model the dependence structure. We obtain a closed-form expression for the channel outage probability using an important family of Copulas named the Farlie-Gumbel-Morgenstern (FGM) Copula. The results show that the negative dependence structure of the channel coefficients has a constructive effect on the outage probability; in other words, it reduces the outage probability values compared to the independent case. In contrast, the positive dependency has a destructive effect on the outage probability. The efficiency of the analytical results is illustrated numerically, and the numerical illustrations confirm our theoretical results.

Keywords: multiple access channel, correlated Rayleigh fading, Copula theory, Farlie-Gumbel-Morgenstern, outage probability

1. Introduction

In the last few decades, wireless communication has become an indispensable part of life and the growth of wireless networks continues. Considering the ever-increasing progress of wireless communication systems, as well as from practical and theoretical viewpoints, it is important to analyze the performance of communication systems using the results of information theory. In addition, the coefficients of communication channels in real wireless systems are dependent, and this dependency affects communication performances.

Copulas are known as a useful and powerful tool for modeling the dependence between random variables, hence these functions play an important role in various fields such as finance, insurance, hydrology, image processing statistics,

machine learning, and various engineering applications [1-6]. Recently the use of Copula functions in the field of wireless communication is increasing.

An important family of Copulas is the FGM Copulas, first studied by Eyraud, Farlie, Gumble, and Morgenstern [7-10]. Some features of this family of Copulas have been studied in the literature [11-13]. FGM Copula is one of the fastest Copulas for statistical data processing and the simplest for calculating joint distributions [12]. Due to the simple form of this Copula, and the coverage of positive and negative values of its dependency parameter, the FGM Copula is very suitable for the analysis of wireless channels with dependent coefficients and provides a deep understanding of the effects of correlation on the performance metrics of multi-user communication systems.

The outage probability is an important performance metric of wireless communication, which has been studied in the literature [14-16]. In [14], assuming the correlation of the channel coefficients, a general closed-form expression of the outage probability for any arbitrary fading distribution has been obtained. Also, for the Rayleigh fading channel in two correlated and independent states, closed-form expressions for the outage probability have been obtained. In [15], for the doubly dirty fading MAC with non-causally known side information at transmitters, using the Copula theory, closed-form expressions for the outage probability and the coverage region are obtained. In [16], the authors have investigated the destructive and constructive effects of coefficients dependence of wireless three-user MAC on the outage probability.

Considering the importance of wireless MAC in communication issues, in this paper, these channels have been studied and the outage probability under the influence of the correlation of channel coefficients has been evaluated. For this purpose, first, using the three-dimensional FGM Copula function, a closed-form expression for the outage probability of wireless fading MAC with three transmitters has been obtained. Then, according to this expression, the

^{*} Manuscript received: 2023 July 17, Revised, 2023 August 16, Accepted, 2023 September 30.

¹ PhD Candidate, Department of Electrical Engineering, Ferdowsi University of Mashhad, Mashhad, Iran.

² Corresponding Author. Professor, Department of Electrical Engineering, Ferdowsi University of Mashhad, Mashhad, Iran.

Email: ghodtani@gmail.com

influence of the dependence of the channel coefficients on the outage probability has been evaluated. Considering that the FGM Copula includes both positive and negative dependencies, therefore, by changing the dependency parameter of this Copula within the allowed range, the effect of both types of dependencies on the outage probability can be investigated.

The structure of this paper is as follows: channel model and a brief review of Copula theory are described in sections 2. The outage probability is obtained in Section 3. Numerical results are in section 4 and the paper is concluded in section 5.

2. Channel and Basic Definitions

2.1. Channel

Our desired channel in this paper is a three-transmitter wireless multiple access channel with independent sources and dependent Rayleigh fading coefficients (Figure 1). The received signal is

$$Y = h_{1D}X_1 + h_{2D}X_2 + h_{3D}X_3 + Z \quad (1)$$

X_1 , X_2 , and X_3 are the signals sent by the first, second, and third transmitters, respectively.

Z is independent identically distributed (iid) Additive White Gaussian Noise (AWGN) with zero mean and variance N . h_{iD} ; $i \in \{1,2,3\}$ are the fading coefficients of the communication channel.

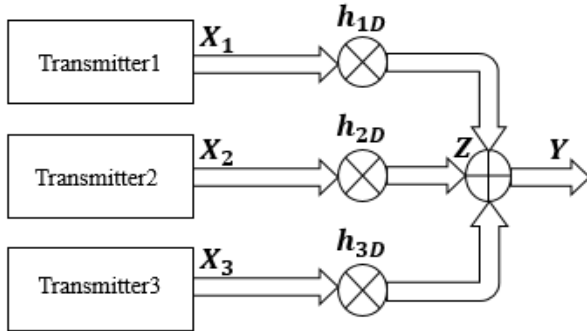


Figure 1. A three-user wireless Rayleigh fading MAC

We can extend the capacity region of two-user MAC with independent sources, determined by Ahlswede and Liao [17,18], to three-user MAC with independent sources.

The capacity region of a three-transmitter wireless MAC with block fading and coherent is equal to:

$$R_1 \leq \frac{1}{2} \log_2 \left(1 + \frac{P_1 |h_{1D}|^2}{N} \right)$$

$$R_2 \leq \frac{1}{2} \log_2 \left(1 + \frac{P_2 |h_{2D}|^2}{N} \right)$$

$$R_3 \leq \frac{1}{2} \log_2 \left(1 + \frac{P_3 |h_{3D}|^2}{N} \right)$$

$$R_1 + R_2 \leq \frac{1}{2} \log_2 \left(1 + \frac{P_1 |h_{1D}|^2 + P_2 |h_{2D}|^2}{N} \right)$$

$$R_1 + R_3 \leq \frac{1}{2} \log_2 \left(1 + \frac{P_1 |h_{1D}|^2 + P_3 |h_{3D}|^2}{N} \right)$$

$$R_2 + R_3 \leq \frac{1}{2} \log_2 \left(1 + \frac{P_2 |h_{2D}|^2 + P_3 |h_{3D}|^2}{N} \right)$$

$$R_1 + R_2 + R_3 \leq \frac{1}{2} \log_2 \left(1 + \frac{P_1 |h_{1D}|^2 + P_2 |h_{2D}|^2 + P_3 |h_{3D}|^2}{N} \right) \quad (2)$$

Where R_1 , R_2 and R_3 are the desired transmission rates of the first, second, and third transmitters, respectively.

2.2. A Brief Review of Copula Theory

What is a Copula? a brief review [1].

Definition 1 A d-dimensional Copula, $C: [0,1]^d \rightarrow [0,1]$ is a function with the following properties:

1. $C(u_1, \dots, u_d) = 0$; if any $u_j = 0, j \in \{1, \dots, d\}$

2. $C(1, \dots, 1, u_j, 1, \dots, 1) = u_j$

3. C is d-increasing on $[0,1]^d$, that is:

$$\sum_{i_1=1}^2 \dots \sum_{i_d=1}^2 (-1)^{i_1+\dots+i_d} C(u_{1i_1}, \dots, u_{di_d}) \geq 0 \quad (3)$$

For all $0 \leq u_{j1} \leq u_{j2} \leq 1$ and $j \in \{1, \dots, d\}$.

Theorem 1 Assuming that F is a d-dimensional cumulative distribution function (CDF) and F_1, \dots, F_d are its margins, then there exists a Copula, C , such that:

$$F(x_1, \dots, x_d) = C(F_1(x_1), \dots, F_d(x_d)) \quad (4)$$

If F_i is continuous for all $i = 1, \dots, d$, then C is unique.

Using Sklar's theorem, the joint probability density function (PDF) corresponding to F is as follows:

$$f(x_1, \dots, x_d) = f_1(x_1) \dots f_d(x_d) c(F_1(x_1), \dots, F_d(x_d)) \quad (5)$$

Where $f_i(x_i)$; $i \in \{1, \dots, d\}$ are the marginal PDFs and c is the Copula density function and is calculated as follows:

$$c(u_1, \dots, u_d) = \frac{\partial^d C(u_1, \dots, u_d)}{\partial u_1 \dots \partial u_d} \quad (6)$$

Definition 2 A d-dimensional FGM Copula is defined as follows [11]:

$$C(u_1, \dots, u_d) = \left(\prod_{j=1}^d u_j \right) \left(1 + \sum_{k=2}^d \sum_{1 \leq j_1 < \dots < j_k \leq d} \theta_{j_1 \dots j_k} \bar{u}_{j_1} \dots \bar{u}_{j_k} \right) \quad (7)$$

Where $(u_1, \dots, u_d) \in [0,1]^d$ and $\bar{u}_j = 1 - u_j, j \in \{1, \dots, d\}$

3. Outage Probability

Outage probability is the probability that the channel capacity is less than a required threshold information rate, R_{th} .

Theorem 2 The outage probability of a three-user wireless Rayleigh correlated fading MAC is:

$$P_{out} = 1 - \left[\frac{\bar{\lambda}_2 e^{-\frac{L}{\bar{\lambda}_2}}}{(\bar{\lambda}_2 - \bar{\lambda}_3)} + \theta_{FGM} \left(\frac{\bar{\lambda}_2 e^{-\frac{L}{\bar{\lambda}_2}} \left(1 + e^{-\frac{L}{\bar{\lambda}_2}} \right)}{(\bar{\lambda}_2 - \bar{\lambda}_3)} - \frac{2\bar{\lambda}_2 e^{-\frac{L}{\bar{\lambda}_2}}}{(2\bar{\lambda}_2 - \bar{\lambda}_3)} - \frac{\bar{\lambda}_2 e^{-\frac{2L}{\bar{\lambda}_2}}}{(\bar{\lambda}_2 - 2\bar{\lambda}_3)} \right) \right] \quad (8)$$

Proof: According to the definition given above for the outage probability, we have:

$$P_{out} = Pr(R_2 + R_3 \leq R_{th}) \quad (9)$$

$$= 1 - Pr(R_2 + R_3 > R_{th}) \quad (10)$$

$$= 1 - P_{out}^c \quad (11)$$

Where R_{th} represents the total required threshold information rates and P_{out}^c is the complement of the outage probability. Now we calculate P_{out}^c .

Any of the inequalities in (2) can be used to calculate P_{out}^c . In this paper, we use the following inequality:

$$R_2 + R_3 \leq \frac{1}{2} \log_2 \left(1 + \frac{P_2|h_{2D}|^2 + P_3|h_{3D}|^2}{N} \right) \quad (12)$$

We have:

$$P_{out}^c = Pr \left(\frac{1}{2} \log_2 \left(1 + \frac{P_2|h_{2D}|^2 + P_3|h_{3D}|^2}{N} \right) > R_{th} \right) \quad (13)$$

$$= Pr \left(\frac{P_2|h_{2D}|^2 + P_3|h_{3D}|^2}{N} > 2^{2R_{th}} - 1 \right) \quad (14)$$

$$= Pr(\lambda_2 + \lambda_3 > L) \quad (15)$$

$$= \int_0^\infty \int_{L-\lambda_3}^\infty f(\lambda_2, \lambda_3) d\lambda_2 d\lambda_3 \quad (16)$$

Where $\lambda_2 = \frac{P_2|h_{2D}|^2}{N}$, $\lambda_3 = \frac{P_3|h_{3D}|^2}{N}$, $L = 2^{2R_{th}} - 1$ and $f(\lambda_2, \lambda_3)$ is the joint PDF of λ_2 and λ_3 .

To calculate the double integral (16), we need to have the

joint PDF of λ_2 and λ_3 . Also, according to (5), to calculate $f(\lambda_2, \lambda_3)$, we must have the marginal PDFs, $f(\lambda_2)$ and $f(\lambda_3)$, and the Copula density function, $c(u_1, u_2)$.

As stated in subsection A, the channel coefficients, $h_{iD}; i \in \{1,2,3\}$, have a Rayleigh distribution and we know that the square of the Rayleigh distribution is an exponential distribution, so $|h_{iD}|^2; i \in \{1,2,3\}$ and consequently $\lambda_i; i \in \{1,2,3\}$ have an exponential distribution:

$$f(\lambda_i) = \frac{1}{\bar{\lambda}_i} \exp \left(-\frac{\lambda_i}{\bar{\lambda}_i} \right); i \in \{1,2,3\} \quad (17)$$

$$F(\lambda_i) = 1 - \exp \left(-\frac{\lambda_i}{\bar{\lambda}_i} \right); i \in \{1,2,3\} \quad (18)$$

Where $f(\lambda_i); i \in \{1,2,3\}$ and $F(\lambda_i); i \in \{1,2,3\}$ are PDFs and CDFs of $\lambda_i; i \in \{1,2,3\}$, respectively. $\bar{\lambda}_i; i \in \{1,2,3\}$ are average SNRs and are given as:

$$\bar{\lambda}_i = \frac{P_i \mathbb{E}[|h_i|^2]}{N}; i \in \{1,2,3\} \quad (19)$$

Considering $d = 2$ in (7), two-dimensional FGM Copula is obtained as:

$$C_{FGM}(u_1, u_2) = u_1 u_2 (1 + \theta_{FGM} \bar{u}_1 \bar{u}_2); \theta_{FGM} \in [-1,1] \quad (20)$$

Where $\bar{u}_1 = 1 - u_1$, $\bar{u}_2 = 1 - u_2$ and θ_{FGM} is the FGM Copula parameter.

Now, according to (6) and (20), the density function of the two-dimensional FGM Copula is:

$$c_{FGM}(u_1, u_2) = 1 + \theta_{FGM}(1 - 2u_1)(1 - 2u_2) \quad (21)$$

According to (5), (17) and (21), $f(\lambda_2, \lambda_3)$ is obtained as follows:

$$f(\lambda_2, \lambda_3) = \frac{e^{-\frac{\lambda_2}{\bar{\lambda}_2} - \frac{\lambda_3}{\bar{\lambda}_3}}}{\bar{\lambda}_2 \bar{\lambda}_3} \left[1 + \theta_{FGM} \left(1 - 2e^{-\frac{\lambda_2}{\bar{\lambda}_2}} \right) \left(1 - 2e^{-\frac{\lambda_3}{\bar{\lambda}_3}} \right) \right] \quad (22)$$

By putting (22) in (16), it is easy to calculate the double integral and the outage probability is obtained as (8) and the proof is complete.

4. Numerical Results

In this section, we investigate the effect of the correlation of channel coefficients on the outage probability.

Figure 2 and Fig. 3 show the outage probability versus the average SNR. As can be seen, as average SNR increases, the outage probability decreases, because the channel conditions improve.

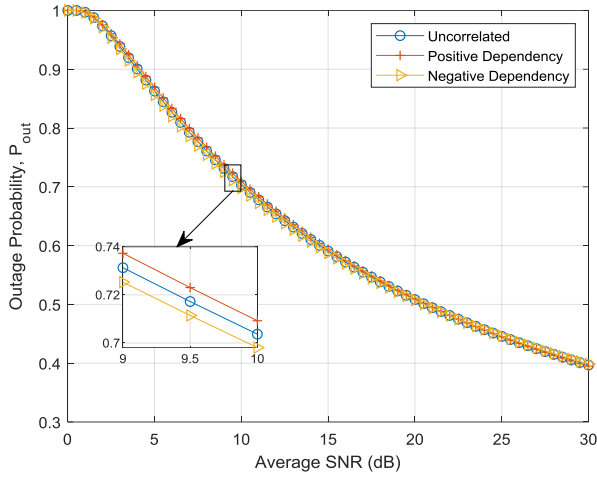


Figure 2. Outage probability versus average SNR, $\theta_{FGM} = 1/8$ (positive dependency), $\theta_{FGM} = 0$ (uncorrelated), and $\theta_{FGM} = -1/8$ (negative dependency)

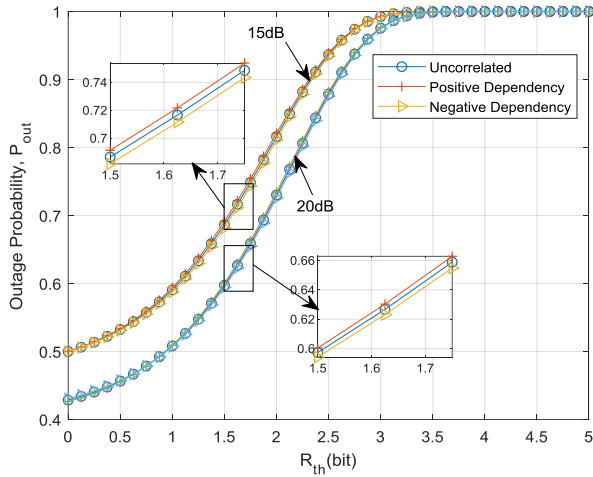


Figure 3. Outage probability versus R_{th} , $\theta_{FGM} = 1/8$ (positive dependency), $\theta_{FGM} = 0$ (uncorrelated), and $\theta_{FGM} = -1/8$ (negative dependency)

Also, we can see that the negative dependence structure ($\theta_{FGM} \in [-1, 0)$) improves the outage probability performance compared to the uncorrelated state. In contrast, the positive dependence structure has a destructive effect on the outage probability performance.

Figure 3 and Figure 5 show the behavior of outage probability in terms of the total required threshold information. It can be seen that as R_{th} increases, the outage probability also increases and finally tends to one, and this means that it is impossible to transfer information. According to these figures, it is clear that the outage probability under the influence of the negative dependence structure has lower values compared to the uncorrelated state.

The curves in Figure 3 are plotted for two different values of average SNR, and according to these two groups of curves, we find that for a fixed value of the total required threshold information rate, R_{th} , as the average SNR increases, the outage probability improves.

By comparing Figure 2 with Figure 4 and Figure 3 with Figure 5, we find that as the positive dependence increases,

its destructive effect on the outage probability increases.

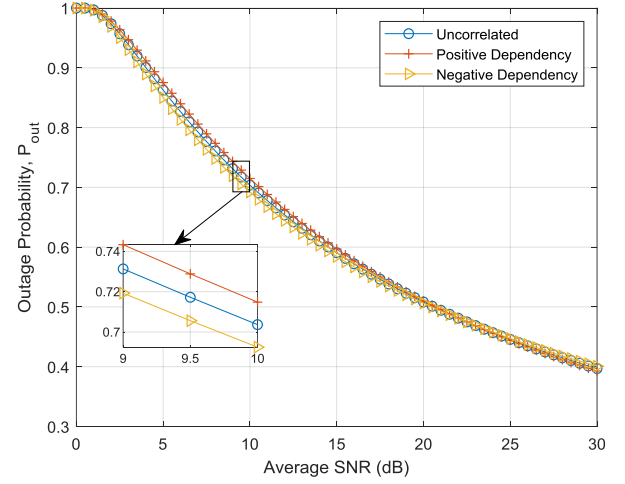


Figure 4. Outage probability versus average SNR, $\theta_{FGM} = 1/4$ (positive dependency), $\theta_{FGM} = 0$ (uncorrelated), and $\theta_{FGM} = -1/4$ (negative dependency)

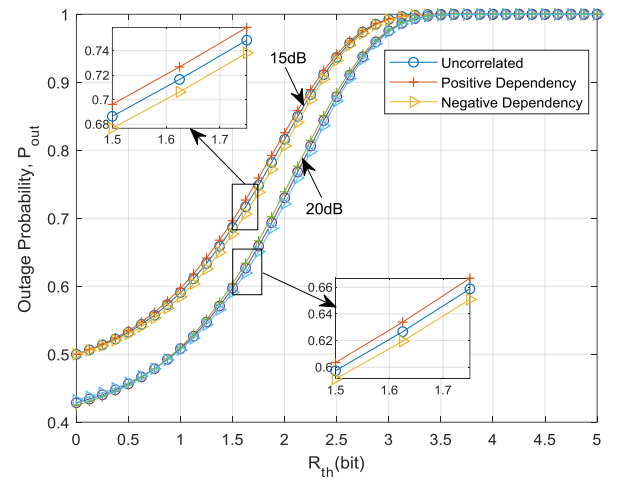


Figure 5. Outage probability versus R_{th} , $\theta_{FGM} = 1/4$ (positive dependency), $\theta_{FGM} = 0$ (uncorrelated), and $\theta_{FGM} = -1/4$ (negative dependency)

Similarly, as the negative dependence increases, its constructive effect on the outage probability increases.

5. Conclusion

In this paper, wireless three-user MAC with independent sources and Rayleigh fading was investigated. Using the FGM Copula, a closed form expression for the outage probability was obtained. Then we analyzed the impact of positive and negative dependencies on the outage probability performance. According to the obtained results, it is clear that negative dependence, compared to the independent state, reduces the outage probability, while positive dependency increases the outage probability compared to the non-dependent case.

6. References

- [1] Nelsen, R. B., "An introduction to copulas", Springer, 2006.

- [2] Lee, R., Kim, G., Hur, J., and Shin, H., "Advanced Probabilistic Power Flow Method Using Vine Copulas for Wind Power Capacity Expansion," *IEEE Access*, vol. 10, pp. 114929-114941, 2022.
- [3] Pambudi, A. D., Ahmad, F., and Zoubir, A. M., "Robust Copula-Based Detection of Shallow-Buried Landmines With Forward-Looking Radar", *IEEE Transactions on Aerospace and Electronic Systems*, vol. 58, no. 2, pp. 1187-1198, 2022.
- [4] Trigui, I., Shahbaztabar, D., Ajib, W., and Zhu, W. P., "Copula-Based Modeling of RIS-Assisted Communications: Outage Probability Analysis", *IEEE Communications Letters*, vol. 26, no. 7, pp. 1524-1528, 2022.
- [5] Wang, T., Yang, X., Ren, X., Yu, W., and Yang, S., "Locally Private High-Dimensional Crowdsourced Data Release Based on Copula Functions," *IEEE Transactions on Services Computing*, vol. 15, no. 2, pp. 778-792, 2022.
- [6] Zhou, N., Xu, X., Yan, Z., and Shahidehpour, M., "Spatio-Temporal Probabilistic Forecasting of Photovoltaic Power Based on Monotone Broad Learning System and Copula Theory", *IEEE Transactions on Sustainable Energy*, vol. 13, no. 4, pp. 1874-1885, 2022.
- [7] Eyraud, H., "Les principes de la mesure des correlations", *Ann. Univ. Lyon, III. Ser., Sect. A*, vol. 1, no. 30-47, pp. 111, 1936.
- [8] Farlie, D. J., "The performance of some correlation coefficients for a general bivariate distribution", *Biometrika*, vol. 47, no. 3/4, pp. 307-323, 1960.
- [9] Gumbel, E. J., "Bivariate exponential distributions," *Journal of the American Statistical Association*, vol. 55, no. 292, pp. 698-707, 1960.
- [10] Morgenstern, D., "Einfache beispiele zweidimensionaler verteilungen", *Mitt. Math. Statist.*, vol. 8, pp. 234-235, 1956.
- [11] Bekrizadeh, H., "Generalized FGM copulas: Properties and applications", *Communications in Statistics - Simulation and Computation*, pp. 1-12, 2022.
- [12] Sriboonchitta, S., and Kreinovich, V., "Why Are FGM Copulas Successful? A Simple Explanation", *Advances in Fuzzy Systems*, vol. 2018, pp. 5872195, 2018/05/14 2018.
- [13] Stoica, E., "A stability property of farlie-gumbel-morgenstern distributions", 2013.
- [14] Rostamighadi, F., and Abed Hodtani, G., "Copula-based Analysis of Interference Channels: Outage Probability", *Iran Workshop on Communication and Information theory*, 2022.
- [15] Ghadi, F. R., Hodtani, G. A., and López-Martínez, F. J., "The role of correlation in the doubly dirty fading mac with side information at the transmitters", *IEEE Wireless Communications Letters*, vol. 10, no. 9, pp. 2070-2074, 2021.
- [16] Mohsenzadeh, M. S., Hodtani, G. A., "FGM Copula based Analysis of Outage Probability for Wireless Three-User Multiple Access Channel with Correlated Channel Coefficients", *31st Iranian Conference on Electrical Engineering (ICEE)*, 2023.
- [17] Ahlswede, R., "Multi-way communication channels", 1973.
- [18] Liao, H. "Multiple Access Channels," Ph.D. dissertation, Department of Electrical Engineering, University of Hawaii, Honolulu, 1972.



Ferdowsi
University of
Mashhad

Journal of Computer and Knowledge Engineering

<https://cke.um.ac.ir>



Information and
Communication Technology
Association of Iran

Optimization of FlexiTP Energy-Aware Algorithm in Wireless Sensor Networks *

Research Article

Hamid Mirvaziri¹

DOI: [10.22067/cke.2023.82926.1082](https://doi.org/10.22067/cke.2023.82926.1082)

Abstract: Maximizing WSN longevity besides maintaining their efficiency and proper performance, is one of the most important challenges that researchers of this field encounter. FlexiTP, is a protocol that was designed and made for optimizing energy consumption and maximizing longevity of these kinds of networks. This study presents improved version of FlexiTP protocol using Harmony Search algorithm with the objective of optimizing energy consumption of FlexiTP protocol. The suggested method, HS-FlexiTP, is able to choose the best parent for each sensor node, using Harmony Search algorithm, based on three criteria including; distance of parent node from child node, number of the hops of the parent node and remaining energy of the parent node. Obtained results of the simulations indicate that HS-FlexiTP is able to decrease 25 percent of the consumed energy per each node throughout the various scenarios of simulation in comparison with FlexiTP protocol. In comparison with ZMAC protocol, it has much better efficiency in decreasing consumed energy as well and in comparison with both protocols, ZMAC and FlexiTP, the suggested procedure is able to maximize network longevity and optimize other efficiency criteria including average packet delay, throughput and productivity of the channel.

Keywords: FlexiTP, Harmony Search algorithm, Energy Optimization, Network Longevity, WSNs.

1. Introduction

In the recent years, wireless sensor networks have been able to bring efficiency and speed in data aggregation, wonderful analysis and decision making for people, while provides required security in unsafe environments and control or analyze various issues) for them. In wireless sensor networks, several wireless sensor nodes are used in a way that each of them, considering the mission they have, is equipped with several sensors and use to collect data. Each node sends data to the sink node after data collection and then, through various communication systems send them into the communication center to make required and urgent decision. Because of that, wireless sensor networks have numerous applications including military operations, rescue and assistance, industrial activities, controlling railways, agricultural activities and etc.

In wireless sensor networks each node has a limited computing and energy saving ability bearing in mind that it is often employed in the environment with no

communication or power supply and its tiny dimension. Also, a node can remain at the network until its power supply has not been finished and when the energy of that node finishes, its undercover region will encounter failure which may have drastic consequences. One of the protocols that has been designed and implemented for optimizing energy consumption and maximizing longevity of wireless sensor networks is FlexiTP protocol, whose features and attributes has been examined in the third section. Meta-heuristic Harmony Search algorithm, whose feature has been fully explained in the fourth section of this article, is an algorithm derived from music orchestra. In this algorithm the best harmony would be chosen and played and by repeating and testing various sounds performed by artists. The objective of this article is optimizing consumed energy of wireless sensor networks through redesigning FlexiTP using Harmony Search algorithm. It's expected that this reconstruction would be able to decrease energy consumption significantly in the FlexiTP and increase longevity of the network while provides performance, reliability and scalability of the network. This article organized as follows: Previous works and researches are presented in the second section. Third section is dedicated to introducing FlexiTP protocol; the fourth section has covered Harmony Search algorithm and implementation of improved protocol is offered in the fifth section. Obtained results and comparison with other protocols are presented in the sixth section and finally seventh part is about conclusion and suggestions for future investigations.

2. Review of related works

In order to optimize energy consumption in sensor networks, many solutions and protocols are presented in MAC layer that is divided into three categories of contention based MAC protocol, contention free MAC protocol and synthetic protocol.

2.1. Contention based MAC protocols

Protocols such as [1], [2], [3] and [4] are among contention based MAC protocols that have used multiple access mechanism with (CSMA) (Carrier Sensing Multiple Access) collision detection in their structure. In SMAC (sensor MAC) [1] and T-MAC (Timeout MAC) [2] RTS/CTS (request to send/clear to send) handshaking mechanisms is used as well

* Manuscript received: 2023 March 13, Revised, 2023 July 9, Accepted, 2023 September 20.

¹Computer Department, Faculty of Engineering, Shahid Bahonar University, Kerman, Iran, Email: hmirvaziri@uk.ac.ir

as CSMA mechanism to prevent contention in multi-hop communications. SMAC has alternative waking and sleeping periods with fixed length. In this protocol, nodes go to periodic sleep. This feature causes decreasing of energy consumption but on the other side, this behavior increases delay as the transmitter node has to wait for the listening period of the receiver node when the receiver node wakes up. In order to overcome the problem of using stable duty cycle in SMAC, T-MAC was suggested with dynamic duty cycle. In this protocol the period of nodes' activity and sleeping is determined based on the network traffic and it overcome former protocol problem significantly. Other protocols such as D-MAC [4], B-MAC [5] and WiseMAC [6] are contention based protocols that try to decrease SMAC duty cycle in some way respectively by offering tree structure of data aggregation and utilizing features of the channel and header sampling simultaneously. IEEE 802.15.4 [7] is a popular standard for the medium access control (MAC) scheme of low-powered, low data rate wireless sensors. Further IEEE 802.15.4 has a MAC protocol in MAC layer which defines the slotted medium access control scheme for the short-range communication devices with low data rates. MAC mechanism in IEEE 802.15.4 extends the number of Guaranteed Time Slot (GTS) and the single channel operation of IEEE 802.15.4 standard multichannel operation for data transmission. Performance analysis of the slotted CSMA/CA scheme in the IEEE 802.15.4 standard is discussed in [8]. FROG-MAC [9] as a new algorithms, introduced a new fragmentation strategy for heterogeneous traffic in WSNs that enables high-priority data packets to interrupt any unimportant ongoing transmissions. Main problems of these protocols are contention in multi hop communications and lack of guarantee to deliver packets in these types of communications which causes increasing in energy consumption ratio and decreasing in maximum of network longevity which degrade their performance.

2.2. Contention free MAC protocols

Time division multiple access (TDMA), is one of the contention free protocols in which various users can share channel bandwidth through dividing signal into various time slots. Problem of listening in leisure time and contention is solved by assigning a time frame to each node in this protocol. However, the most important problem that these protocols encounter is the way of specifying length and measurement of each frame or time slot which could influence energy consumption of the network to a great extent [10].

[13], [14] and [15] are synthetic protocols with combination of TDMA/CSMA communicational models. It was tried to use the advantages of those models and remove their defects. In [15], reduced frame protocol TDMA, has been suggested which is a synthetic model of TDMA/CSMA and each time period has been spread with a short time period of competitive mechanisms based on CSMA. In comparison with traditional TDMA that solve the problem of contention by a separate piece of time period, a potential part of these contentions was limited and remains which was resolved using CSMA protocol in a dynamic way. Two strategies are introduced; first one includes the length adjustment of TDMA frame and second one is related to decreasing the number of conflicts over time periods using an Exploration

Concession Period algorithm.

ZMAC is a MAC synthetic protocol in which the power of two CSMA and TDMA protocols are combined [14]. The most important feature of this protocol is its performance under low contention situation (similar to CSMA) and it has a performance under high node density (similar to TDMA). ZMAC has two basic components. First is called neighbor detection and slot assignment, and second called local framing and synchronization. Also this protocol is resistant against dynamic changes of topology, breakage in time synchronization and breakage in assigning time slot to nodes. In the worst case it has a performance similar to CSMA. In spite of abilities that were discussed above, it is not able to have a proper performance in high node density and its performance is degraded down to the performances of contention based protocols. Also, high contention in this protocol leads increasing of node's energy, decreasing of network longevity, increasing degree of packet delay and efficiency reduction of the whole network. Consequently this protocol could not be a proper option in network with high node density. Both ZMAC and IEEE 802.15.4 are hybrid protocols which can save higher energy and supply better scalability and flexibility in comparison to their ancestors.

In the primary stage of FlexiTP protocol, each node performs data aggregation related to their neighbor in order to construct data aggregation tree and determining parent, child, and ancestors using CSMA mechanism. In the next stage a time slot is assigned to each node similar to TDMA mechanism by which it would be able to activate with its parent or child node to send/receive data [13]. In this protocol there are not only maintaining performance criteria but also ability of increasing network longevity by taking proper advantage of features related to CSMA/TDMA. Also, this protocol is able to guarantee packet delivery while removing contention in the channel. It has also a fairness mechanism to ensure fair access of channel among nodes. In order to maximize the longevity of the network in [4], [9], [11], [12] and [14] and reducing energy consumption, nodes timing techniques in [9], clustering technique in [4] [12] [14] and distributed troubleshooting technique in [11] was employed. In [13], Harmony Search algorithm based on clustering was used in order to reduce energy consumption in wireless sensor networks. IEEE 802.15.4 is a low-rate Wireless Personal Area Networks (WPAN). To maintain the synchronization of time-frames it takes coordinator which is operating in the beaconed mode. It has concept of super-frame structure uses TDMA-based period for access so even for energy conservation there is no special design method except a typical duty cycle controlling scheme [16].

There are some major design challenges in wireless sensor networks due to lack of energy and many researchers are looking for those algorithms that can optimize the use of this energy [17]. Harmony Search algorithm is used to optimize energy consumption in many applications. For example a mutated harmony search algorithm (MHSA) is used to improve the energy efficiency of an optimized routing protocol [18]. HS algorithm is also employed for charge scheduling of an energy storage system [19]. In another work a comprehensive review on the applications of HS method is inspected [20]. They evaluate researches in energy systems by using HS will be analysed. In the next sections we discuss only about contention free protocols for non-guaranteed

access and then compare proposed method with them.

3. FlexiTP protocol

The main performance of FlexiTP protocol includes management in generating the routs (forming data aggregation tree), generation of nodes' timing (assigning time period), synchronization in order to reduce clock drift and local fixing and reconstructing for the network in the time of error occurrence (bearing the error). This protocol contains two main phases: First one includes primary initiation of the network that is performed thoroughly by CSMA/CA protocol and second one is data aggregation cycles that is performed by TDMA protocol. Data aggregation structure is located in the primary phase and performs in three levels which include: choosing parents, improving network communication by aggregation of the neighborhood data which eventually leads into formation of clusters and identification of components and network topology. As it was already mentioned, this stage is performed by CSMA/CA mechanism. The following stage (assigning time periods) is performed by the same protocol as well and nodes are allocated in various time periods based on their structure that eventually leads into formation of required routs for data transport. Because of using TDMA in the second stage, time synchronization is very considerable as nodes are required to wake up in a specific time and transport their data. In FlexiTP hierarchy structure is used in formation of data aggregation tree and by performing this synchronization the clock difference is minimized at each cycle. In this protocol, nodes listen to environmental activities to resolve the errors when they occur, without previous notification. This happens at disconnection time of the link or network topology alteration. The architecture of protocol is static in WSNs here. Nodes may suffer from temporary or permanent errors and new nodes may be added to the network at any moment or a node may exit from the network because of a reason such as termination of the energy or movement, but despite of this situation, nodes are placed on their previous position. In this protocol, inner group consulting mechanism is used to categorize all types of defects [5].

4. Harmony Search algorithm

Harmony search (HS) algorithm in this context uses the best chosen process by repeating numerous processes. Each process is tested and would be considered as a solution (solution vector) if it has the required harmony. In order to achieve a better and more beautiful harmony, this work is repeated and in case of success, previous work is replaced with new one. Obtained results will be compared with the fitness function to achieve a better solution in HS algorithm. In order to achieve the best solutions, Harmony Memory (HM) is used. This memory is implemented in the form of a matrix and each row is indicative of a solution in it. Number of columns in that matrix, indicates the number of possible solutions and last column is used for computing and saving degree of fitness function of each row. Number of rows in the matrix is called Harmony Memory Size (HMS) and measurement of the matrix will be equal to $HMS \times (N+1)$. In HS algorithm choosing value of variables is through using one of these three procedures including employment of existing variables in the memory (considering Harmony

Memory or HMC), generating minor alteration in the value of the variables and applying them (Pitch Adjustment or PA) and generating a random value for variable (randomization). Every note is fixed in music; thus, values of variables in the whole memory of the harmony would be alike as well. Probability of choosing a variable from the Harmony Memory is between 0 and 1 and known as Harmony Memory Considering Rate (HMCR). In case that this value is close to 1, desirable solutions would not be gained and if it is close to 0, convergence would be slow down as it should try numerous solutions. Consequently, the value of this probability can be a number between 0.7 and 0.9. This procedure guarantee solutions close to the optimized one, are always kept in the Harmony Memory and they could be more optimized by adjusting the pitch or applying minor alterations in the future usages. It is possible to apply some changes in HS variables after choosing existing variables in the memory. The objective of these minor changes would be optimizing of local solutions. Value alteration of PA occurs with the probability of Pitch Adjustment Rate (PAR) and it has a value between 0 and 1. This value should be chosen carefully i.e. algorithm have to try more local solutions in low values which would slow down the whole complex. For high values, several solutions remain unexamined so the solution is mostly random. This value is often chosen from 0.1 to 0.5. Randomization in HS Algorithm means choosing random values for variables and causes more variation in the solutions. The difference between pitch setting procedures and randomization is that PA tries to find the optimized local solutions by limiting the search area while randomization takes the algorithm performance toward overall searches and finding overall optimized solutions by expanding the search area.

5. Proposed Method

In FlexiTP algorithm, generating of data aggregation tree is performed only once and that would be in primary initialization phase of the network. Also, considering the fact that this protocol is utilized assigning time slots for sleeping and waking periods of the nodes, and each node knows its parents, ancestor and children and their sleeping and waking periods are adjusted together, flat routing algorithm is practically used. In the other word if a node is a few hops far from the sink node and attempt to transmit data to it, this data is delivered to the destination node through the same time slots and in forwarding data mode by its parent and ancestor. On the other hand, the main usage of the HS algorithm is in multi hop routings among wireless sensor nodes; thus HS algorithm can only be used to generate optimal data aggregation tree at the beginning time when this protocol starts. Therefore the best middle nodes (ancestor and parent/parents) are chosen for the node which is a few steps far from the sink node. When the structure of the data aggregation tree is shaped by this algorithm, the protocol enters its next phase which is assigning time slot into the nodes. As a result, it would be assured that the best structure of the data aggregation tree is dedicated to the protocol to continue. In order to achieve this, the suggested method is implemented in three stages and accuracy in the data aggregation is evaluated at the evaluation stage.

In the first stage all of the existing nodes in the network identify the nodes which are located in their direct radio

range by using mass effect and propagation of its data (the remaining energy, positional data and the number of the pitches in relation to the sink node) and save them in their neighbors' table.

In the second stage all nodes know their neighbors, along with the number of their hops to the sink node. They also have gained their distance to the neighbor nodes. In this case they proceed to fill the Harmony Memory and then choose the best node as their parent based on determined fitness function.

In the third stage the structure of the data aggregation tree is generated and dedicated to the second phase of the protocol which is assigning the time slots into the nodes. In this situation, the nodes receive data of time slots in a state that they have chosen the best parent in terms of distance, energy consumption and the number of hops to sink node according to the fitness function defined for them. In the following details of each stage have been explained.

6.1. Identifying the neighbor nodes

In the beginning, the sink node produces a packet contains: identification of a node, its remaining energy and distance to the sink node and then propagates it in the network pervasively.

Those nodes that are located in the direct communicational zone of the sink node receive this message and save it in their direct neighbors table.

After receiving this message, the neighbor nodes specify their status and distance from the sink node. This is happened by adding one hop to the received distance from transmitter node and then responding to the node that has sent the message in order to make the transmitter node able to identify its neighbor nodes and register them in its neighbors table.

It's worth mentioning that determining status and distance to the sink node is only performed once so a node evaluate its status at first by the time of receiving the message and then in case the node's status is initialized, it uses that value and otherwise, adds one hop to the field of received distance from transmitted node and propagate it as its status.

The direct and first surface nodes of the sink node propagate their data pervasively similar to the previous stage. The neighbor nodes of transmitter node receive transmitted data and save it in their neighbors' node table.

The receiver node first looks at its neighbors' table during registration of transmitted node's data. In case that data of this node is in the table, it updates them and otherwise, it registers new data in its neighbors table. Algorithm 1 illustrates Pseudo-code of the above stages. When neighbor's aggregation data phase is finished, phase of generating data aggregation tree structure is started by HS algorithm.

Step1. aggregating neighbors information nodes

```

Start
Initial My Status packet (my id ,remaining energy ,hop to sink ,
distance to base station , my position )
Generate My Status packet
Broadcast My Status packet
End

```

a. Base Station

```

If (received NewPacket )
{
Read NewPacket (Node ID ,Remain Energy ,hop to sink, D2S
and Position of sender node )
If (node id of NewPacket Not exist in my Neighbor Tables
)
{
Add sender node to my Neighbor Tables
If (My Status not initialized )
Initial My Status packet (my id , Remain
Energy , hop to sink of sender node +1 , D2S and my Position)
Broadcast My Status packet
}
Else
Ignore received packet
}

```

b. Receiver nodes

Algorithm 1. Neighbor's Data aggregation to a. basic station b. transmitter groups

6.2. Generating of data aggregation tree by Harmony Search algorithm

In this stage, each node has the data related to its all neighbors. In order to implement the HS algorithm, Harmony Memory matrix is used with size of HMS * (n+1) in which HMS (Harmony Memory Size) indicates the number of rows and solutions and n indicates maximum of the route length between source and destination and could be equal to all of the existing nodes in the network in its maximized status. n is chosen to be 2 in this sample because the objective of choosing the best parent from the neighbors of a node is by using HS algorithm. For computing the fitness of each row, fitness function is used. In order to implement HS algorithm, it is required to fill the memory of the harmony randomly based on the condition that orbit have not been shaped in each row or solution. If there are only two nodes in the solution then the orbit will never been shaped. Also, HMS would be equal to the number of the neighbors of a node in its maximum state. This value is obtained even less after optimization by the algorithm. In this article HSA has been used for choosing the best parent for every existing node in the network. In this state, each node has access to data related to its neighbors so the number of rows in the Harmony Memory is filled randomly and the value of each obtained solution is computed based on the following fitness function in equation (1).

$$fitness\ function = \frac{ER}{D} / h \quad (1)$$

Here, *ER* indicates remaining energy of the node which is considered as the parent node. *D* is indicator of the distance of two nodes from each other and *h* is indicator of the number of hops between the candidates of parent with the sink node. The main objective of choosing this fitness function is because of the following properties: 1-candidate node would have the least distance with the chosen parent 2-parent node would have a proper remaining energy and 3-candidate node would have the minimum number of hops with the sink node for becoming a parent. These three features are able to optimize the consumed energy to a great extent. In this fitness function not only proper distribution of the consumed energy but also the distance between two neighbor nodes are considered. The first one increases longevity of the network and the less distance between two nodes; the less energy is consumed to reinforce the signal. In addition to the

mentioned points, distance of neighbor node has been considered in terms of the number of hops to the sink node which can save energy by eliminating additional and futile forwards by the middle nodes so increase network longevity. Consequently, the bigger the degrees of the fitness function; the more valuable it gets. In order to implement Harmony Memory, each node is filled randomly and based on its neighbor table so the maximum number of rows in the Harmony Memory matrix, can be equal to the number of the neighbors of that node which was demonstrated before by HMS. This matrix would have three columns. The first column contains the source node, the second column is one of the neighbors of the node which has been chosen randomly and the third one is computed degree of the fitness function for each solution.

A. Harmony Search algorithm Optimization

Rows of Harmony Search matrix are filled randomly and each of them is considered as a solution or a vector and fitness function should be computed for each one so if these solutions are not being optimized, the algorithm will work slowly as it have to examine solutions that are not optimized and therefore, in order to optimize functionality of the algorithm, neighbors nodes and the node that is in the state of choosing parent and have equal hops (distance from the sink node) are eliminated from the solution set.

As a result, only those nodes will remain in the Harmony Memory that is closer to the sink node in comparison to the node itself, this could eliminate the solutions which are not optimized and allocate a faster convergence to the algorithm. After the end of this period, each of the nodes has gained the best position in the structure of data aggregation tree, so, in the third stage of the time slots assignment mechanism, FlexiTP protocol is used to determine the activation and sleeping time of nodes. Pseudo-code in Algorithm 2 demonstrates this stage.

Step 2. data aggregation tree using HSA per nodes

```

Parent=-1
HMS=Length of Neighbor Tables
N=2
HS[HMS][N+1]
For (i=0; i<HMS;i++)
{
  If (my hop to sink > hop to sink Neighbor Tables[i]
  Compute fitness function (parent candidate)
  If fitness function(parent candidate > current parent)
    Parent = node id Neighbor Tables[i]
}
Set best parent id

```

Algorithm 2. Optimizing the performance of the HS algorithm and choosing the best parent for each node

6.3. Implementation of the proposed algorithm

In order to implement the proposed algorithm some changes are done in the base code of FlexiTP and HS algorithm has been added to that code in NS2 simulator. The end results are based on the results of 20 various network topologies each them has owned about 100 nodes and are distributed in 300 square meters area.

In the performed simulations the status of the base station

has been fixed on the central spot of the map and the parameters of the simulation have been based on Mica2Mote hardware. Table 1 indicates the parameters of simulation used in the simulation. Also, the gap length of the data aggregation FlexiTP and MFS has been chosen 26 milliseconds. 23.3 milliseconds of that time is for transporting the packet and 2.45 milliseconds for the status of the radio of the node and transition from sleeping to idle state and 0.25 milliseconds is for the transition of the nodes from the idle to sleeping state [5]. The radio range of the existing nodes in the network is 60 meters and the used traffic is CBR (Constant Bit Rate) with the packet production rate of 5 in every second. Also, the whole time span of simulation was 1000 seconds. Choosing this duration make it possible for required outputs to perform optimization and functionality comparisons between proposed and other protocols in a reasonable time.

Table 1. Simulation parameters of FlexiTP in NS2 [5]

Default value	Simulation parameters
19.2 Kbps	bandwidth
56 byte (36 byte for payload and 20 byte for header)	Packet size
60 meter	Communicational zone (radio range of sensor nodes)
63 microwatt	Power consumption in transportation status
30 microwatt	Power consumption in receiving status
30 microwatt	Power consumption in idle status
0.003 microwatt	Power consumption in sleeping state
30 microwatt	Power consumption for transition of sleeping to idle state and vice versa
2.45 microsecond	Transition point
27 microseconds	The size of the FlexiTP gap
100 microseconds	Period of FlexiTP FTS
54000 joule	Primary energy of the node
1000 seconds	Simulation time

7. Obtained results and Evaluation

In this section we discuss about the results obtained from simulation in NS2. We also evaluate proposed algorithm with other protocols and compare them by known criterion.

7.1. Installation of time span and initialization of the network

Figure 1 indicates performance comparison of HS-FlexiTP and FlexiTP protocols in the installation time span and initialization of the network. As it is indicated in this figure, the suggested method is able to have a better performance in the installation time span and initialization in comparison with FlexiTP protocol and improve this time up to 5 percent.

In any case in the proposed algorithm, neighbor's nodes data are saved in the table of their neighbors in the beginning time of data aggregation of the neighbor's period and computing features such as the distance of the neighbor node from the main node and computing the degree of their fitness

function will be occurred in the node itself. This process decreases network traffic and the number of the exchanging packets between existing nodes in the network so this period will be finished in a less time span.

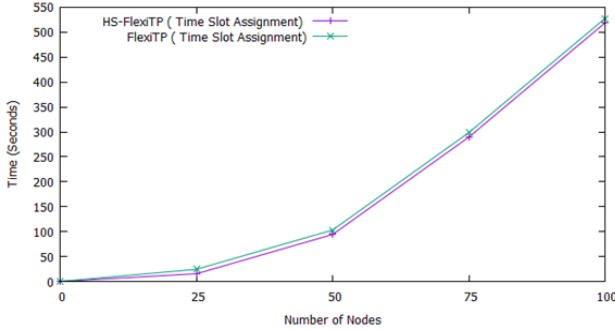


Figure 1. Network installation and initialization time span in FlexiTP and HS-FlexiTP protocols

7.2. Assigning time slot to the nodes

After installation and initialization of the network, the nodes which have been able to receive the neighbors data aggregation messages from the other nodes at the first stage, are positioned in the tree structure and specific time slots is assigned to them in order to enable them to transmit the aggregated data to the parent nodes and eventually to the base station based on these gaps. Figure 2, indicates a comparison between HS-FlexiTP and FlexiTP protocol in assigning time slot to the nodes.

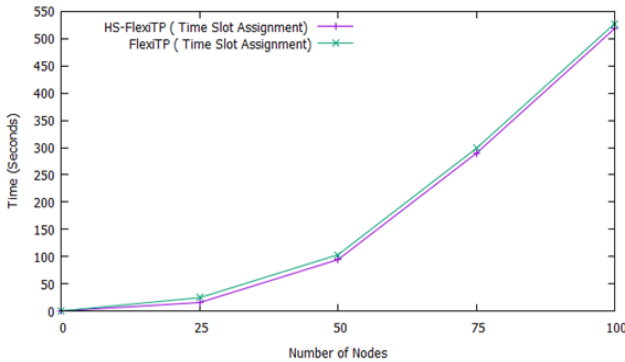


Figure 2. Assigning time slots to the nodes in FlexiTP and HS-FlexiTP protocols

It is obvious in Figure 2 that suggested method is able to reduce the time span of time slots assigned to the existing nodes in the network up to 3.2 percent. Reduction of this time in the two mentioned phases leads to reduce energy consumption for each node and eventually leads to optimize consumed energy and increase network longevity while maintain its efficiency. One of the reasons that the suggested procedure has been able to reduce the duration of time slot assignment, is choosing appropriate parents for each node through computing fitness function of the neighbor nodes and choosing the best node as the parent node of each node.

Of course, optimization of the suggested method in the installation and initialization phase of the network is performed in order to eliminate the neighbor nodes, whose number of hops to the sink node is more than the node which

is in the states of choosing the parent itself. This fact leads to choose the node as the parent node that firstly, has the least hops to the sink node and secondly, the number of its hops is less than the node itself and thirdly, the distance between the parent node and the current node has the minimum value. These points cause to reduce the distance that the packets move between the nodes, and as a result, total energy consume less than before.

7.3. Energy consumption Criteria

In order to evaluate the performance and efficiency of the suggested method in energy consumption and comparing it with FlexiTP and ZMAC protocols, three criteria has been considered: average degree of consumed energy for each node, average degree of consumed energy for each packet and maximum network longevity [5]. As the suggested method is able to reduce the required time in the installation hops and initialization of the network and time slot assignment in comparison with FlexiTP, it is expected that the suggested method has a better performance in the degree of energy consumption. FlexiTP protocol has the capability of using or not using time slots assigned to the nodes by the other nodes. Both features have been considered in the simulations. Figures 3 and 4 show the average consumed energy in activation and deactivation states of capability of reusing time slots for the proposed method, FlexiTP and ZMAC respectively. It is worth mentioning that this capability is not defined in ZMAC.

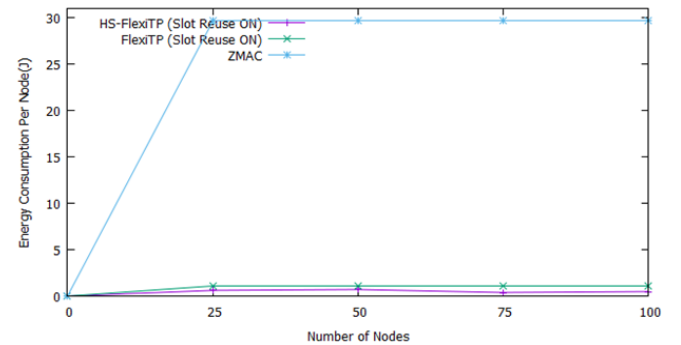


Figure 3. The degree of the consumed energy for each node with the capability of reusing time slots in HS-FlexiTP, FlexiTP and ZMAC protocol

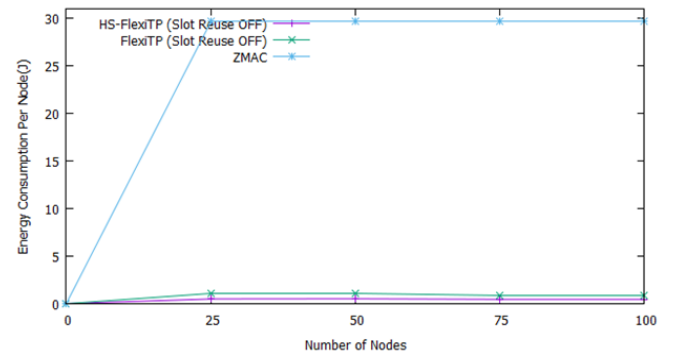


Figure 4. Consumed energy for each node without the capability of reusing time slots in HS-FlexiTP, FlexiTP and ZMAC protocols

As it is clear, the proposed method can have relatively

reduced consumed energy for different scenarios of simulations with different node density, in activation and deactivation of capability of reusing time slots up to 26.3 and 24.9 percent respectively. Performance of the suggested method is significant. It is able to reduce the energy consumption up to 98 percent in comparison with ZMAC protocol. In the activation and deactivation of capability of reusing time slots, FlexiTP is able to optimize the degree of energy consumption about 97.4 and 97.8 respectively in comparison with ZMAC protocol. FlexiTP protocol consumes 2.51 (0.746 joule) and 2.18 (0.649 joule) percent of the whole energy that ZMAC protocol consumes in activation and deactivation of capability of reusing time slots states respectively with the similar simulation parameters. This was predictable, as the performance of this synthetic protocol in the busy and crowded networks is being degraded to a contention based protocol. In such status, nodes compete with each other in order to obtain the channel and transmit their data which increases the degree of contention in the channel excessively and leads to increase the energy consumption and reduces the network longevity. In addition to this, average packet delay increases because of the contentions and competitions happening in the channel. On the other side, FlexiTP protocol has excessively reduced the probability of contention in this channel by assigning the accurate time slots specified to each node and their waking up according to this timing. Therefore nodes have no competition for gaining the channel and considering the timing that has been allocated to them. They activate and deliver their data to their parent nodes and go to sleep again and wait for the next period of transmission. The proposed method has consumed 1.85% (0.55 joule) and 1.58% (0.487 joule) of the whole consumed energy of ZMAC in activation and deactivation of capability of reusing time slots respectively. Consequently it is able to optimize the degree of energy consumption in those states up to 98.1 and 98.4 respectively in comparison with ZMAC protocol. Obtained energy of each node indicates the performance of proposed method in optimization of energy consumption in comparison with the other protocols. In fact, optimization of energy consumption is due to its efficiency and appropriate performance in two installation and initialization and time slot assignment phases.

One of the most important features exist in FlexiTP protocol is guarantying of the packet delivery through assigning time slots to the nodes. In fact, this protocol is able to eliminate contention in the channel, while making fairness between the nodes to access the network using this mechanism. In comparison with Flexi-TP, the mechanism of proposed method just differs in the generating of data aggregation tree therefore it is expected that they wouldn't be different much in this criterion. However, there is completely different story in ZMAC protocol. The degree of contention in the channel is excessively high while applying ZMAC protocol at networks with heavy traffic. Occurring contention in the channel, in addition to have negative effects on the degree of energy consumption have negative effect on the network efficiency and increase the delay for packet delivery. Obtained results indicate that ZMAC protocol has the highest energy consumption for each node. The difference between FlexiTP protocol and the suggested one is not significant in this criterion but the performance of the

suggested one is better than the mentioned protocol based on this criterion. Obtained results indicate that the suggested method is able to increase the network longevity in comparison with FlexiTP and ZMAC protocol. Possibility of contention in ZMAC has excessively increased the degree of energy consumption of the sensor nodes in this protocol. This feature increases the delay of the packets in this protocol as well.

Figures 5 and 6 indicate the network longevity in activation and deactivation of reusing time slots in previous protocols. Network longevity is increasing up to when nodes are equal to 25 but after that there is a decline when nodes are 50 and it will increase again. It is clear that capability of reusing time slot assignment made a more sharply decline when the number of nodes are increasing up to 50. The main reason of such variation is reaching to a local optimum value of nodes. Obviously reusing time slot assignment make network to have more longevity in general. For the rest of figures the optimal value is 25 nodes for activation of reusing time slot assignment and this value is 50 for deactivation of reusing time slot assignment. Obviously reusing time slot assignment make network to have more longevity in general. As it is obvious from these figures, the proposed method is able to optimize the network longevity in comparison with FlexiTP protocol. Also, the presented longevity of the network by the suggested one is significant in comparison with ZMAC protocol.

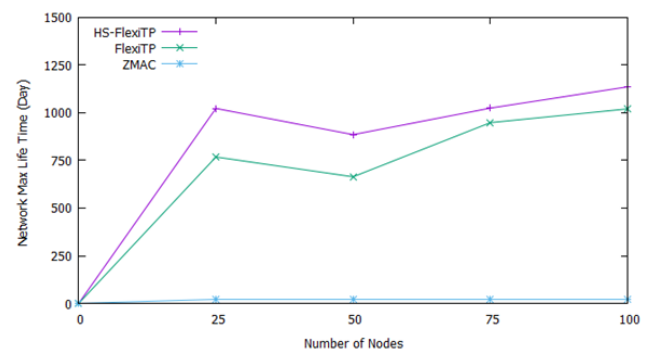


Figure 5. The longevity of the network with the capability of reusing time slot assignment in FlexiTP, HS-FlexiTP and ZMAC

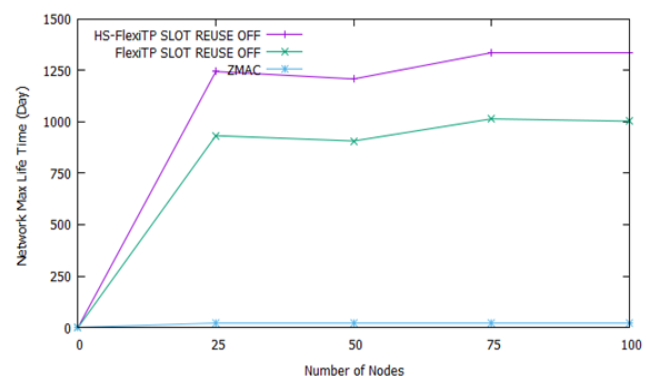


Figure 6. Maximum of the longevity of the network without the capability of reusing time slot assignment in FlexiTP, HS-FlexiTP and ZMAC

7.4. The criteria of efficiency evaluation

In order to evaluate the performance and efficiency of our method and compare it with FlexiTP and ZMAC protocols, four criteria are considered. These criteria are average throughput, percentage of channel efficiency, average packet delay and fairness in nodes accessing to the communicational channel. [5]. In FlexiTP time slots are assigned to each node with equal length. This make it possible for the aggregated data to be transmitted based on the determined timing so fairness is guaranteed [5]. Therefore, this criterion is guaranteed in the suggested method as it is derived from this protocol.

Simulation results indicate that when the capability of reusing assigned time slots is deactivated, network throughput is relatively more than when it is activated. Also, the number of received packets in deactivation of capability of reusing time slots is more than its activated status as well. Theoretically, a slot which is assigned to a node at the network can be reused by the other nodes which are a few steps far from that node. Albeit, this will not guarantee the channel being without contention, as the results of the simulation indicate that the throughput of the network is relatively higher when there is no possibility of reusing time slots than when this possibility exists. In [5], the reason of this is said to be contention or irregularity in radio channel. In ZMAC protocol there is not a feature called capability of reusing or not reusing time slots, and this feature has not been defined in the body of this protocol. Simulations results indicate that due to numerous contentions in high node densities, ZMAC protocol has the worst performance among the three mentioned protocols in the field of average packet delay, channel efficiency and throughput. According to the predictions, in scenarios with high nodes density, the number of the packets that are delivered to the base station in the suggested method in activation of the capability of reusing time slots assignment is more in comparison with the main protocol. This happened due to decreasing of packet delay through choosing parent nodes with less distance in the proposed method. Anyway, the degree of the delivered packets has been equal in the time of deactivation of the capability of reusing time slots assignment, as in this status, the probability of contention in channel, has been zero. The degree of the packet delay of the suggested procedure, in both activation and deactivation of reusing time slots has been improved. Also, based on predictions, the degree of the throughput and efficiency of the channel is equal for both protocols in both activated or deactivated status of reusing time slots assignment. Although, in scenarios with high node density and active reusing time slots assignment, considering high contention rate and radio interactions and also high packet delay in FlexiTP protocol, based on the expectations, the suggested method has a better performance in throughput. In figures 7 to 12, there is a performance comparison of the suggested method with ZMAC and FlexiTP protocols in criteria such as average packet delay, average throughput and average channel efficiency for both being activation and deactivation of the capability of reusing

time slot assigned to the existing nodes at the network. As it is mentioned before, the optimal value is 25 nodes for activation of reusing assigned time slot and this value is 50 for deactivation of reusing assigned time slot. For example, in figure 12, throughput is maximum when the number of nodes are 25 without reusing assigned time slot for Flexi-TP and the proposed method while it is maximized in figure 11 with reusing assigned time slot for the same protocols.

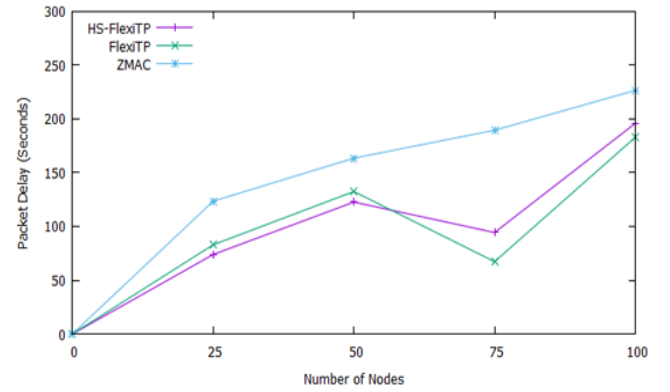


Figure 7. The degree of the packet delay with the capability of reusing assigned time slots for ZMAC, FlexiTP and HS-FlexiTP protocols

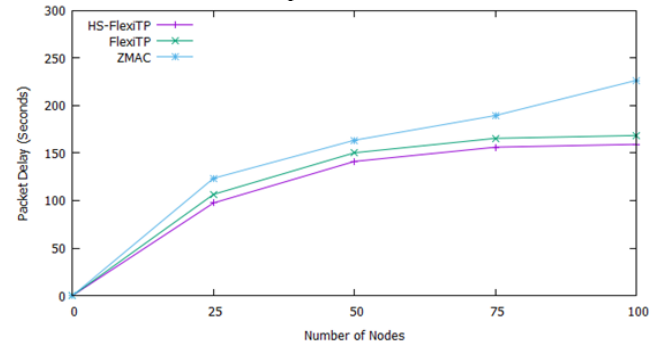


Figure 8. The degree of packet delay without the capability of reusing assigned time slots for ZMAC, FlexiTP and HS-FlexiTP protocols

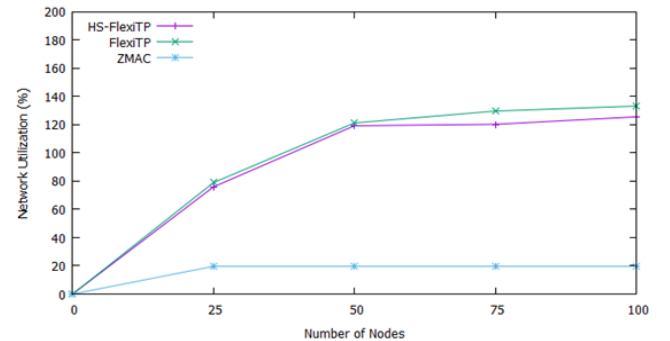


Figure 9. Efficiency percentage of the network with the capability of reusing assigned time slots for ZMAC, FlexiTP and HS.FlexiTP protocols

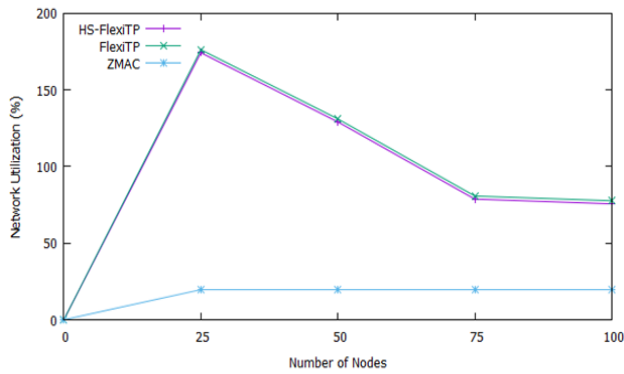


Figure 10. The percentage of the network efficiency without the capability of reusing assigned time slot for ZMAC, FlexiTP and HS.FlexiTP

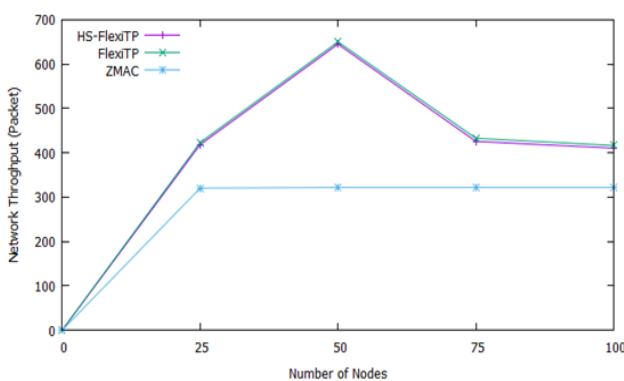


Figure 11. Average network throughput with the capability of reusing assigned time slots for ZMAC, FlexiTP and HS.FlexiTP protocols

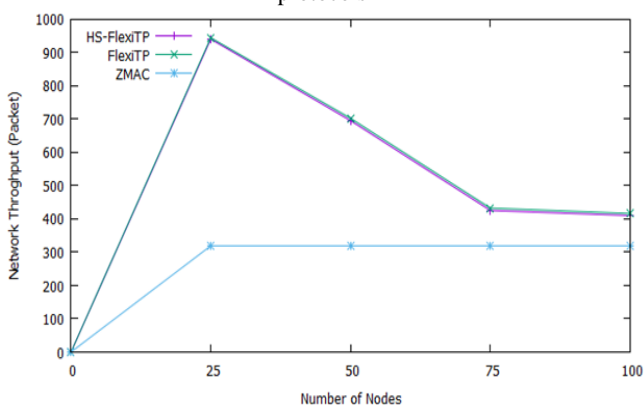


Figure 12. Average network throughput without the capability of reusing the assigned time slots for ZMAC, FlexiTP and HS.FlexiTP protocols

8. Discussion

In this study some important parameters such as distance, hops and energy are considered to evaluate proposed method with relevance works in this field. Some other parameters are measured for comparison such as average packet delay, throughput, residual energy of the node and its parents and productivity of the channel. Also average throughput, percentage of channel efficiency and fairness in nodes accessing to the communicational channel are considered in this evaluation. Longevity of the network is shown in Figure 5 and 6. Improvement of the packet delay of suggested procedure, in both activation and deactivation of reusing time slots is shown in Figure 7 and 8 respectively. Figure 9

and 10 show the efficiency of suggested method in above cases which is computed based on the average degree of consumed energy for each node, the average degree of consumed energy for each packet and the maximum of the network longevity. Throughput Improvement of suggested method is shown in Figure 11 and 12 in both activation and deactivation of reusing time slots respectively. Obtained results indicate that the suggested method is able to increase the network longevity in comparison with FlexiTP and ZMAC protocol. Also suggested method have a better performance in throughput. By comparing the results of simulations, HS.FlexiTP reduces installation time span, initialization, and assigning time slot and energy consumption for each node and increase longevity of the network by choosing the best parent node for existing nodes in the network.

9. Conclusion and future works

In this research Harmony Search algorithm has been applied in FlexiTP protocol in a way that for each node, the parent would be computed according to the degree of the fitness function based on the distance, the remaining energy of the node and the number of hops of the neighbor nodes. Applying this procedure cause the reduction in the time span of installation and initialization, time slot assignment and energy consumption for each node and increasing the longevity of the network while maintaining efficiency parameters in comparison with FlexiTP protocol. In order to perform and depict the results of the suggested method and comparing it with FlexiTP and ZMAC protocol, simulation is performed in NS2 software. Generally by comparing the results of simulations for two FlexiTP and ZMAC protocols and the suggested one, HS.FlexiTP cause reduction in the installation time span and initialization, assignment of time slot and energy consumption for each node and increasing longevity of the network by choosing the best parent node for any node exist in the network.

The performance of the HS algorithm in FlexiTP protocol clearly indicated that this protocol can be used for modification of the protocols which use the structure of the data aggregation tree. Consequently, the objectives of the future research could be redesigning the protocols which have used hierarchy procedures and data aggregation tree structure in order to make it possible to see the obtained results after implementation of the algorithm in optimization of their energy consumption.

conflicts of interest: The author declares that he has no conflict of interest.

9. References


- [1] Ye, W., Heidemann, J., Estrin, D., "An energy-efficient MAC protocol for wireless sensor Networks", in: *21th Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM)*, Vol. 3, pp. 1567–1576, 2002.
- [2] Van Dam, T., Lengendoen, K., "An adaptive energy efficient MAC protocol for wireless sensor networks", in: *1st ACM Conference on Embedded Networked Sensor Systems*, pp. 171–180, 2003.

- [3] Polastre, J., Hill, J., Culler, D., "Versatile low power media access for wireless sensor networks", in: *Proc. of the Second ACM Conference on Embedded Networked Sensor Systems*, pp. 95–107, 2004.
- [4] Kumar, S. S., Kumar, M. N., Sheeba, V. S., and Kashwan, K. R., "Power Efficient Dynamic MAC Protocol (D-MAC) for Wireless Sensor Network", *Journal of Information & Computational Science*, Vol. 9, No. 7, pp. 1795–1805, 2012.
- [5] Lee, W. L., Datta, A., Cardell-Oliver, R., "Flexitp: A Flexible-Schedule-Based TDMA Protocol For Fault-Tolerant And Energy-Efficient Wireless Sensor Networks", *IEEE Transactions On Parallel And Distributed Systems*, Vol. 19, No. 6, 2008.
- [6] El-Hoiydi, A., and Decotignie, J. D., "WiseMAC: An Ultra-Low Power MAC Protocol for Multi-Hop Wireless Sensor Networks", In *Proceedings of the first Springer Verlag International Workshop on Algorithmic Aspects of Wireless Sensor Networks (ALGOSENSORS'04)*, July 2004.
- [7] IEEE Standard. 802.15.4-2011: Wireless Medium Access Control (MAC) and Physical Layer (PHY) Specifications for Low Rate Wireless Personal Area Networks (LR-WPANs); IEEE Standard: New York, NY, USA, 2011.
- [8] Park, P., Marco, P. D., Soldati, P., Fischione, C., Johansson, K. H., "A Generalized Markov Chain Model for Effective Analysis of Slotted IEEE802.15.4", In *Proceedings of the IEEE 6th International Conference on Mobile Adhoc and Sensor Systems, Macau, China*, 12–15 October; pp. 130–139, 2009.
- [9] Khan, A. A., Ghani, S., Siddiqui, S. A., "Preemptive Priority-Based Data Fragmentation Scheme for Heterogeneous Traffic in Wireless Sensor Networks", *Sensors* 2018, Vol. 18, No. 4473, 2018.
- [10] Rajendran, V., Obraczka, K., and Garcia-Luna-Aceves, J. J., "Energy-Efficient Collision-Free Medium Access Control for Wireless Sensor Networks", *Proc. First ACM Int'l Conf. Embedded Networked Sensor Systems (SenSys '03)*, Mar. 2003.
- [11] Mengjie, Y., Mokhtar, H., and Merabti, M., "Fault management in wireless sensor networks", *IEEE Wireless Communications*, Vol. 14, pp. 13–19, 2007.
- [12] Alagoz, I. B. F., "Energy Efficient Delay Sensitive Fault Tolerant Wireless Sensor Network for Military Monitoring", *International Journal of Distributed Sensor Networks*, Vol. 5 pp. 729–747, 2009.
- [13] Dhawan, A., Parks, M., "Fault-tolerant Coverage in Dense Wireless Sensor Networks", *2nd International Conference on Sensor Networks (SENSORNETS)*, Barcelona, Spain, February, 2013.
- [14] Rhee, I., Warrier, A., Aia, M., Min, J., and Sichitiu, M. L., "Z-MAC: Hybrid MAC for Wireless Sensor Networks", *IEEE/ACM Transactions on Networking*, Vol. 16, No. 3, June 2008.
- [15] Jovanovic, M. D., Goran, L. J., Jordjevic, D., "Reduced-Frame TDMA Protocols for Wireless Sensor Networks", *International Journal of Communication Systems*, Vol. 27, No. 10, pp 1857–1873, 2014.
- [16] Ramya, R., Saravanakumar, S., and Ravi, S., "MAC Protocols for Wireless Sensor Networks", *Indian Journal of Science and Technology*, Vol. 8 No. 34, DOI: 10.17485/ijst/2015/v8i34/72318, December, 2015.
- [17] Hoseini, R., Mirvaziri, H., "A New Clustering-Based Approach for Target Tracking to Optimize Energy Consumption in Wireless Sensor Networks", *Wireless Personal Communications*, Vol. 111, pp. 729–751, 2020.
- [18] Poonguzhali, P. K., Ananthamoorthy, N. P., "Design of Mutated Harmony Search Algorithm for Data Dissemination in Wireless Sensor Network", *Wireless Personal Communications* 111", pp.729–751, 2020.
- [19] Geem, Z. W., Yoon, Y., "Harmony search optimization of renewable energy charging with energy storage system", *International Journal of Electrical Power & Energy Systems*, Vol. 86, pp. 120-126, 2017.
- [20] Nazari-Heris, M., Mohammadi-Ivatloo, B., Asadi, S., and Woo Geem, S., "A comprehensive review on the applications of HS method to energy systems", *Journal of Experimental & Theoretical Artificial Intelligence*, Vol. 31, No. 5, pp. 723-749, 2019.



Dynamic Security Risk Management Considering Systems Structural and Probabilistic Attributes*

Research Article

Masoud Khosravi-Farmad¹, Abbas Ghaemi-Bafghi² 

DOI: [10.22067/cke.2023.83744.1102](https://doi.org/10.22067/cke.2023.83744.1102)

Abstract: Today's cyber-attacks are getting more sophisticated and their volume is consistently growing. Organizations suffer from various attacks in their lifetime each of which exploiting different vulnerabilities, therefore, preventing them all is not affordable nor effective. Hence, selecting the optimal set of security countermeasures to protect IT assets from being compromised is a challenging task which requires various considerations such as vulnerabilities characteristics, countermeasures effectiveness, existing security policies and budget limitations. In this paper, a dynamic security risk management framework is presented which identifies the optimal risk mitigation plans for preventing ongoing cyber-attacks regarding limited budget. Structural and probabilistic analysis of system model are conducted in two parallel and independent aspects in which the most probable system's risk hotspots are identified. Suitability of countermeasures are also calculated based on their ability in covering vulnerabilities and organizational security policies. Moreover, a novel algorithm for dynamically conducting cost-benefit analysis is proposed which identifies optimal security risk mitigation plans. Finally, practical applicability is ensured by using a case study.

Keyword: Attack Graph, Bayesian Networks, Cost-Benefit Analysis, Countermeasure Analysis, Security Risk Management.

1. Introduction

By evolving technology and increasing the number of sophisticated IT related threats and growth of cybercriminals capabilities in exploiting security vulnerabilities, reducing security risks by protecting valuable assets is becoming the greatest concern for digitized companies. Risk is defined as the net negative impact of the exploitation of security vulnerabilities and is determined by considering the probability of successful exploit of vulnerabilities and the impact they incur on the confidentiality (C), integrity (I), and availability (A) requirements of assets, known as CIA requirements [1]. Security risk management generally is the process of

identifying, assessing and mitigating risks to an organization's IT assets [2, 3]. The risk identification stage is the process of identifying assets and their significance in bringing an organization closer to its goals. Vulnerabilities putting the CIA requirements of these assets at risk are also identified in this stage. The risks are then determined using probability of exploiting assets vulnerabilities combined with their overall consequences in the risk assessment stage. After that, in the risk mitigation stage, the highest-ranked risks are selected to be treated by lessening the probability and/or impact of them.

A variety of security risk assessment methodologies have been proposed in the literature which can be broadly categorized into qualitative, quantitative and semi-qualitative methods [4, 5]:

1. In qualitative methods such as [2, 6], Information security risks are assessed using relative non numerical values (e.g. low, medium, and high). These methods are useful for dealing with situations which are not well defined. While qualitative methods are simple and easy to understand and implement, they lack enough accuracy and precision in calculations involved. Also, these methods are based on the knowledge and experience of assessors, making them more subjective and error prone than quantitative methods [2]. Moreover, since the range of qualitative values are relatively small, risk prioritization and comparison is comparatively difficult [1].
2. In quantitative methods such as [7-9], Information security risks are assessed using numbers. These methods are based on objective measurements, hence, the results are more accurate and clear. While these methods have advantages, they meet several problems. For instance, because of limited time, budget, and human resources available, their implementation complexity is more than their qualitative counterparts. Moreover, exact detailed information about system attributes may not always be easily extractable from experts when not enough accurate historical data is

* Manuscript received: 2023 August 12, Revised, 2023 September 9, Accepted, 2023 November 1.

¹ Corresponding author. Associate Professor, Data and Communication Security Lab., Computer Engineering Department, Ferdowsi University of Mashhad, Mashhad, Iran. **Email:** m.khosravi@mail.um.ac.ir.

² PhD Student, Computer Engineering Department, Ferdowsi University of Mashhad, Mashhad, Iran.

available [10].

3. Semi-qualitative methods such as [11, 12], try to combine advantages of both quantitative and qualitative methods. They can benefit the simplicity and understandability of the qualitative methods, while taking advantage of the accuracy of quantitative methods.

There are several standards available for assessing individual IT systems security vulnerabilities. One of the most common and widely adopted standards is Common Vulnerability Scoring System (CVSS) [13] developed by the Forum of Incident Response and Security Teams (FIRST) [14] which assigns both numerical scores and relative values to identified vulnerabilities.

Most of existing approaches take into account the overall scores of vulnerabilities for assessing systems security risks. These values usually represent the security level of coarse-grained attributes of IT systems. This viewpoint makes the process of risk assessment done straightforward, but it faces obstacles in risk mitigation process when cost-benefit analysis is required to determine appropriate risk control recommendations [5]. For instance, consider three vulnerabilities existing on .NET Framework developed by Microsoft which are listed in Table 1. Suppose that, in case of budget limitation, the security administrator tries to specify a proper countermeasure for covering only one of these vulnerabilities regarding their scores. As can be seen, these vulnerabilities are hard to be distinguished based on only their CVSS Base Scores. Because, despite the major difference, all of them have the same score equal to 7.5. To overcome this problem, the values of CIA parameters of vulnerabilities should also be taken into account in decision making. For example, if confidentiality is more important for an asset, then the security administrator should identify a countermeasure covering CVE-2016-0047 and leave other vulnerabilities unpatched. For more information about CVSS scoring system refer to [13].

Since during modern sophisticated attacks such as Advanced Persistent Threats [15-17] sequences of vulnerabilities are usually exploited to perform multi-step attacks to achieve particular goals, utilizing only individual scores or values of these standards is not sufficient, because they do not consider the interactions between vulnerabilities. In order to be able to clearly demonstrate multi-step attacks, one can use graph-based security models such as Attack Graph (AG) [18-20]. AG is a powerful model that can encode causal relationships between vulnerabilities and give description about the correlated attacks.

Risk mitigation is a crucial stage in the process of risk management which is required for successfully reducing systems security risks [2]. It includes prioritizing, implementing and maintaining the most suitable security countermeasures. The input data for risk mitigation is provided from risk assessment results. Therefore, a risk assessment report is beneficial only when it is compatible

with risk mitigation processes.

In this paper, we present a method for managing IT systems security risks which uses both numerical and relative values. In cases when input values are supplied by experts and security administrators, relative values are used to ease the process of data extraction. In cases when enough data is available in existing security databases or repositories, exact values are used. We use AG as a graphical security model for modeling different attack scenarios targeting IT assets. We analyze AGs in two ways:

1. **Structural analysis:** Each AG contains several attack paths each of which represents an attack scenario. Therefore, it is an important source for extracting attackers' behavioral information to identify existing risk hotspots. We identify these risk hotspots using defined metrics over the structure of AG.
2. **Probabilistic analysis:** by assigning a probability to each node of AG and applying Bayesian theory we can compute unconditional probability (UP) of attackers reaching to different states in the graph. These probabilities will play an important role in final risk reducing decision making.

Moreover, a parametric solution for countermeasure analysis is presented which first calculates the coverage level of vulnerabilities by countermeasures based on fine-grained attributes. After that, considering organizational security policies on assets, the suitability of countermeasures for implementation is identified.

Briefly, the main contributions of this work are:

1. A dynamic security risk management framework is presented which uses exact values when enough data is available and uses relative values when data need to be extracted from experts' knowledge.
2. Structural and probabilistic analysis of AG model are conducted in two parallel ways. In structural analysis, AGs risk hotspots are identified and in probabilistic analysis the UP of attackers reaching their goals are calculated utilizing Bayesian theory.
3. The countermeasures' ability in reducing vulnerabilities impact are calculated in terms of C, I and A parameters.
4. Utility of countermeasures are calculated based on their ability in reducing vulnerabilities, their negative effects on the service level agreements, organizational security policies and systems risk hotspots.
5. A novel algorithm for dynamically conducting cost-benefit analysis is presented which identifies optimal security risk mitigation plans.

The paper is structured as follows. Section 2 reviews related work. Section 3 presents the concepts used for modeling systems attributes. The proposed risk management framework is presented in Section 4. Experimental results are given in Section 5. Finally, Section 6 concludes the paper and discusses future work.

Table 1. Sample .Net Framework Vulnerabilities

CVE ID	Confidentiality (C)	Integrity (I)	Availability (A)	CVSS v3 Base Score	Vector
CVE-2016-0047	High	None	None	7.5 (High)	CVSS:3.0/AV:N/AC:L/PR:N/UI:N/S:U/C:H/I:N/A:N
CVE-2017-0248	None	High	None	7.5 (High)	CVSS:3.0/AV:N/AC:L/PR:N/UI:N/S:U/C:N/I:H/A:N
CVE-2016-0033	None	None	High	7.5 (High)	CVSS:3.0/AV:N/AC:L/PR:N/UI:N/S:U/C:N/I:N/A:H

2. Related Work

A variety of cybersecurity risk management methodologies have been developed for assessing risks in IT systems, thereby, enabling systems security administrators to make correct decisions towards mitigating the most important risks in the operational environments.

Qualitative security risk management methods rate both input and output attributes of a system using a scale of usually three or five levels (e.g., very low, low, moderate, high, very high) [2, 21]. Since these methods have several disadvantages, including their inappropriateness in making a cost-benefit analysis of recommended controls, quantitative methods are preferred. The major advantage of a quantitative method is that it most effectively supports cost-benefit analysis of alternative risk-reducing measures [1]. Quantitative methods such as [9, 22-25] typically employ sets of methods, principles or rules for managing risks based on the use of numbers. Semi-qualitative methods can provide the benefits of quantitative and qualitative methods [1].

Because AG-based security models can properly model multi-step attacks, they are popular in both qualitative and quantitative risk management activities [18, 19, 26-31]. Some approaches apply Bayesian concept over AG to represent information about causal relationships between vulnerabilities and capture uncertainties about probabilities of attacker actions. One of the first researches in this field is [32] which models attack paths using Bayesian networks and quantitatively represents the security of computer networks. Frigault and Wang [33] used Bayesian networks with AGs to calculate security metrics. They named their model Bayesian attack graph. After that, Poolsappasit et al. [34] extended their model to be able to dynamically analyze existing risks in networked systems. In [35], a security risk analysis model based on Bayesian networks and ant colony optimization algorithm is proposed which estimates risk values. In [9, 36] authors used Bayesian networks to implement Factor Analysis of Information Risk (FAIR) as one of the most popular models for quantitative security risk assessment. There are several works that use Bayesian networks for assessing security risks of IT systems and capturing uncertainties in attacker actions, such as [37-42]. Aforementioned methodologies use Bayesian inference results for calculating the risk level of systems and do not consider anatomy of attack scenarios and their interactions in forming successful attacks. The methodology proposed in this paper not only uses Bayesian inference results, but also takes into account the topology and structure of security

model in risk assessment calculations.

While most of existing researches focus on risk assessment and vulnerability analysis, fewer studies proposed methods for risk mitigation and countermeasure analysis. The reason is that there is no standard and comprehensive database for security countermeasures [43]. Moreover, most of risk mitigation processes are largely dependent to expert's knowledge. In [44], minimum-cost countermeasures are identified using exploit dependency graphs. In [45], the minimal subset of attacks that are necessary for reaching a goal in the network is determined. After that, the minimal subset of countermeasures that covers the subset of attacks is identified. Dewri et al. [46] used a multi-objective optimization problem on the security model of the network to determine if a given set of security hardening measures effectively secures the system. In [34], authors proposed a Bayesian attack graph model which assigns cost and outcome values to each countermeasure. After that, by applying genetic algorithm solutions, countermeasures with the highest outcome given a specific budget are identified. Authors in [47] proposed Bayesian decision networks to manage security vulnerabilities and conduct cost-benefit analysis by using a variable elimination-based algorithm to identify the optimal subset(s) of security countermeasures. Authors in [48] assign an effectiveness value to each countermeasure. This value represents the percentage of probability reduction of vulnerabilities for which the countermeasure is implemented on. In [49], authors define safeguard effectiveness as the ability of safeguards in reducing the criticality of threats. Authors in [50] define countermeasure effectiveness as risk mitigation level after the countermeasure implementation. The mentioned methods are useful, but the main problem with such methods is that countermeasures effectiveness is assigned statically and security experts are responsible to assign numeric outcomes to each countermeasure solely, regardless of systems vulnerabilities and ongoing attacks. This, indeed, is not an easy task, because, outcome of a countermeasure is dependent on its ability in remedying its covered vulnerabilities and dynamic state of the system. Moreover, most of existing researches neglect countermeasures' negative impact on service quality and service level agreements which may lead to select inappropriate countermeasures and as a result, reducing network performance.

In this paper, a dynamic security risk management framework is presented which utilizes fine-grained attributes of IT systems to handle the aforementioned drawbacks in

existing methods. Using relationships between security policies on assets, security vulnerabilities existing on assets and countermeasures covering these vulnerabilities, we can manage IT systems security risks properly.

3. Modeling system attributes

Application of the proposed framework requires a keen understanding of the system-related information. Hence, we need to model the attributes of the system under assessment appropriately. For this reason, we define security attributes as necessary system-related information sources for doing risk management. We categorize the security attributes into three main classes, namely, Assets, Vulnerabilities and Security Countermeasures. Each attribute has three requirements, namely, confidentiality (C), integrity (I) and availability (A), known as CIA requirements. Asset's CIA represent the importance of CIA requirements of assets which are determined according to the organizations security policies. Each vulnerability can impact assets in terms of CIA parameters. Security countermeasures can reduce this impact and therefore protect assets from being compromised. The relations between the security attributes is depicted in Figure 1.

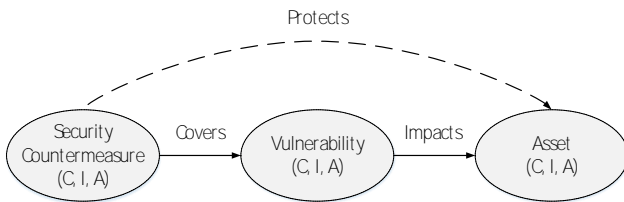


Figure 1. Relations between Assets, Vulnerabilities and Security Countermeasures

The information for each of the attributes are modeled using a vector. A brief description about each vector and its components is presented below:

- **Asset (A):** we define assets as any hardware or software component in the system under assessment which supports information-related activities. Therefore, assets could be hosts existing in the system, operating systems, services and software running on them. The CIA requirements of assets are also included in vector. Moreover, each asset is assigned a unique identifier. Hence, the Asset Vector is defined as below:

Asset = {Identifier, Asset Details (Name, Operating System, Service, Software, Hardware, Protocol, etc.), CIA Requirements}

- **Vulnerability (V):** we define vulnerability as any weakness or flaw existing on an asset configuration which could be exploited by malicious attackers and result in violation of the system's CIA requirements. Each vulnerability is associated with a CVE ID which is a unique identifier for publicly disclosed information security vulnerabilities. The impact on CIA requirements of successful exploitation of vulnerabilities and the exploitation probability of vulnerabilities are also included in the vector. Therefore, the Vulnerability Vector is defined as below:

Vulnerability={CVE_{ID}, Vulnerable Asset Configurations

(Operating System, Service, Software, Hardware, Protocol, etc.), Impact on CIA, Exploitation Probability}

- **Security Countermeasure (SC):** a security countermeasure or a security control is a protecting measure which reduces the vulnerability of an asset by protecting its CIA requirements. Each SC is assigned a unique identifier and also the IDs of covered vulnerabilities. We separate the efficacy of SCs into two classes; 1- a SC could reduce the impact of a vulnerability on CIA requirements of an asset (impact reduction (IR)), and/or 2- it could reduce the exploitation probability of a vulnerability (probability reduction (PR)). Implementation of SCs may bring negative effects on the service level agreements. Therefore, we define intrusiveness (I) which reflects this effect. Moreover, each SC has a cost of implementation (IC). Hence, the Security Countermeasure Vector is defined as below:

Security Countermeasure = {Identifier, IDs of Covered Vulnerabilities, Impact Reduction on CIA (IR), Probability Reduction (PR), Intrusiveness (I), Implementation Cost (IC)}

In the next Section, we express how these vectors are used in the proposed security risk management framework.

4. The Proposed Risk Management Framework

The proposed dynamic security risk management framework, consists of four activities, namely, vulnerability scanning, modeling network attacks, countermeasure analysis and dynamic cost-benefit analysis. These activities contain seven processes, namely, vulnerability scanning, attack graph generation, attack graph analysis, Bayesian inference, vulnerability coverage assessment, policy conformance assessment and cost-benefit analysis. It starts by identifying the vulnerabilities existing on system assets. After that, there are two activities which could be done in parallel. The first activity is modeling network attacks, in which the AG model is used to model attack scenarios. The generated model is then analyzed and some structural metrics are extracted from it. These metrics will be used in dynamic cost-benefit analysis process. Moreover, to consider the interconnections between vulnerabilities, Bayesian inference algorithm is applied on the AG model. Therefore, UP of compromising each network state is calculated. These probabilities are also used in dynamic cost-benefit analysis process. In the second activity, i.e. countermeasure analysis, a mapping analysis is conducted to calculate the coverage level of vulnerabilities impact by SCs in terms of C, I and A parameters. After that, the conformance between coverage levels of vulnerabilities with the importance of C, I and A parameters of assets are calculated regarding organizational security policies. Therefore, the suitability of countermeasures for covering vulnerabilities existing on specific assets can be calculated. Finally, considering extracted metrics from AG model, UP of network states and suitability of each SC, a cost-benefit analysis is conducted to identify the optimal security risk mitigation plans to reduce the overall risk level of the system. The data flow diagram of the proposed dynamic security risk management framework is depicted in Figure 2.

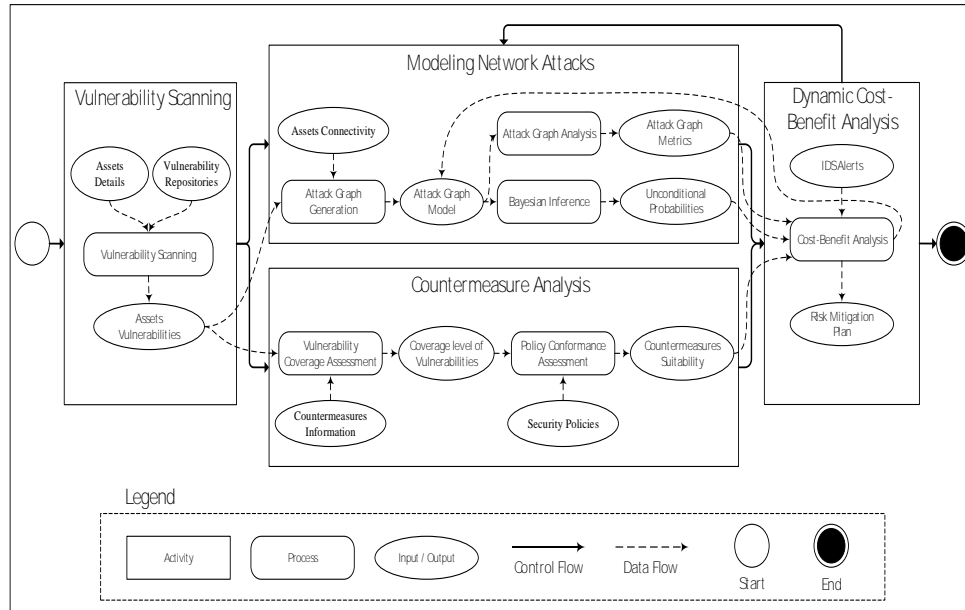


Figure 2. The Proposed Security Risk Management Framework

In this section, each of the mentioned activities and processes of the proposed risk management framework is described in detail.

4.1. Vulnerability Scanning

In this activity, all of the system's assets such as hardware, software, operating systems and services are scanned to search for security vulnerabilities. There exist several vulnerability scanners that can be used for this purpose such as Nessus [51], OpenVAS [52] and Retina [53]. After finding a vulnerability, the information about it can be extracted from existing vulnerability repositories such as the US National Vulnerability Database (NVD) [54] and MITRE's Common Vulnerabilities and Exposures (CVE) [55]. The vulnerabilities information containing their technical description, CVE ID and metrics values gathered in this activity will be used in the next steps. After discovering the system's assets vulnerabilities, we define an Asset Vector for each vulnerability. Moreover, vulnerabilities existing on each asset are listed in a table called Assets Vulnerabilities for future use. This table represents a mapping between assets and vulnerabilities existing on them.

4.2. Modeling Network Attacks

This activity consists of three processes, namely, attack graph generation, attack graph analysis and Bayesian inference. At first, the AG model of the system under assessment is generated. After that, the generated model is processed in two parallel processes. In the attack graph analysis process, some structural metrics of the graph are extracted which represent the risk hotspots in the graph. In the Bayesian inference process, the AG model is converted into the Bayesian attack graph. Therefore, by applying Bayesian theorem, the UP of each network state being compromised by attackers is calculated.

The three mentioned processes are further explained in the following subsections.

A. Attack Graph Generation

This process aims at generating a model based on the information about assets vulnerabilities and their connectivity. Here, the assets connectivity is an important factor in modeling process, because, most of modern sophisticated attacks utilize several vulnerabilities existing on different assets in various sequences. These attacks are called multi-step attacks which can be properly represented by attack graph model. Attack graph is a powerful tool that can demonstrate all attack scenarios an adversary can utilize to compromise a system by modeling vulnerabilities and interactions between them. Therefore, each node of the attack graph represents a state in which a vulnerability exploits.

Attack Graph Definition: An attack graph (AG) is a tuple $AG = (S, s_0, s_g, \tau)$, where:

- S is a set of states in the network. Each state represents an exploitation of a vulnerability.
- $s_0 \subseteq S$ denotes the attacker's entry point to the network and hence is the initial state in the graph.
- $s_g \subseteq S$ is the set of potential goals for attackers.
- $\tau \subseteq S \times S$ is the set of directed arcs that change the states of the network.

To generate an AG for a given system, information about vulnerabilities existing on assets and connectivity of assets are required. Information about assets vulnerabilities are provided from output of Vulnerability Scanning process. To determine how assets are connected together, one can refer to the documentations about system topology provided by system administrator or use available tools like Nmap [56] security scanner. After identifying assets vulnerabilities and their connectivity, AG model of the system can be generated using existing tools such as MulVAL [57] and TVA [58].

B. Attack Graph Analysis

After generating AG model of the system under assessment, its risk hotspots should be identified. We define risk hotspots as the most important nodes within the structure of an AG.

To find these hotspots, four AG structural metrics, namely exposure, path length, closeness centrality and betweenness centrality are used in this paper which are defined as follows:

Exposure for $node_i$ is defined as the summation of indegree and outdegree of $node_i$ in the AG model.

Path length for $node_i$ is defined as the length of the shortest path between the leaf node (i.e. attacker's entry point) and the root node of the AG model (i.e. attacker's goal) visiting $node_i$.

Closeness centrality for $node_i$ is defined as the length of the shortest path between $node_i$ and the root node of the AG model. It represents how close $node_i$ is to attacker's goal.

Betweenness centrality for $node_i$ is defined as the number of paths that pass between the leaf node and the root node of the AG model visiting $node_i$.

Using the defined metrics, we present Equation 1 to identify the importance level, i.e. centrality, of each state node in the attack graph:

$$importance = \left(\frac{exposure \times betweenness\ centrality}{path\ length \times closeness\ centrality} \right) \quad (1)$$

As can be seen, in Equation 1, importance of a node have a direct relation with its exposure and betweenness centrality, but have an indirect relation with its path length and closeness centrality. In fact, if a node gets involved in more attack scenarios and if these scenarios are more critical, then this node is at higher importance and represents a more at-risk spot in the system.

C. Bayesian Inference

A Bayesian network is a probabilistic graphical model which uses Bayesian inference techniques for probability computations [59]. In order to be able to apply Bayesian inference over the generated AG model, conditional probability tables should be added to each of its nodes [8, 60]. Each conditional probability table represents conditional probability of a network state with respect to its parents. The entries of conditional probability tables are filled with the probabilities of vulnerabilities exploitations. To calculate the exploitation probability of vulnerabilities, their CVSS Base scores are used in this paper. CVSS is an open framework which provides a way to assess the severity level of IT vulnerabilities [13]. Since the CVSS's scores are in the interval of [0 – 10], we divide them by 10. As the result, it is possible to calculate the UP of compromising network states by attackers [60]. UP of a network state indicates the likelihood that this state gets compromised independent of whether any other states are compromised by attackers. States with higher unconditional probability represent gears that attackers can easily take advantage of. More detailed information about Bayesian inference techniques and algorithms can be found in [59, 61].

4.3. Countermeasure Analysis

This activity consists of two processes, namely, vulnerability coverage assessment and policy conformance assessment which are explained in the following subsections.

A. Vulnerability Coverage Assessment

Before conducting vulnerability coverage assessment, we need to find SCs covering system vulnerabilities and define

a Security Countermeasure Vector (see Section 3) for each one of them. This information can be gathered from vulnerability repositories, publicly available security reports and documents and also security administrator knowledge.

After defining Security Countermeasures Vectors, we need to assess the efficacy of SCs coverage, i.e. impact reducing ability, over their covered vulnerabilities. For this reason, we propose a table called Coverage Table for each one of the C, I and A requirements separately which is shown in Table 2.

Table 2. Coverage Table Definition for C/I/A Requirement

Vulnerability (C/I/A)	Security Countermeasure (C/I/A)	Coverage Level
None	None	Equal coverage
None	Partial	Extra coverage
None	Complete	Extra coverage
Partial	None	No coverage
Partial	Partial	Equal coverage
Partial	Complete	Extra coverage
Complete	None	No coverage
Complete	Partial	Little coverage
Complete	Complete	Equal coverage

The first column of this table represents the magnitude of impact on C/I/A caused by exploitation of a vulnerability. The second column represents the ability of a SC to mitigate the C/I/A impact of its covered vulnerabilities. And the third column represents the coverage level resulted by implementing a SC on its corresponding vulnerability in terms of C/I/A.

For instance, the first row of this table can be interpreted as follow: If exploitation of vulnerability V_i doesn't have any impact on C/I/A requirement of an asset, then existing a security countermeasure SC_i with no coverage on C/I/A requirement (or even absence of security countermeasure) results in equal coverage. As another example, the second row of the table says that, if exploitation of vulnerability V_i doesn't have any impact on C/I/A requirement of an asset, then existing a security countermeasure SC_i with partial coverage on C/I/A requirement results in extra coverage. The same interpretations go for the rest of the table.

The output of this process is the C/I/A coverage level for each pair of $\langle Vulnerability, Security\ Control \rangle$.

B. Policy Conformance Assessment

After conducting vulnerability coverage assessment, the coverage levels of vulnerabilities by their covering SCs are available. But this criterion is not sufficient for selecting appropriate SCs, because it doesn't reflect importance of security policies over the CIA requirements of the organizational assets. Hence, we need to consider assets CIA requirements importance in decision making process. To do so, we propose a table called Suitability Table for each one of the C, I and A requirements separately which is shown in Table 3. SCs with higher suitability are more appropriate to be selected to mitigate the risks.

Table 3. Suitability Table Definition for C/I/A Requirement

		Assets C/I/A Requirement Importance		
		None	Partial	Complete
Coverage Level	No coverage	5	3	1
	Little coverage	4	4	2
	Equal coverage	3	7	9
	Extra coverage	2	6	8

The rows of Table 3 represent the coverage levels of SCs over their covered vulnerabilities acquired from Table 2. The columns of this table represent the importance of security policies over the C/I/A requirement of assets. The values of table's entries represent relative suitability of countermeasures implementation based on their ability in covering vulnerabilities and organizational security policies.

For example, if an asset's C/I/A requirement importance is ranked as Complete, then leaving it unprotected, i.e. providing No Coverage, results Suitability of 1 (lowest suitability). But, if we provide Equal Coverage for this asset, we gain the Suitability of 9 (highest suitability).

4.4. Dynamic Cost-Benefit Analysis

During the lifetime of IT systems, they could be targeted to many attacks. It's important to dynamically respond to these attacks to stop attackers' progress and their further intrusions. In this activity, we conduct a dynamic cost-benefit analysis to find the optimal security risk mitigation plans (SRMPs). A SRMP is a subset of SCs which are selected for implementation to cover the most important vulnerabilities and mitigate the overall security risk level of a system. As the result of this step, SRMP with the highest utility and total implementation cost lower than the allocated security hardening budget is identified.

Algorithm 1 represents the steps for dynamically identifying the optimal SRMP for a given attack scenario. The inputs to the algorithm are generated alert by IDS, representing a network state is compromised, AG model of the system under assessment, SC Vectors, SC suitability tables, importance level of AG nodes and allocated budget for system hardening. The output of the algorithm is the optimal SRMP considering budget limitation.

The algorithm starts by selecting the corresponding state to the generated IDS alert, i.e. $state_{Compromised}$ (line 1). Since an alert is generated only when the attack has taken place, we set the probability of $state_{Compromised}$ to 1 and after that, new probabilities of other states in the model are recalculated by applying Bayesian inference over the AG model (lines 2 & 3). If the change in the UP of the goal node in the AG, i.e. ($state_{Goal}$), is less than a predefined threshold, it means that the attack has not a significant effect in compromising the system and hence, we do not perform countermeasure selection. All the descendant nodes of $state_{Compromised}$, including $state_{Compromised}$, are collected into a set T (line 7). After that, for each state in the model, all applicable countermeasures are selected and new probabilities are calculated based on the ability of countermeasures in reducing the probability of exploitations (lines 11 & 12). The influence of each countermeasure is

calculated based on the change in the UP of $state_{Goal}$ (line 13). This value represents the total influence of countermeasures in preventing attackers reaching $state_{Goal}$. After calculating influence level of countermeasures on their corresponding states, the utility of them should be calculated based on Equation 2 (line 19). The utility represents the ultimate usefulness and advantageousness of countermeasures based on countermeasures influence on states, suitability of countermeasures, importance level of states and countermeasures intrusiveness. Finally, while there is enough budget, using a greedy method, in each iteration, a countermeasure with the most utility is selected as part of the risk mitigation plan (line 23) and its cost of implementation is subtracted from budget value (line 24).

Algorithm 1. Dynamic Cost-Benefit Analysis

Input: IDS alert, AG model, SC Vectors, suitability tables, importance level and budget

Output: optimal SRMP

1. Let $state_{Compromised} = state$ node corresponding to the IDS Alert
2. Set $Pr(state_{Compromised}) = 1$
3. $UpdateProbabilities()$
4. **if** ($\Delta UP(state_{Goal}) < threshold$) **then**
5. **return**
6. **end if**
7. Let $T = Descendants(state_{Compromised}) \cup state_{Compromised}$
8. Let $influence[|T|, |SC|] = \emptyset$
9. **for each** $t \in T$ **do**
10. **for each** $sc \in SC$ **do**
11. $Pr(t) = Pr(t) \times (1 - sc.probabilityReduction)$
12. $UpdateProbabilities()$
13. $influence[t, sc] = \Delta UP(state_{Goal}) \times 100$
14. **end for**
15. **end for**
16. Let $utility[|T|, |SC|] = \emptyset$
17. **for each** $t \in T$ **do**
18. **for each** $sc \in SC$ **do**
19. $utility[t, sc] = \frac{influence[t, sc] \times suitability_{sc} \times importance_t}{intrusiveness_{sc}}$
20. **end for**
21. **end for**
22. **while** ($budget > MIN(sc.cost)$)
23. $SRMP = SelectOptimalSC(utility)$
24. $budget = budget - sc.cost$
25. **end while**
26. **return SRMP**

5. Experimental Results

In this section, we study a hypothetical network to validate the rationality, feasibility and efficacy of the proposed method.

In the experimental network shown in Figure 3, there are seven hosts, namely, Web server, Mail server, DNS server, Gateway server, SQL server, Administrative server and local desktops which are located within two zones, namely, DMZ zone and Trusted zone. A firewall is used to separate the

DMZ zone (which is accessible to the public) from the trusted zone. Policies allow Web server to send SQL queries to the SQL server. Local desktops and administrative server use remote desktop service which allows remote communication of employees. Moreover, SSHD protocol is installed on the gateway server to monitor remote connections.

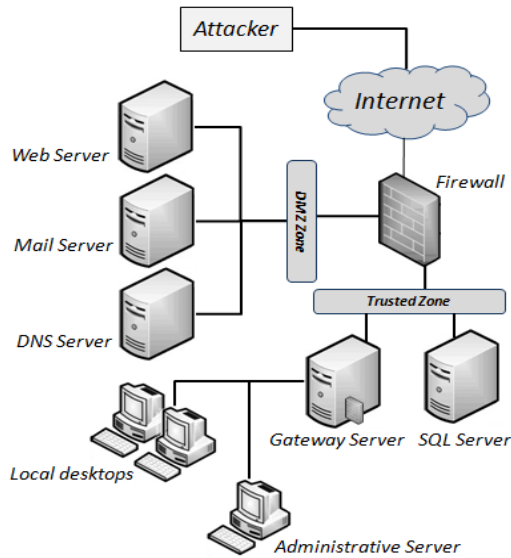


Figure 3. Topology of the Test Network

5.1. Vulnerability Scanning

Having information about the system under assessment and its topology, Security administrator can define an Asset Vector for each asset. These vectors for the test network of Figure 3 are as follow. The definition of Asset Vector is explained in Section 3.

$A_1 = \{\text{Local Desktops, Windows Server, enabled remote login, (Partial, Partial, Partial)}\}$

$A_2 = \{\text{Administrative Server, Windows Server, enabled remote login, (Complete, Complete, Complete)}\}$

$A_3 = \{\text{Gateway Server, Windows Server, Remote Desktop Services, OpenSSH 3.7, (None, None, Complete)}\}$

$A_4 = \{\text{SQL Server, Windows Server, Microsoft SQL Server, (Complete, Complete, Partial)}\}$

$A_5 = \{\text{(Mail Server, Windows Server, Microsoft Exchange Services, (Complete, None, Partial))}\}$

$A_6 = \{\text{DNS Server, Windows Server, DNS protocol, (None, Partial, Complete)}\}$

$A_7 = \{\text{Web Server, Windows Server, Microsoft Internet Information Services (IIS), (Partial, Partial, Complete)}\}$

Each Asset Vector is comprised of the name of an asset, its running operating system and installed services, software and protocols on it. Moreover, each Asset Vector consists of the values of C, I and A requirements which are assigned by network security administrator. For instance, the availability requirement of asset A_4 is *Partial*, while its confidentiality and integrity requirements have *Complete* importance.

After defining Asset Vectors, these assets should be scanned to find their security vulnerabilities. Here, Nessus [51] vulnerability scanner is used. Using the results, we define Vulnerability Vectors as follow. The definition of Vulnerability Vector is explained in Section 3.

$V_1 = \{\text{CVE 2019 - 0708, Remote Desktop Services, (Complete, Complete, Complete), 1.00}\}$

$V_2 = \{\text{CVE 2018 - 17706, Foxit PhantomPDF, (Partial, Partial, Partial), 0.68}\}$

$V_3 = \{\text{CVE 2016 - 3207, Internet Explorer 11, (Complete, Complete, Complete), 0.76}\}$

$V_4 = \{\text{CVE 2017 - 16381, Adobe Acrobat and Reader, (Complete, Complete, Complete), 0.93}\}$

$V_5 = \{\text{CVE 2008 - 0166, OpenSSL 0.9.8c - 1, (Complete, None, None), 0.78}\}$

$V_6 = \{\text{CVE 2016 - 7407, Dropbear SSH, (Complete, Complete, Complete), 1.00}\}$

$V_7 = \{\text{CVE 2007 - 4752, SSH in OpenSSH before 4.7, (Partial, Partial, Partial), 0.7}\}$

$V_8 = \{\text{CVE 2017 - 11509, Firebird SQL Server version 3.0.2, (Complete, Complete, Complete), 0.90}\}$

$V_9 = \{\text{CVE 2019 - 11682, MailCarrier 2.51, (Partial, Partial, Partial), 0.75}\}$

$V_{10} = \{\text{CVE 2008 - 3060, V - webmail 1.5.0, (Partial, None, None), 0.50}\}$

$V_{11} = \{\text{CVE 2001 - 1030, Squid Proxy Server, (Partial, Partial, Partial), 0.75}\}$

$V_{12} = \{\text{CVE 2010 - 0290, ISC BIND 9.7.0, (None, Partial, Partial), 0.40}\}$

$V_{13} = \{\text{CVE 2017 - 7269, Internet Information Services (IIS) 6.0 in Microsoft Windows Server 2003 R2, (Complete, Complete, Complete), 1.00}\}$

The mapping between corresponding vulnerabilities for each asset is listed in Assets Vulnerabilities table, shown in Table 4.

Table 4. Assets Vulnerabilities Table

Asset	Vulnerability
A ₁	V ₁ , V ₂ , V ₃
A ₂	V ₄
A ₃	V ₅ , V ₆ , V ₇
A ₄	V ₈
A ₅	V ₉ , V ₁₀ , V ₁₁
A ₆	V ₁₂
A ₇	V ₁₃

5.2. Attack Graph Generation

Using information about assets vulnerabilities and their

connectivity, AG model of the network can be automatically generated using MulVAL network security analyzer [57]. The simple representation of attack graph of the test network is shown in Figure 4.

Attack graph shown in Figure 4 is a directed acyclic graph in which remote attacker's entry point is represented using plain text, security vulnerabilities are represented using ovals and possible attacker's goals are represented using dashed shapes. A directed edge represents a transition from one state to another.

As can be seen, there are several possible goals an attacker can choose to compromise. For simplicity, in this experiment, a scenario of compromising administrative server is considered as the attackers' goal.

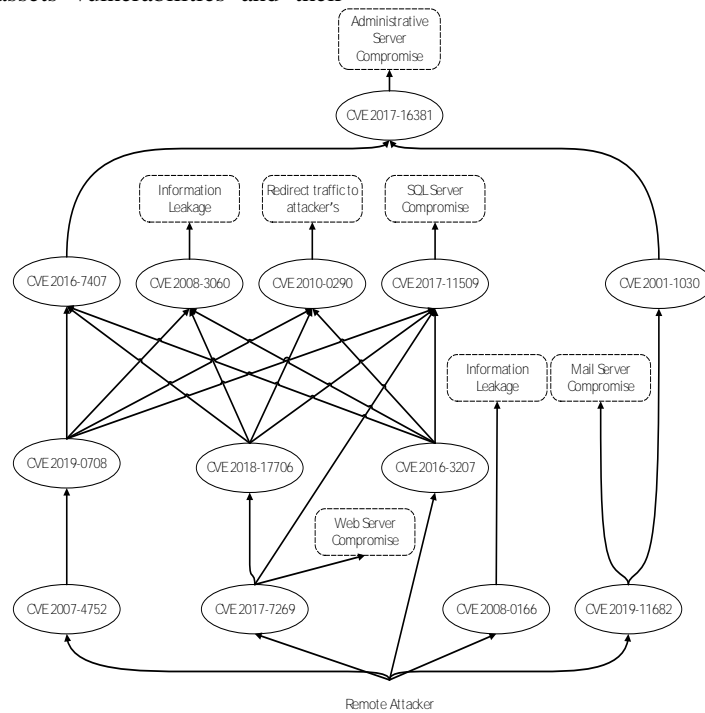


Figure 4. Attack Graph of the Test Network

Table 5. Attack Graph's States Importance

State ID	State Name	Importance
S1	CVE 2007-4752	0.1
S2	CVE 2017-7269	0.2
S3	CVE 2016-3207	0.42
S4	CVE 2008-0166	NA
S5	CVE 2019-11682	0.25
S6	CVE 2019-0708	0.33
S7	CVE 2018-17706	0.33
S8	CVE 2001-1030	0.25
S9	CVE 2016-7407	1.2
S10	CVE 2008-3060	NA
S11	CVE 2010-0290	NA
S12	CVE 2017-11509	NA
S13	CVE 2017-16381	1.5

5.3. Attack Graph Analysis

In order to identify risk hotspots in the generated AG model, exposure, path length, closeness centrality and betweenness centrality metrics should be extracted from its structure as discussed in Section 4.2.2. After assigning metrics values, AG states importance can be calculated using Equation 1. The results of analyzing test network's AG are represented in Table 5. Detailed information about metrics values are provided in Supplementary Table 1 in Appendix. As can be seen, states S_{13} , S_9 and S_3 have the highest importance among other nodes, respectively.

5.4. Bayesian Inference

In order to calculate UP of compromising network states, GeNIe Modeler [62] is used to apply Bayesian inference over the generated AG model. As the result of Bayesian inference, initial UP of compromising network states when no countermeasures is implemented are calculated. These probabilities are listed in second column of Table 6. As can be seen, states S_9 , S_2 and S_{12} have the highest UP among other nodes, meaning that they are the most probable stepping stones for attackers.

Table 6. Unconditional Probabilities of AG States

State	Initial UP	UP After S_2 Compromise
S1	0.528	0.747
S2	0.696	1.00
S3	0.535	0.757
S4	0.549	0.777
S5	0.528	0.747
S6	0.527	0.742
S7	0.476	0.680
S8	0.401	0.563
S9	0.702	0.978
S10	0.531	0.744
S11	0.461	0.646
S12	0.687	0.959
S13	0.405	0.570

5.5. Vulnerability Coverage Assessment

In this process, a Security Countermeasure Vector is defined for each SC covering identified vulnerabilities. List of SCs can be acquired from online repositories, reports, documents and knowledge of security administrator. Vectors of SCs covering the test network vulnerabilities are listed as follow. The definition of Security Countermeasure Vector is explained in Section 3.

$SC_1 = \{\text{Filtering external traffics}, (V_1, V_9, V_{13}),$
 $IR: (Partial, Partial, Partial), PR: 0.60, I: Partial, IC: 70\}$

$SC_2 = \{\text{Apply MS workaround}, (V_3),$
 $IR: (Complete, Complete, Complete), PR: 0.65,$
 $I: None, IC: 30\}$

$SC_3 = \{\text{Disable WebDAV}, (V_{13}),$
 $IR: (Complete, Complete, Complete), PR: 0.95,$

$I: Complete, IC: 120\}$

$SC_4 = \{\text{Patch OpenSSH}, (V_6, V_7),$
 $IR: (Complete, Complete, Complete), PR: 0.75,$
 $I: None, IC: 63\}$

$SC_5 = \{\text{Disable port scan}, (V_{11}),$
 $IR: (Partial, Partial, Partial), PR: 0.45,$
 $I: Partial, IC: 21\}$

$SC_6 = \{\text{Add network IDS}, (V_{10}, V_{11}),$
 $IR: (Partial, Partial, Partial), PR: 0.68,$
 $I: Partial, IC: 102\}$

$SC_7 = \{\text{Gateway firewall}, (V_2),$
 $IR: (Partial, Partial, Partial), PR: 0.43, I: Partial,$
 $IC: 105\}$

$SC_8 = \{\text{Query restriction}, (V_8),$
 $IR: (Partial, Partial, Partial), PR: 0.28, I: Partial, IC: 84\}$

$SC_9 =$
 $\{\text{Disable external UDF libraries from being loaded},$
 $(V_8), IR: (Partial, Partial, Partial), PR: 0.43, I: Partial,$
 $IC: 31\}$

$SC_{10} = \{\text{Upgrade firebird packages}, (V_8),$
 $IR: (Partial, Partial, Partial), PR: 0.65, I: None, IC: 45\}$

$SC_{11} = \{\text{Apply bind security update}, (V_{12}),$
 $IR: (Complete, Complete, Complete), PR: 0.56, I: None,$
 $IC: 34\}$

$SC_{12} = \{\text{Limit access to DNS server}, (V_{12}),$
 $IR: (Complete, Complete, Complete), PR: 0.8, I: Partial,$
 $IC: 53\}$

$SC_{13} = \{\text{Digital signature}, (V_{10}),$
 $IR: (None, Complete, None), PR: 0.3, I: None, IC: 33\}$

$SC_{14} = \{\text{Use POP3}, (V_{10}),$
 $IR: (None, None, Partial), PR: 0.25, I: None, IC: 153\}$

Having Vulnerability Vectors and Security Countermeasure Vectors, we can assess the efficacy of SCs coverage over vulnerabilities using Coverage Tables. The integrated Coverage Table of countermeasures over their covered vulnerabilities for each one of the C, I and A requirements is shown in Table 7. The detailed Coverage Table containing the values of C, I and A metrics of vulnerabilities and security countermeasures is provided in Supplementary Table 2 in Appendix.

5.6. Policy Conformance Assessment

Using Asset Vectors and Coverage Tables created in vulnerability coverage assessment process, Suitability Tables for security countermeasures can be created according to Section 4.3.2. The summarized Suitability Table of countermeasures based on assets policies is shown in Table 8. The detailed Suitability Table for each one of the C, I and A requirements is presented in Supplementary Table 3 in Appendix.

Table 7. Coverage Table for C, I and A Requirements

Security Countermeasure	Vulnerability	Confidentiality Coverage Level	Integrity Coverage Level	Availability Coverage Level
SC ₁	V ₁	Little coverage	Little coverage	Little coverage
	V ₉	Equal coverage	Equal coverage	Equal coverage
	V ₁₃	Little coverage	Little coverage	Little coverage
SC ₂	V ₃	Equal coverage	Equal coverage	Equal coverage
SC ₃	V ₁₃	Equal coverage	Equal coverage	Equal coverage
SC ₄	V ₆	Equal coverage	Equal coverage	Equal coverage
	V ₇	Extra coverage	Extra coverage	Extra coverage
SC ₅	V ₁₁	Equal coverage	Equal coverage	Equal coverage
SC ₆	V ₁₀	Equal coverage	Extra coverage	Extra coverage
	V ₁₁	Equal coverage	Equal coverage	Equal coverage
SC ₇	V ₂	Equal coverage	Equal coverage	Equal coverage
SC ₈	V ₈	Little coverage	Little coverage	Little coverage
SC ₉	V ₈	Little coverage	Little coverage	Little coverage
SC ₁₀	V ₈	Little coverage	Little coverage	Little coverage
SC ₁₁	V ₁₂	Extra coverage	Extra coverage	Extra coverage
SC ₁₂	V ₁₂	Extra coverage	Extra coverage	Extra coverage
SC ₁₃	V ₁₀	No coverage	Extra coverage	Equal coverage
SC ₁₄	V ₁₀	No coverage	Equal coverage	Extra coverage

Table 8. Suitability Values of Security Countermeasures Covering Asset's Vulnerabilities

Security Countermeasure	Vulnerability	Asset	Total Suitability
SC ₁	V ₁	A ₁	12
	V ₉	A ₅	19
	V ₁₃	A ₇	10
SC ₂	V ₃	A ₁	21
SC ₃	V ₁₃	A ₇	23
SC ₄	V ₆	A ₃	15
	V ₇	A ₃	12
SC ₅	V ₁₁	A ₅	19
SC ₆	V ₁₀	A ₅	17
	V ₁₁	A ₅	19
SC ₇	V ₂	A ₁	21
SC ₈	V ₈	A ₄	8
SC ₉	V ₈	A ₄	8
SC ₁₀	V ₈	A ₄	8
SC ₁₁	V ₁₂	A ₆	16
SC ₁₂	V ₁₂	A ₆	16
SC ₁₃	V ₁₀	A ₅	10
SC ₁₄	V ₁₀	A ₅	10

Table 9. Countermeasures Utility in Preventing $state_{Goal}$ Compromise

State in T	Covering SC	Influence	suitability	importance	intrusiveness	Utility
S2	SC1	1	10	0.2	0.7	2.86
	SC3	1.6	23	0.2	1	7.36
S7	SC7	0.7	21	0.33	0.7	6.93
S9	SC4	1.3	15	1.2	0.3	78
S13	NA	0.0	NA	1.5	NA	NA

5.7. Dynamic Cost-Benefit Analysis

Finally, by conducting a dynamic cost-benefit analysis, SRMP with the highest utility and total implementation cost lower than the allocated security hardening budget should be identified. As seen in Section 5.4, in case where no countermeasures is implemented, initial UP of compromising network states is calculated using Bayesian inference technique and the results are listed in the second column of Table 6. It can be seen that the UP of reaching $state_{Goal}$ by attackers, i.e. compromising Administrative Server, is equal to 0.405.

By an assumption that an IDS alert is generated representing state S_2 is compromised, the proposed dynamic cost-benefit analysis algorithm (algorithm 1) is employed to find the optimal SRMP. In this case, first of all, the probability of state S_2 is changed to 1 (line 2) and the UP of graph states are updated by applying Bayesian inference algorithm (line 3). The result is shown in the third column of Table 6.

Since $\Delta UP(state_{Goal})$ is equal to $0.570 - 0.405 = 0.165$ and is more than the predefined threshold, let's say 0.1, the procedure of countermeasure selection continues. In the next step, the set $T = \{S_2, S_7, S_9, S_{13}\}$ is created according to line 7. In lines 9 to 15, the influence of countermeasures in preventing attackers reaching $state_{Goal}$ is calculated.

In lines 17 to 21, the utility of countermeasures is calculated using Equation 2. SCs utility values and metrics values used to calculate them are shown in Table 9.

Finally, assuming that the allocated budget for system hardening is equal to 200 units, the output of the algorithm, according to the utility values of countermeasures, is $SRMP = \{SC_3, SC_4\}$ with the total implementation cost of 183 units.

6. Conclusions and Future Works

By increasing the volume and sophistication of today's cyber-attacks, the need for a method to identify the optimal set of security countermeasures is indispensable. This paper presents a dynamic security risk management framework to identify the optimal security risk mitigation plans considering vulnerabilities characteristics, countermeasures effectiveness, existing security policies and budget limitations. Systems risk hotspots are identified by conducting structural analysis of attack graph. By conducting probabilistic analysis of attack graph, the most probable stepping stones for attackers are determined. Countermeasures suitability are calculated according to their ability in covering vulnerabilities and assets security policies. Moreover, a dynamic cost-benefit analysis algorithm is proposed to identify the optimal security risk mitigation plans. Finally, the feasibility and applicability of the proposed framework is ensured using a case study. In future, we try to further extend the proposed framework by considering attackers' capabilities and intentions in bypassing countermeasures and exploiting vulnerabilities.

7. References

- [1] Ross, R., "Guide for conducting risk assessments NIST special publication 800-30 revision 1", US Dept. Commerce, NIST, Gaithersburg, MD, USA, Tech. Rep, 2012.
- [2] Wheeler, E., Security risk management: Building an information security risk management program from the Ground Up. Elsevier, 2011.
- [3] Kuzminykh, I., Ghita, B., Sokolov, V., and Bakhshi, T., "Information security risk assessment", *Encyclopedia*, Vol. 1, No. 3, pp. 602–617, 2021.
- [4] Shameli-Sendi, A., Cheriet, M., and Hamou-Lhadj, A., "Taxonomy of intrusion risk assessment and response system", *Computers & Security*, Vol. 45, pp. 1–16, 2014.
- [5] Shameli-Sendi, A., Aghababaei-Barzegar, R., and Cheriet, M., "Taxonomy of information security risk assessment (ISRA)", *Computers & security*, Vol. 57, pp. 14–30, 2016.
- [6] Erdogan, G., and Refsdal, A., "A method for developing qualitative security risk assessment algorithms", in *International Conference on Risks and Security of Internet and Systems*, pp. 244–259, Springer, 2017.
- [7] Dobaj, J., Schmittner, C., Krisper, M., and Macher, G., "Towards integrated quantitative security and safety risk assessment", in *International Conference on Computer Safety, Reliability, and Security*, pp. 102–116, Springer, 2019.
- [8] Khosravi-Farmad, M., Rezaee, R., Harati, A., and Bafghi, A. G., "Network security risk mitigation using Bayesian decision networks", in *2014 4th International Conference on Computer and Knowledge Engineering (ICCCKE)*, pp. 267–272, IEEE, 2014.
- [9] Wang, J., Neil, M., and Fenton, N., "A bayesian network approach for cybersecurity risk assessment implementing and extending the fair model", *Computers & Security*, Vol. 89, pp. 101659, 2020.
- [10] Hulitt, E., and Vaughn, R. B., "Information system security compliance to FISMA standard: a quantitative measure", *Telecommunication Systems*, Vol. 45, No. 2, pp. 139–152, 2010.
- [11] Lo, C.-C., and Chen, W.-J., "A hybrid information security risk assessment procedure considering interdependences between controls", *Expert Systems with Applications*, Vol. 39, No. 1, pp. 247–257, 2012.
- [12] Figueira, P. T., Bravo, C. L., and López, J. L. R., "Improving information security risk analysis by including threat-occurrence predictive models", *Computers & Security*, Vol. 88, pp. 101609, 2020.
- [13] CVSS, "Common vulnerability scoring system v3.0: Specification document".
- [14] FIRST, "Forum of incident response and security teams". <https://www.first.org/>.
- [15] Khosravi-Farmad, M., Ramaki, A. A., and Bafghi, A. G., "Moving target defense against advanced persistent threats for cybersecurity enhancement", in *2018 8th International Conference on Computer and Knowledge Engineering (ICCCKE)*, pp. 280–285, IEEE, 2018.
- [16] Ouassini, A., and Hunter, M., "Advanced Persistent Threats (APTs)", *The Handbook of Homeland Security*, CRC Press, pp. 163–165, 2023.
- [17] Chen, Z., Liu, J., Shen, Y., Simsek, M., Kantarci, B., Mouftah, H. T., and Djukic, P., "Machine learning-enabled IoT security: Open issues and challenges under advanced persistent threats", *ACM Computing Surveys*, Vol. 55, No. 5, pp. 1–37, 2022.
- [18] Hong, J. B., Kim, D. S., Chung, C.-J., and Huang, D.,

- "A survey on the usability and practical applications of graphical security models", *Computer Science Review*, Vol. 26, pp. 1–16, 2017.
- [19] Kaynar, K., "A taxonomy for attack graph generation and usage in network security", *Journal of Information Security and Applications*, Vol. 29, pp. 27–56, 2016.
- [20] Lallie, H. S., Debattista, K., and Bal, J., "A review of attack graph and attack tree visual syntax in cyber security", *Computer Science Review*, Vol. 35, pp. 100219, 2020.
- [21] Shameli-Sendi, A., and Dagenais, M., "Arito: Cyber-attack response system using accurate risk impact tolerance", *International journal of information security*, Vol. 13, No. 4, pp. 367–390, 2014.
- [22] Zahid, M., Inayat, I., Daneva, M., and Mehmood, Z., "A security risk mitigation framework for cyber physical systems", *Journal of software: Evolution and Process*, Vol. 32, No. 2, pp. e2219, 2020.
- [23] Li, S., Tryfonas, T., Russell, G., and Andriotis, P., "Risk assessment for mobile systems through a multilayered hierarchical bayesian network", *IEEE transactions on cybernetics*, Vol. 46, No. 8, pp. 1749–1759, 2016.
- [24] Shameli-Sendi, A., Louafi, H., He, W., and Cheriet, M., "Dynamic optimal countermeasure selection for intrusion response system", *IEEE Transactions on Dependable and Secure Computing*, Vol. 15, No. 5, pp. 755–770, 2016.
- [25] Li, S., Zhao, S., Yuan, Y., Sun, Q., and Zhang, K., "Dynamic security risk evaluation via hybrid bayesian risk graph in cyber-physical social systems", *IEEE Transactions on Computational Social Systems*, Vol. 5, No. 4, pp. 1133–1141, 2018.
- [26] He, W., Li, H., and Li, J., "Unknown vulnerability risk assessment based on directed graph models: a survey", *IEEE Access*, Vol. 7, pp. 168201–168225, 2019.
- [27] Garg, U., Sikka, G., and Awasthi, L. K., "Empirical analysis of attack graphs for mitigating critical paths and vulnerabilities", *Computers & Security*, Vol. 77, pp. 349–359, 2018.
- [28] Hermanowski, D., and Piotrowski, R., "Network risk assessment based on attack graphs", in *International Conference on Dependability and Complex Systems*, pp. 156–167, Springer, 2021.
- [29] Rezaee, R., and Ghaemi Bafghi, A., "A risk estimation framework for security threats in computer networks", *Journal of Computing and Security*, Vol. 7, No. 1, pp. 19–33, 2020.
- [30] Rezaee, R., Bafghi, A. G., and Khosravi-Farmad, M., "A threat risk estimation model for computer network security", in *2016 6th International Conference on Computer and Knowledge Engineering (ICCKE)*, pp. 223–228, IEEE, 2016.
- [31] Presekal, A., Ștefanov, A., Rajkumar, V. S., and Palensky, P., "Attack graph model for cyber-physical power systems using hybrid deep learning", *IEEE Transactions on Smart Grid*, 2023.
- [32] Liu, Y., and Man, H., "Network vulnerability assessment using bayesian networks", in *Data mining, intrusion detection, information assurance, and data networks security 2005*, Vol. 5812, pp. 61–71, International Society for Optics and Photonics, 2005.
- [33] Frigault, M., and Wang, L., "Measuring network security using bayesian network-based attack graphs", in *2008 32nd Annual IEEE International Computer Software and Applications Conference*, pp. 698–703, IEEE, 2008.
- [34] Poolsappasit, N., Dewri, R., and Ray, I., "Dynamic security risk management using bayesian attack graphs", *IEEE Transactions on Dependable and Secure Computing*, Vol. 9, No. 1, pp. 61–74, 2011.
- [35] Feng, N., Wang, H. J., and Li, M., "A security risk analysis model for information systems: Causal relationships of risk factors and vulnerability propagation analysis", *Information sciences*, Vol. 256, pp. 57–73, 2014.
- [36] Le, A., Chen, Y., Chai, K. K., Vasenev, A., and Montoya, L., "Incorporating fair into bayesian network for numerical assessment of loss event frequencies of smart grid cyber threats", *Mobile Networks and Applications*, Vol. 24, No. 5, pp. 1713–1721, 2019.
- [37] Al-Hadhrani, N., Collinson, M., and Oren, N., "A subjective network approach for cybersecurity risk assessment", in *13th International Conference on Security of Information and Networks*, pp. 1–8, 2020.
- [38] Ramaki, A. A., Khosravi-Farmad, M., and Bafghi, A. G., "Real time alert correlation and prediction using Bayesian networks", in *2015 12th International Iranian Society of Cryptology Conference on Information Security and Cryptology (ISCISC)*, pp. 98–103, IEEE, 2015.
- [39] Chen, Y. Y., Xu, B., and Long, B., "Information security assessment of wireless sensor networks based on bayesian attack graphs", *Journal of Intelligent & Fuzzy Systems*, Vol. 41, No. 3, pp. 4511–4517, 2021.
- [40] Meyur, R., "A bayesian attack tree based approach to assess cyber-physical security of power system", in *2020 IEEE Texas Power and Energy Conference (TPEC)*, pp. 1–6, IEEE, 2020.
- [41] Khosravi-Farmad, M., Ramaki, A. A., and Bafghi, A. G., "Risk-based intrusion response management in ids using bayesian decision networks", in *2015 5th International Conference on Computer and Knowledge Engineering (ICCKE)*, pp. 307–312, IEEE, 2015.
- [42] Behbehani, D., Komninos, N., Al-Begain, K., and Rajarajan, M., "Cloud enterprise dynamic risk assessment (CEDRA): a dynamic risk assessment using dynamic bayesian networks for cloud environment", *Journal of Cloud Computing*, Vol. 12, No. 1, 2023.
- [43] Nespoli, P., Papamartzivanos, D., Mármol, F. G., and Kambourakis, G., "Optimal countermeasures selection against cyber attacks: A comprehensive survey on reaction frameworks", *IEEE Communications Surveys & Tutorials*, Vol. 20, No. 2, pp. 1361–1396, 2017.
- [44] Noel, S., Jajodia, S., O'Berry, B., and Jacobs, M., "Efficient minimum-cost network hardening via exploit dependency graphs", in *19th Annual Computer Security Applications Conference*, 2003. Proceedings., pp. 86–95, IEEE, 2003.
- [45] Jha, S., Sheyner, O., and Wing, J., "Two formal analyses of attack graphs", in *Proceedings 15th IEEE Computer Security Foundations Workshop*. CSFW-15,

- pp. 49–63, IEEE, 2002.
- [46] Dewri, R., Poolsappasit, N., Ray, I., and Whitley, D., "Optimal security hardening using multi-objective optimization on attack tree models of networks", in *Proceedings of the 14th ACM conference on Computer and communications security*, pp. 204–213, 2007.
- [47] Khosravi-Farmad, M., and Ghaemi-Bafghi, A., "Bayesian decision network-based security risk management framework", *Journal of Network and Systems Management*, Vol. 28, No. 4, pp. 1794–1819, 2020.
- [48] Chung, C.-J., Khatkar, P., Xing, T., Lee, J., and Huang, D., "Nice: Network intrusion detection and countermeasure selection in virtual network systems", *IEEE transactions on dependable and secure computing*, Vol. 10, No. 4, pp. 198–211, 2013.
- [49] Schilling, A., and Werners, B., "Optimal selection of it security safeguards from an existing knowledge base", *European Journal of Operational Research*, Vol. 248, No. 1, pp. 318–327, 2016.
- [50] Kotenko, I., and Doynikova, E., "Selection of countermeasures against network attacks based on dynamical calculation of security metrics", *The Journal of Defense Modeling and Simulation*, Vol. 15, No. 2, pp. 181–204, 2018.
- [51] Nessus, "Nessus vulnerability scanner", Available on, <https://www.tenable.com/products/nessus>.
- [52] OpenVAS, "Open vulnerability assessment scanner", Available on, <http://www.openvas.org/>.
- [53] Retina, "Retina network security vulnerability scanner", Available on, <https://www.beyondtrust.com/products/retinanetwork-security-scanner/>.
- [54] NVD, "NIST US national vulnerability database (NVD)", Available on, <https://nvd.nist.gov/>.
- [55] CVE, "Common vulnerabilities and exposures (CVE)", Available on, <https://cve.mitre.org/>.
- [56] Nmap, "Nmap, the network mapper", Available on, <https://nmap.org/>.
- [57] Ou, X., Govindavajhala, S., Appel, A. W., et al., "Mulval: A logic-based network security analyzer", in *USENIX security symposium*, Vol. 8, pp. 113–128, Baltimore, MD, 2005.
- [58] Jajodia, S., and Noel, S., "Topological vulnerability analysis", in *Cyber situational awareness*, pp. 139–154, Springer, 2010.
- [59] Russell, S., and Norvig, P., "Artificial intelligence: A modern approach, global edition 4th", Foundations, Vol. 19, pp. 23, 2021.
- [60] Khosravi-Farmad, M., Rezaee, R., and Bafghi, A. G., "Considering temporal and environmental characteristics of vulnerabilities in network security risk assessment", in *2014 11th International ISC Conference on Information Security and Cryptology*, pp. 186–191, IEEE, 2014.
- [61] Koller, D., and Friedman, N., Probabilistic graphical models: principles and techniques. MIT press, 2009.
- [62] GeNIe, "GeNIe modeler, bayesfusion, llc", Available on, <https://www.bayesfusion.com/>.

8. Appendix

Supplementary Table 1 represents metrics used for states importance calculation. Detailed coverage table for C, I and A requirements is presented by Supplementary Table 2. Supplementary Table 3 represents the Suitability Table metrics for C, I and A requirements.

Supplementary Table 1. Metrics for Calculating States Importance

State	Exposure (E)	Path Length (PL)	Closeness Centrality (CC)	Betweenness Centrality (BC)	Importance ($I = E \times BC / PL \times CC$)
S1	2	5	4	1	0.1
S2	4	5	4	1	0.2
S3	5	4	3	1	0.42
S4	2	NA	NA	NA	NA
S5	3	4	3	1	0.25
S6	5	5	3	1	0.33
S7	5	5	3	1	0.33
S8	2	4	2	1	0.25
S9	4	5	2	3	1.2
S10	4	NA	NA	NA	NA
S11	4	NA	NA	NA	NA
S12	5	NA	NA	NA	NA
S13	3	4	1	2	1.5

Supplementary Table 2. Detailed Coverage Table for C, I and A Requirements

Security Countermeasure				Vulnerability				Coverage Level		
ID	C	I	A	ID	C	I	A	C	I	A
SC ₁	P	P	P	V ₁	C	C	C	Little	Little	Little
				V ₉	P	P	P	Equal	Equal	Equal
				V ₁₃	C	C	C	Little	Little	Little
SC ₂	C	C	C	V ₃	C	C	C	Equal	Equal	Equal
SC ₃	C	C	C	V ₁₃	C	C	C	Equal	Equal	Equal
SC ₄	C	C	C	V ₆	C	C	C	Equal	Equal	Equal
				V ₇	P	P	P	Extra	Extra	Extra
SC ₅	P	P	P	V ₁₁	P	P	P	Equal	Equal	Equal
SC ₆	P	P	P	V ₁₀	P	N	N	Equal	Extra	Extra
				V ₁₁	P	P	P	Equal	Equal	Equal
SC ₇	P	P	P	V ₂	P	P	P	Equal	Equal	Equal
SC ₈	P	P	P	V ₈	C	C	C	Little	Little	Little
SC ₉	P	P	P	V ₈	C	C	C	Little	Little	Little
SC ₁₀	P	P	P	V ₈	C	C	C	Little	Little	Little
SC ₁₁	C	C	C	V ₁₂	N	P	P	Extra	Extra	Extra
SC ₁₂	C	C	C	V ₁₂	N	P	P	Extra	Extra	Extra
SC ₁₃	N	C	N	V ₁₀	P	N	N	No	Extra	Equal
SC ₁₄	N	N	P	V ₁₀	P	N	N	No	Equal	Extra

Supplementary Table 3. Suitability Table Metrics for C, I and A Requirements

Security Countermeasure	Vulnerability	Coverage Level			Asset Policies				Suitability			
ID	ID	C	I	A	ID	C	I	A	C	I	A	Total
SC ₁	V ₁	Little	Little	Little	A ₁	P	P	P	4	4	4	12
	V ₉	Equal	Equal	Equal	A ₅	C	N	P	9	3	7	19
	V ₁₃	Little	Little	Little	A ₇	P	P	C	4	4	2	10
SC ₂	V ₃	Equal	Equal	Equal	A ₁	P	P	P	7	7	7	21
SC ₃	V ₁₃	Equal	Equal	Equal	A ₇	P	P	C	7	7	9	23
SC ₄	V ₆	Equal	Equal	Equal	A ₃	N	N	C	3	3	9	15
	V ₇	Extra	Extra	Extra	A ₃	N	N	C	2	2	8	12
SC ₅	V ₁₁	Equal	Equal	Equal	A ₅	C	N	P	9	3	7	19
SC ₆	V ₁₀	Equal	Extra	Extra	A ₅	C	N	P	9	2	6	17
	V ₁₁	Equal	Equal	Equal	A ₅	C	N	P	9	3	7	19
SC ₇	V ₂	Equal	Equal	Equal	A ₁	P	P	P	7	7	7	21
SC ₈	V ₈	Little	Little	Little	A ₄	C	C	P	2	2	4	8
SC ₉	V ₈	Little	Little	Little	A ₄	C	C	P	2	2	4	8
SC ₁₀	V ₈	Little	Little	Little	A ₄	C	C	P	2	2	4	8
SC ₁₁	V ₁₂	Extra	Extra	Extra	A ₆	N	P	C	2	6	8	16
SC ₁₂	V ₁₂	Extra	Extra	Extra	A ₆	N	P	C	2	6	8	16
SC ₁₃	V ₁₀	No	Extra	Equal	A ₅	C	N	P	1	2	7	10
SC ₁₄	V ₁₀	No	Equal	Extra	A ₅	C	N	P	1	3	6	10



Ferdowsi
University of
Mashhad

Journal of Computer and Knowledge Engineering

<https://cke.um.ac.ir>



Information and
Communication Technology
Association of Iran

A Lightweight Secure Scheme for Data Aggregation in Large-Scale IoT-Based Smart Grids

Research Article

Mohammad J. Abdolmaleki¹, Amanj Khorramian², Mohammad Fathi³

DOI: [10.22067/cke.2024.82734.1080](https://doi.org/10.22067/cke.2024.82734.1080)

Abstract. With the emergence of IoT devices, data aggregation in the area of smart grids can be implemented based on IoT networks. However, the communication and computation resources of IoT devices are limited so it is not possible to apply conventional Internet protocols directly. On the other hand, gathering data from smart meters in the advanced metering infrastructure faces challenges such as privacy-preserving and heavy-loaded authentication and aggregation schemes. In this paper, we propose an improved lightweight, secure, and privacy-preserving scheme for aggregating data of smart meters in large-scale IoT-based smart grids. The proposed scheme adopts lightweight operations of cryptography such as exclusive-OR, hash, and concatenation functions. In comparison with the schemes in the literature, the analysis and simulation results show that the proposed scheme satisfies the same security levels, while at the same time burdens lower computation and communication overheads. This observation makes the proposed scheme more suitable to be employed in large-scale and IoT-based smart grids for data aggregation.

Keywords: Index Terms— internet of things, smart grids, large-scale networks, light-weight security, data aggregation.

1. Introduction

The Internet of Things (IoT) is a set of smart devices communicating with each other through wired or wireless channels in order to achieve a specific goal [1]. IoT consists of four general environments as Internet of Vehicles (IoV), Internet of Sensors (IoS) or wireless sensor networks (WSN), machine-to-machine (M2M) communications, and Internet of energy (IoE) which is called smart grid (SG) [2].

The growth in population faces the world with challenges of supplying the energy needed by industrial units, offices, houses, etc. Moreover, environmental issues arise due to excessive consumption of fossil fuels. Therefore, it sounds necessary to supply some part of the energy demand from other resources. In order to mitigate these challenges, a smart network is needed to coordinate both suppliers and consumers. The aim of SG is to achieve a trade-off between the supply and demand of electrical energy and also to optimize the consumed energy through power and demand-side management programs.

Using digital information and communication infrastructures, SG establishes a platform to integrate consumers and energy sources such as renewable energies

and power plants. Using the aforementioned infrastructure, SG can also provide two-way communication between the power supplier and consumers. In other words, unlike traditional distribution systems, the power supplier can also send data to the consumer. This feature can sometimes cause the consumer to act as a supplier and generate electricity for other consumers. This happens when the supplied energy is more than the energy needed for a consumer [3].

The structure of SG consists of several operators and devices including maintenance personnel, security officers, advanced metering infrastructure (AMI), data aggregator (DA), and intelligent electronic devices. A typical structure of SG data aggregation is shown in Figure 1. Here, AMI's task is to measure and report the amount of consumed energy by each customer smartly using a smart meter (SM). The reported data of SMs is then collected by a trusted third-party DA and then forwarded to the power supplier (PS). Finally, PS generates the power according to the reports of the SMs. This cycle occurs periodically in order to monitor, control, and predict the amount of power consumption, which respectively eventuates cost reduction for both the PS and customers [4].

Data aggregation in SGs faces a number of challenges. One of the most important challenges is posed by the limited processing and communication resources of SMs. In order to address this challenge, protocols of the SG infrastructure have to be designed computationally and communicationally lightweight while at the same time providing the security of the connections [5]. Another challenge is to preserve the privacy of costumers, for which the measured data is transmitted throughout the network. In this paper, a lightweighted secure scheme for data aggregation in SG is proposed.

A. Related Work

So far, different schemes have been proposed for data aggregation in various networks. In 2018, Zhang et al. [6] proposed a light-weight privacy-preserving data aggregation for resource-constraint edge terminals and edge computing systems that uses online/offline digital

* Manuscript received: 2023 June1, Revised, 2023October7, Accepted, 2023 November7.

¹ MSc Student, Department of Electrical Engineering, University of Kurdistan, Sanandaj, Iran.

² Assistant Professor, Department of Computer Engineering and Information Technology, University of Kurdistan, Sanandaj, Iran.

³ Corresponding author. Associate Professor, Department of Electrical Engineering, University of Kurdistan, Sanandaj, Iran

Email: mfathi@uok.ac.ir.

signature, Paillier homomorphic cryptosystem, and double trapdoor Chameleon hash function. Their scheme provides data confidentiality and keeps the privacy of the network users from the control center and also edge server. Shen et al. [7] proposed a privacy-preserving cube-data aggregation scheme for electricity consumption in smart grids. The proposed scheme is based on Horner's Rule and Paillier cryptosystem. In 2018, Zheng He et al. [8] presented a privacy-preserving multi-functional data aggregation without a trusted third party in smart grids. They have used Paillier homomorphic cryptography to design their scheme. Likewise, Lu et al. [9] proposed a privacy-preserving data aggregation protocol that uses the Paillier homomorphic cryptosystem which also causes higher computation overhead. Li et al. [10] proposed a privacy-preserving multi-subset scheme for data aggregation in smart grids. They have used the Paillier homomorphic cryptosystem in order to prevent the access of trusted third-party aggregators to the private information of consumers. In 2019, Chen et al. [11] proposed a scheme for aggregating the data of power consumed by an SM which uses the Paillier homomorphic cryptosystem. This scheme enables the power supplier to achieve the whole data consumption of SMs, while it has no access to the data of individual SMs.

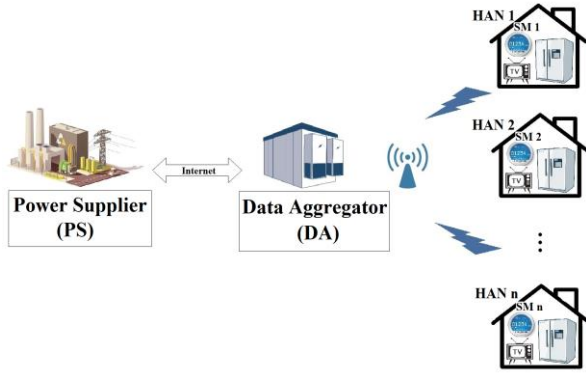


Figure 1. Network model of SG data aggregation

Jo et al. [12] have proposed efficient and privacy-preserving metering protocols for SGs. Their proposed protocols are based on bilinear mapping, hard problem, and the Paillier homomorphic encryption. In 2016, Abdallah et al. [13] presented a lightweight privacy-preserving electricity consumption aggregation scheme that employs a lightweight lattice-based homomorphic cryptosystem. In their proposed scheme, smart domestic facilities aggregate their readings without involving the SM. In [14], a data aggregation scheme is designed based on the discrete logarithm problem, in which a substation has access to the private data of consumers.

In [15], Jo et al. presented two data aggregation schemes based on Paillier homomorphic encryption and elliptic curve digital signature algorithms (ECDSA), respectively. Their schemes cannot warrant the anonymity of the consumers and are also computationally heavy. In 2017, Vahedi et al. [16] designed a privacy-preserving data aggregation scheme for smart grids based on ECDSA. Their scheme cannot assure the privacy of the customers. In 2018, Liu et al. [17] presented a practical privacy-preserving method for aggregating data that exploits EC-

ElGamal to encrypt the data of SM. In their approach, although there is no trusted third party to aggregate the data, the control center is assumed to be honest but curious.

He et al. [18] proposed an efficient and privacy-preserving data aggregation scheme for SGs against internal adversaries. They have proposed the scheme using the Boneh-Goh-Nissim public key cryptography method. Unlike most proposed schemes for data aggregation, their proposed scheme does not use bilinear pairing. In 2015, Abdallah et al. [19] presented a lightweight security and privacy-preserving scheme for customer-side networks which uses n th degree truncated polynomial ring units (NTRU) cryptosystem. Their scheme is based on forecasting the electricity demand for a cluster of houses in the same residential areas. Mustafa et al. [20] have proposed a secure and privacy-preserving protocol for smart metering operational data collection. In their proposed protocol, power suppliers and grid operators are allowed to collect the consumption data of SMs securely while SMs' privacy is protected. Their protocol uses Multiparty Computations as the underlying cryptographic primitive. In 2016, Knirsch et al. [21] proposed an approach for privacy-preserving data aggregation based on symmetric cryptographic systems and advanced encryption standard (AES). Their proposed scheme prevents error occurrence in the aggregation process. Gope et al. [22] in 2018 proposed a lightweight and privacy-friendly spatial data aggregation for secure power supply and demand management in smart grids. Their scheme uses only light-weight cryptographic primitives like hash function, XOR, etc.

Recent work has also focused on secure aggregation protocols optimized for federated learning systems. So et al. [25] proposed LightSecAgg, a lightweight and versatile secure aggregation protocol for federated learning that encodes/decodes aggregate mask values instead of dropped user masks. This reduces overhead with increasing dropped users compared to prior art. Experiments showed significantly reduced training time for diverse models and datasets. In other work, Zhang et al. [26] proposed a lightweight multi-dimensional data aggregation scheme for IoT using the Chinese Remainder Theorem and simple additive encryption. It has lower communication and storage than homomorphic encryption schemes while still ensuring security properties like confidentiality, integrity, and robustness. Additional relevant work includes the two aggregation schemes by Qian et al. [27] using lightweight homomorphic encryption resilient to quantum attacks, and the edge-assisted aggregation scheme by Wu et al. [28] using Paillier encryption and virtual name-based verification to reduce overhead.

B. Preliminaries

PRNG: Pseudo-random number generator is a function which its algorithm generates a sequence of numbers with attributes like really random numbers. Its algorithm starts generating numbers with an initial value called the PRNG seed. By having a seed, the PRNG creates a sequence of numbers that is unique to the seed and can be generated one by one or even a whole sequence at a time. Furthermore, numbers generated in a one-by-one

procedure are the same as those generated in a sequence with the same specified order.

C. Contribution

In this paper, firstly, a procedure to achieve mutual authentication between PS, DA, and SMs is presented. Then a data aggregation scheme is proposed to aggregate the data of power consumption of the SMs periodically. In this regard, lightweight cryptographic operations such as Exclusive-OR (XOR), PRNG, one-way non-collision hash, and concatenation functions are employed. The motivation behind the proposed scheme in this paper is to address the vulnerability in the similar scheme proposed in [22]. The scheme in [22] employs PRNG to generate a sequence of random numbers with a given seed in PS and then transmit these numbers via the network to the SMs. If a sequence of random numbers is sniffed in the network by an attacker, then it would be possible to detect the adopted seed and determine the next numbers. This certainly results in information disclosure. Besides, transmitting the numbers to SMs burdens high computation and communication overhead in the network.

To address the aforementioned vulnerability in [22], in this paper, the same random numbers are generated simultaneously in the PS and individual SMs. In other words, rather than transmitting random numbers sequentially in the network, these numbers are generated with the same seeds at both PS and SMs. Therefore, having this approach causes significant improvements in security regards due to the inaccessibility of the random numbers and the seed by an intruder. Moreover, computation and communication overheads are highly decreased due to the generation of random numbers at both sides of the network.

In the proposed scheme, the personal information of each consumer is private and unachievable by the others. Nevertheless, the SG is able to monitor the summation of the consumed power in the network. The security of the proposed scheme is evaluated in terms of authentication, secure key establishment, data confidentiality, data integrity, and consumer privacy. Finally, the communication and computation overhead of this scheme is evaluated and compared with those in the literature.

In this work, rather than transmitting random numbers from the power supplier to smart meters, we generate matching random numbers simultaneously at both entities using pre-shared seeds. This enhances security by eliminating the transmission of random numbers that could be intercepted. Unlike the prior scheme that transmits pseudorandom numbers over the network, the proposed one generates matching random numbers at the supplier and meter sides using shared seeds. The scheme requires fewer hash function evaluations, encryptions/decryptions, and concatenations compared to the existing protocols in the literature. Our scheme is suitable for resource-constrained IoT devices and can support large-scale smart grid networks.

2. System Model

In this section, the network and the opponent models are explained. The notations and cryptographic functions which are used throughout this paper are shown in Table I.

A. Network Model

The network model considered in this paper for data aggregation consists of PS, DA, home area networks (HAN), and SMs, as shown in Figure 1. The PS distributes the electricity to all HANs. The DA's responsibility is to aggregate the value of power consumption of each SM periodically and then forward it to the PS. The PS takes advantage of reported data to give suitable feedback to the power-generating stations and suppliers. This feedback helps suppliers to make a balance between demand and supply, i.e., how much energy should be generated and distributed over the network. Any fault in reporting the data to the PS may cause wasted energy or lack of supply. Therefore, DA as a trusted third party has a vital role in sustaining the balance between the supply and demand of energy. Each HAN includes a number of electric devices handled by an SM. The SM measures the amount of electricity consumed by the devices and then reports it to the DA using a local network (e.g. WLAN or a cellular network). The PS and DA are connected together using the public Internet (ADSL, cellular network, and so on).

Table 1 Notations and Cryptographic Operations in LWS-DA

Symbols	Definition
SM	Smart Meter
DA	Data Aggregator
PS	Power Supplier
HAN	Home Area Network
s_i	Seed of smart meter SM_i
ID_{SM_i}	Real identity of smart meter SM_i
ID_A	Identity of data aggregator
PID_i	Pseudo identity of SM_i
TID_i	Temporary Identity of SM_i
k_i	Secret key between PS and SM_i
K_{as}	Secret pre-shared key between DA and PS
kh_i	Shared key between SM_i and DA
$E_k[x]$	Plaintext x encrypted using key k
h	Hashing function
\oplus	Bitwise Exclusive-OR (XOR)
\parallel	Concatenation

B. Opponent Model

It is assumed that the PS is an honest partner as it is owned by the government. Moreover, the DA may be owned by a private company to cooperate with the PS. Hence, the DA is assumed as a truthful but curious entity that may be tempted to acquire the data of power consumption by each HAN in the motivation of selling the usage information to other companies. Also, it is assumed that any element on the network connections between DA and PS may behave as an adversary and attempt to perceive the private information of each HAN. Furthermore, there may be an SM that is interested in accessing the data consumption of other SMs from other HANs. Attacks from outside of network are also likely and assumed. For instance, an illegal user may impersonate

itself as legal entity such as SM or even DA.

3. Proposed Scheme

To mitigate the challenges of privacy-preserving and high computation and communication overheads of data aggregation schemes in the literature, here, in this section, we propose a light-weighted secure data aggregation (LWS-DA) scheme for SGs. The proposed scheme is illustrated in two phases: the *authentication phase* and the *data aggregation phase*. A set of n SMs, indexed by i , are considered which are supplied by the PS. In the authentication phase, the PS ascertains the identity of each SM _{i} and DA. Then it proves its legality to both of them, using an encryption key kh_i and a set of temporary identities that can be updated and generated between the SM and the DA.

In the data aggregation phase, the DA collects the aggregate consumption of HANs, without knowing the exact power consumption of each HAN, as explained in subsection B.

In the proposed LWS-DA scheme, a time period consisting of one authentication interval followed by m data aggregation timeintervals are assumed. In other words, for n SMs in a period, one authentication process is done and then the data aggregation process is done for m time intervals. After the m th round of data aggregation, a new period begins. Therefore, the authentication process must be done in order to continue the data aggregation of the next rounds.

A. Authentication Phase

In this phase, we adopt the authentication method used in [22] with a number of modifications, shown in Figure 2. The modifications, shown in red-dashed rectangles, are to define and forward parameters required in the aggregation phase. The following are steps done in the Authentication phase.

Step AUTH1: Each SM _{i} , with the pseudo-identity PID_i and pre-shared secret key k_i with the PS, generates a random number N_s as its nonce. It then computes the hash-output $V_0 = h(PID_i \parallel N_s \parallel k_i)$ by concatenating its pseudo-identity, nonce, and its pre-shared secret key, where $h(\cdot)$ is a hash function. After that, it creates the message $M_{AU1} : \{PID_i, N_s, V_0\}$ and sends it to the DA.

Step AUTH2: On receiving the messages M_{AU1} from each SM, the DA generates its own nonce N_a . Similar to SMs, DA concatenates its real identity ID_A , the nonce N_a , and its pre-shared secret key K_{as} with PS to compute the hash-output $V_1 = h(ID_A \parallel N_a \parallel K_{as})$. Then, the DA sends message $M_{AU2} : \{M_{AU1} \parallel (ID_A, N_a, V_1)\}$ to the PS.

Step AUTH3: When PS receives the message M_{AU2} , firstly, it calculates two has-outputs V_0 and V_1 , and then compares the values of PID_i , V_0 and V_1 of itself with those that are achieved from message M_{AU2} . If all three values in the comparison are matched, then the PS generates PID_i^{new} , which is a new pseudo-identity for each SM _{i} . Thereafter, PS calculates the set of seeds $s_i = PID_i^{new} \oplus k_i$ for SMs in order to generate random numbers in the aggregation phase to the encrypt energy consumption of individual SMs. Then, PS computes parameters $T = h(ID_{SM_i} \parallel k_i \parallel N_s)$, $x = h(k_i \parallel T \parallel N_s) \oplus h(K_{as} \parallel N_a)$, $y = h(T \parallel N_s \parallel k_i) \oplus N_a$, $z = h(T \parallel ID_{SM_i} \parallel k_i) \oplus PID_i^{new}$, $V_2 = h(K_{as} \parallel N_a \parallel x)$ and $V_3 = h(T \parallel y \parallel z \parallel k_i)$, and stores all generated seeds as $S =$

$\{s_1, s_2, \dots, s_n\}$. Details of these parameters are given in Figure 2. Finally, the PS sends message $M_{AU3} : \{x, y, z, V_2, V_3\}$ to the DA.

Step AUTH4: Upon receiving message M_{AU3} , the DA calculates the hash-output V_2 and compares it with V_2 received from the PS. If the comparison is verified, DA computes $TK = x \oplus h(K_{as} \parallel N_a)$ and generates the secret key $kh_i = h(TK \parallel N_a \parallel N_s)$ between DA and each SM _{i} . Then, it generates a set of temporary identities $TID_{i,m} = \{tid_{i,1}, tid_{i,2}, \dots, tid_{i,m}\}$ to be used in the aggregation time intervals. These identities are encrypted with keys kh_i as $TID_{i,m}^* = E_{kh_i}[TID_{i,m}]$. The DA finally computes $V_4 = h(TID_{i,m}^* \parallel kh_i \parallel ID_A)$, and after storing $TID_{i,m}$ and kh_i , sends message $M_{AU4} : \{(y, z, V_3) \parallel (TID_{i,m}^*, V_4)\}$ to each SM _{i} .

Step AUTH5: On receiving M_{AU4} , each SM _{i} computes $T = h(ID_{SM_i} \parallel k_i \parallel N_s)$ and then calculates V_3 to verify if V_3 received from DA is verified. If the verification is succeeded, it computes $N_a = h(T \parallel N_s \parallel k_i) \oplus y$, $TK = h(k_i \parallel T \parallel N_s)$, $kh_i = h(TK \parallel N_s \parallel N_a)$, $PID_i^{new} = h(T \parallel ID_{SM_i} \parallel k_i)$ and the seed $s_i = PID_i^{new} \oplus k_i$. Then, it calculates V_4 and compares it with V_4 received from DA. If the comparison passed, the SM _{i} decrypts $TID_{i,m}^*$ in order to achieve $TID_{i,m}$, and stores $\{s_i, PID_i^{new}, TID_{i,m}, kh_i\}$ for data aggregation. Note that the seed s_i used in each SM _{i} is the same as in the PS.

B. Data Aggregation Phase

This phase is done in two steps at the beginning of each time interval T_j . In step 1, as illustrated in Figure 3, the PS generates a set of random numbers $R_j = \{r_{i,j}\}_{i=1}^n$ at each time interval T_j using the adopted seeds in the authentication phase. Then, the PS computes the \mathfrak{R}^* as shown in Figure 3, and sends it to the DA. The DA then generates a time-stamp t_a and also encrypts the time-interval and sends them together to all SMs. Hereafter, each SM decrypts the received message to verify the time interval and then generates its own random number $r_{i,j}$ using the seed s_i generated in the authentication phase. This completes the first step of this phase. The motivation behind generating random numbers R at both PS and SMs is to mitigate the vulnerability in the similar scheme proposed in [22], in which random numbers are only generated at PS and then sent to individual SMs in the network. If a sequence of these numbers is sniffed by an attacker, it is possible to detect the seeds and upcoming random numbers. This certainly results in information disclosure of the consumed energy of SMs. Moreover, as random numbers are not transmitted in the network, there is a gain of communication overhead in the proposed LWS-DA scheme.

After that, in step 2, as shown in Figure 4, each SM first generates a time-stamp t_i , then chooses a temporary identity $tid_{i,j}$ from the temporary identity set TID_i which was received from DA in the authentication phase. Then, using the measured value M_i as the amount of energy consumed by SM _{i} and the pre-generated random number $r_{i,j}$, it computes a blinded value X_i by adding $r_{i,j}$ to M_i . Finally, SM _{i} sends its temporary identity $tid_{i,j}$, the blinded value X_i , the hash-output H_i , and its time-stamp t_i to the DA. Upon receiving the data from SMs, the DA sums up all the X_i s to send the total amount of consumed energy to the PS.

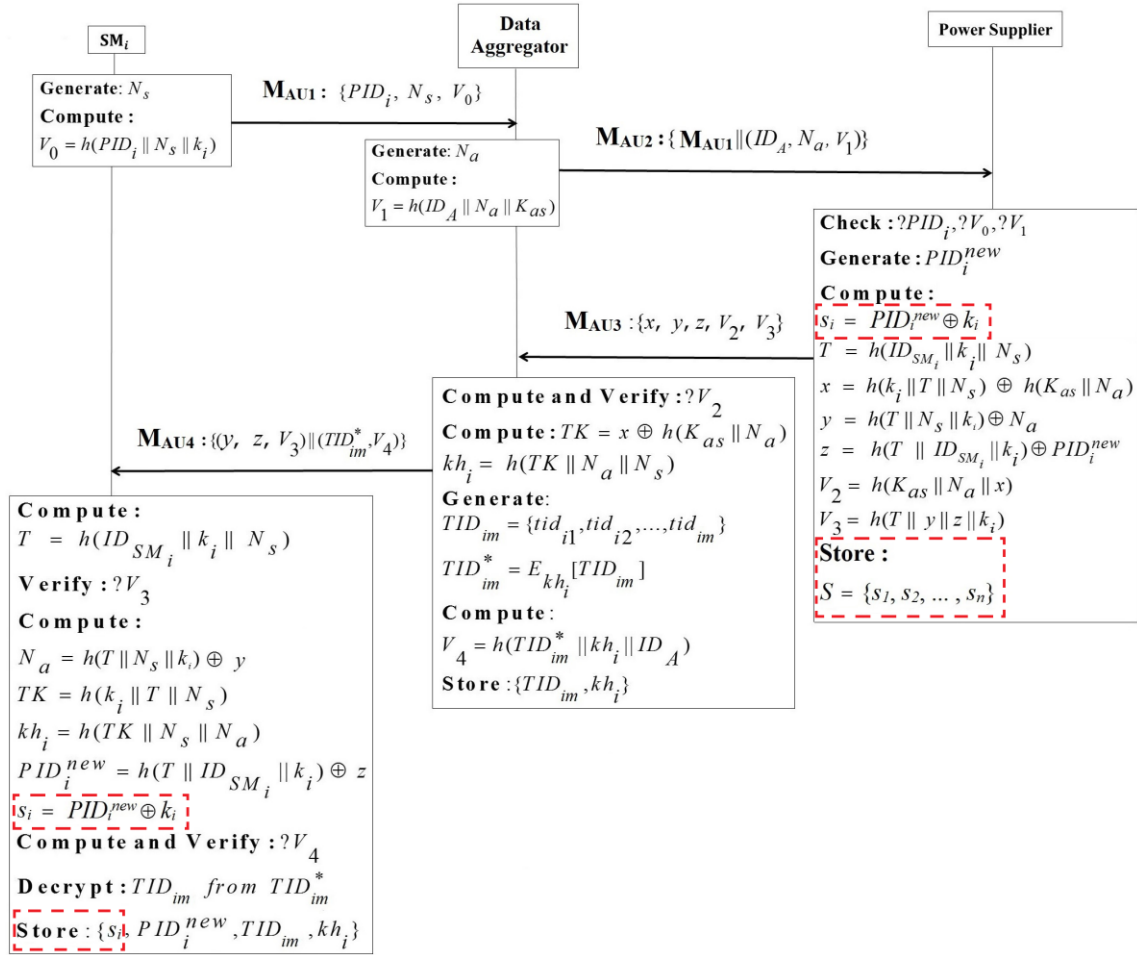


Figure 2. Modified Authentication Phase (modifications are shown in red rectangles)

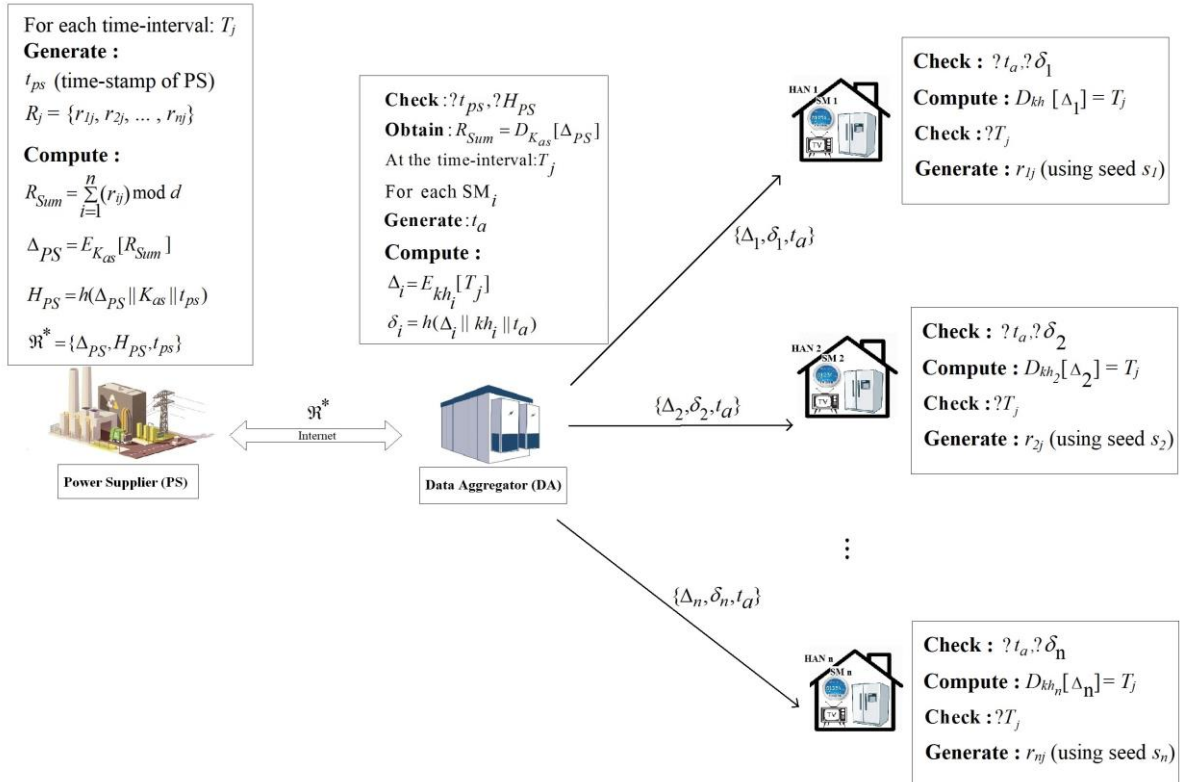


Figure 3. Step 1 in LWS-DA data aggregation phase

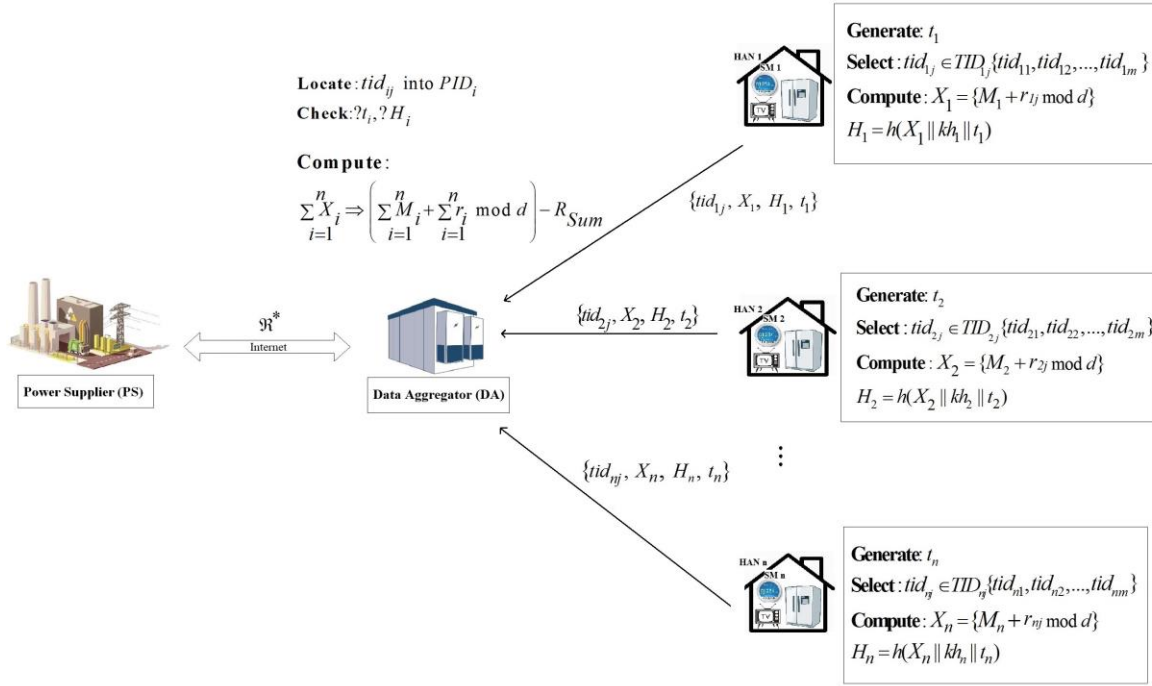


Figure 4. Step 2 in LWS-DA data aggregation phase

4. Security Maintained

The security analysis of the former scheme has been discussed in detail [22]. The modifications made by the proposed LWS-DA scheme do not degrade the security characteristics, as investigated in the following.

A. Authentication

As seen in Figure 2, the PS authenticates each SM_i by checking its PID_i as well as the hash-output V_0 , where only a legal SM has access to them. Also, the PS authenticates the DA by verifying the identity of the DA and the hash-output V_1 which only the legal DA is able to generate. The SM_i authenticates the PS by checking the value of hash-output V_2 and also the DA authenticates the PS by checking V_3 . Therefore, each entity in the considered network authenticates the other two entities.

B. Secure Key-Establishment

As presented in Figure 2, there is no key exchange between the entities of the network. In particular, the keys k_i , K_{as} and kh_i are not transmitted between entities on the communication channels. The key kh_i is generated in DA and SM_i individually. Therefore, the keys remain secret.

C. Data Confidentiality

The data on consumers' energy usage must be secret from others except for the SM_i itself and the PS. Therefore, the value metered by each SM_i is blinded with a random number r_{ij} which is from a long-enough range, i.e., $X_i = M_i + r_{ij} \mod d$. Therefore, the DA has access only to the blinded measurement X_i and cannot perceive the exact amount of consumed power by each SM. So transmitted data remains confidential.

D. Data Integrity

The DA has the option to check the integrity of the information received from the SM of every HAN. Also, DA needs to confirm the respectability of the pertinent data received from the PS during the information gathering. Using a one-way and non-collision hash function to transmit message hash, data integrity is also preserved.

E. Consumer Privacy

As discussed in the proposed scheme, only the PS can map the pseudo-identity PID_i of SM_i to its real identity

ID_{SM_i} . Therefore, the DA or an illegal entity cannot access the identity of consumers.

5. SECURITY ANALYSIS

Suppose an adversary, denoted as **A**, possesses the ability to intercept all forms of communication. This includes the capacity to replay, alter, eliminate, and rearrange messages. Furthermore, assume **A** has the ability to query the random oracle **h**, which produces strings of 256 bits. The security of the LWS-DA scheme is analyzed in the following.

A. Session Key Secrecy

To compute kh_i , **A** must compute $TK = x \oplus h(K_{as} \| N_a)$. **A** can obtain N_a , but not K_{as} since we assumed PS is trusted. **A** can query **h** to try to find an input that outputs x . But this will fail except with negligible probability of 2^{-256} since **h** is a random oracle. Therefore, **A** cannot compute TK or kh_i except with a negligible probability. Thus, **A** cannot compute any session key kh_i .

B. Entity Authentication

SM_i is assured of DA's identity because only DA can produce $E_{kh_i}[TID_{i,m}]$ which can be correctly decrypted by kh_i and is known only by DA and SM_i . DA is assured of SM_i 's identity because SM_i demonstrates knowledge of k_i by producing correct $h(T \| y \| z \| k_i)$ which DA can verify using values from PS. Thus, SM_i and DA are assured of each other's identities.

C. Data Confidentiality

A sees only the blinded meter readings $X_i = \{M_i + r_{ij} \mod d\}$. Without knowing r_{ij} , **A** cannot compute the actual consumption M_i . The value of r_{ij} is generated using seed s_i which we assumed **A** cannot obtain. Therefore, **A** cannot learn the smart meter readings except by guessing r_i which will succeed only with a negligible probability of 2^{-256} . Thus, **A** cannot learn any smart meter's power consumption.

D. Data Integrity

A cannot tamper with the aggregated energy consumption data, because it cannot forge the signatures of SMs on their blinded values. The signatures are based on the secret keys which we assumed **A** cannot obtain. The signatures are also verifiable by the DA using values from PS.

E. Privacy of SMs

A cannot learn the real identities of the SMs, because the SMs use temporary identities in the data aggregation phase. The temporary identities are encrypted with the session keys kh_i which we assumed **A** cannot compute. The temporary identities are also indistinguishable from random values. Thus, **A** cannot learn the real identities of the SMs.

Therefore, the LWS-DA scheme is secure against the attacks mentioned above if the random oracle is secure.

communication overheads are done in this section. Firstly, in subsection A, the comparison of computation overhead between LWS-DA scheme and the scheme in [22] is analyzed. Then, in subsection B, the communication overhead comparison between the two mentioned schemes is done, and finally, the result of simulations which shows the differences between the two mentioned schemes in terms of running time is illustrated in subsection C.

In the aforementioned comparisons, key values, identities, and random numbers are assumed to be of length 256-bits. Moreover, time-stamps, time interval, and R_{sum} are of lengths 64, 8 and $[Log_2^n]+256$ bits respectively, where n is the number of SMs. Calculations are done for a time period consisting of $m = 24$ time intervals. In other words, the data aggregation process of n SMs occurs for m times. After the m th round, the authentication process is repeated in order to continue the data aggregation in the next round.

6. Performance Evaluation

Performance evaluation of the proposed LWS-DA scheme along with evaluating computation and

Table 2 Computation Overhead of The Proposed LWS-DA Scheme VS. [22]

Phase Operation		Aggregation-step1 (one period)					
		[Gope et al.] Scheme			LWS-DA Scheme		
		Number of Operations		Input length	Number of Operations		Input length
Hashing	H_3	$24(n+3)$	$24(n+1)$	1040	$24(n+3)$	$24(n+1)$	328
			24×2	$584 + l^*$		24×2	$584 + l^*$
	H_4	$24(n+1)$		1352	—		
	Tot.	$24(2n+4)$		—	$24(n+3)$		—
Encryption	E_1	—			$24n$		8
	E_2	24		$264 + l^*$	24		$264 + l^*$
	E_4	$24n$		776	—		
	Tot.	$24(n+1)$		—	$24(n+1)$		—
Decryption		24×2	24	$264 + l^*$	24×2	24	$264 + l^*$
			24	776		24	8
Concatenation		$24(9n+3)$			$24(2n+4)$		

* $l = [Log_2^n]$

This paper compares the proposed LWS-DA scheme to Gope et al.'s prior scheme in [22] in terms of communication overhead and overall efficiency. The comparison is comprehensive and robust, covering all the key performance indicators and metrics. The scheme in [22] was selected as a benchmark because it is similar to LWS-DA in that it uses lightweight cryptographic primitives for secure data aggregation in smart grid networks. While several other related schemes have been recently proposed, a further detailed comparison with each of these schemes is not essential to demonstrate the advantages of LWS-DA. The in-depth analysis of security properties and the experimental results validate that LWS-DA accomplishes the key objectives of an efficient and secure data aggregation scheme suitable for resource-constrained smart grid environments. The insights from the comparison sufficiently highlight the technical contributions of LWS-DA without requiring additional comparative analyses at this stage, since the existing scheme, being a well-established and widely accepted benchmark in the field, provides a solid basis for comparison. Furthermore, the methodology used for this comparison is rigorous, ensuring that all relevant aspects are considered. Therefore, additional comparisons with other schemes, including recent works, are not essential. The results obtained from this comparison are sufficient to draw reliable conclusions and make informed decisions about the performance and efficiency of the proposed work. This eliminates the need for further comparisons, allowing us to focus on enhancing and implementing the proposed scheme.

A. Computation Overhead Comparison

Table II shows the comparison of the computation overhead of [22] and LWS-DA scheme in terms of hashing, encryption, decryption, and concatenation operations. The comparisons are done by calculating the number of input bits for each operation and also the number of usages of each operation. It is worth noting that the index of each operation indicates the number of concatenated input arguments and the parameter n is the number of SMs. For the sake of simplicity, we only focus on the items with different values.

Hashing. As shown in Table II, both schemes use $24(n + 3)$ three-input hashing functions (H_3), in which $24(n + 1)$ functions burden the input length of 1040 bits for the scheme in [22] while this value is 328 bits in our scheme, that is less than a third of the scheme in [22]. Moreover, the scheme in [22] uses $24(n + 1)$ four-input hashing functions (H_4) with input length of 1352 bits, while there is no usage of this operation in our scheme LWS-DA. This yields a

significant improvement in the overall computational complexity. Figure 5 shows the number of input-bits of hash functions (H_3 and H_4) in LWS-DA compared to the scheme of [22] versus the number of SMs. While it seems that the length of input bits grows exponentially in [22], this length increases linearly in LWS-DA.

Encryption. In comparison between the schemes, it is apparent that $24n$ number of four-input encryption function (E_4) employed in [22] with an input length of 776 bits is replaced with $24n$ number of one-input encryption function (E_1) in LWS-DA with an input length of 8 bits. This is absolutely a significant improvement in terms of computation. This can also be seen in Figure 6 which shows the number of input bits of encryption functions for both schemes in terms of the number of SMs. Great outperformance of LWS-DA is observed.

Decryption. Considering the concurrency of computations in SMs, we only take into account the computational overhead of the decryption operation for one SM. Hence, as seen in Table II, both schemes use 24×2 decryption operations, in which 24 operations in [22] have an input length of 776 bits while this value is 8 bits in LWS-DA scheme.

Concatenation. The proposed scheme in [22] uses $24(9n + 3)$ number of concatenations, whereas our LWS-DA scheme uses $24(2n + 4)$ concatenations. The improvement becomes significant when the number of SMs becomes large, as observed in Figure 7.

Considering the aforementioned observations, our proposed scheme uses a smaller number of hash functions, encryption functions, decryption functions, and concatenation operations with less input bit-length than that in [22]. This certainly results in a light-weighted data aggregation scheme.

B. Communication Overhead Comparison

Table III shows the difference between the scheme presented in [22] and LWS-DA scheme in terms of communication overhead. As seen in Table III, in the scheme of [22], the number of bits transmitted by a message from PS to DA in step 1 of the data aggregation phase is increasing linearly with respect to the number of SMs. However, in the proposed scheme, the complexity is logarithmic. In other words, as the number of SMs increases, our proposed scheme transmits fewer bits than the scheme proposed in [22], asymptotically. One reason for this improvement is due to generating random numbers simultaneously at both PS and SMs, rather than transmitting them over the network.

Table 3 Communication Overhead of The Proposed LWS-DA Scheme VS. [22]

Scheme		[Gope et al.] Scheme (bits)	LWS-DA Scheme (bits)
Phase			
Aggregation-step 1	PS to DA	$(n \times 1032) + l^* + 584$	$l^* + 584$
	DA to SM _i	1352	328

C. Simulation Results

According to the modification that led to the achieved results towards the overhead reduction, we anticipate a significant reduction in the running time of the data aggregation scheme. To address this point, we conduct simulations using Python 3.7 in order to experimentally show the effectiveness of our proposed LWS-DA scheme in comparison with [22].

As noted in [23], at the same security level, the time performance of AES-CBC encryption method is more than other methods such as Paillier, RSA, ECC-based and, so on. On the other hand, since there is no key exchange between the components of the network in both schemes,

we use symmetric cryptographies in comparison. Therefore, for the simulation of both schemes, we adopt the AES-CBC encryption with a 256-bits key size which is a lightweight cryptographic method. Moreover, all hashing operations are done with SHA-256 which is a non-collision one-way hash function [24]. Figure 8 shows the difference between LWS-DA and the former schemes in terms of aggregation time for a duration of one period including 24 time intervals. As seen, while the running time of [22] grows exponentially with the number of SMs, our scheme demonstrates a linear increase with a gentle slope.

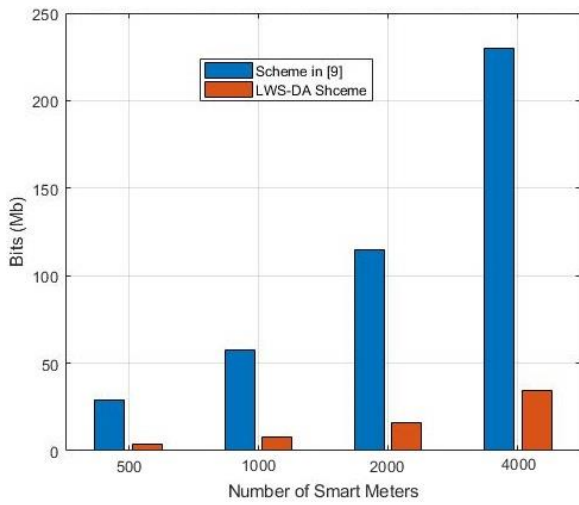


Figure 5. Number of input-bits of hash functions versus number of SMs.

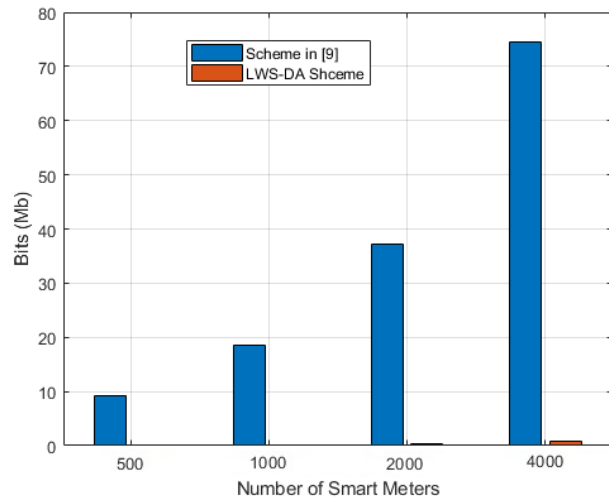


Figure 6. Number of input-bits of encryption functions versus number of SMs.

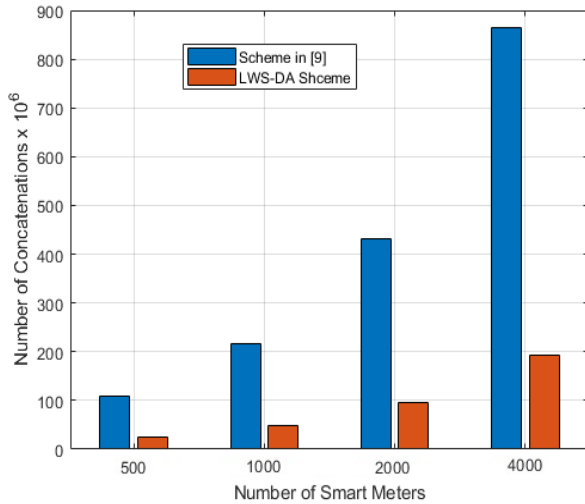


Figure 7. Number of concatenations versus number of SMs

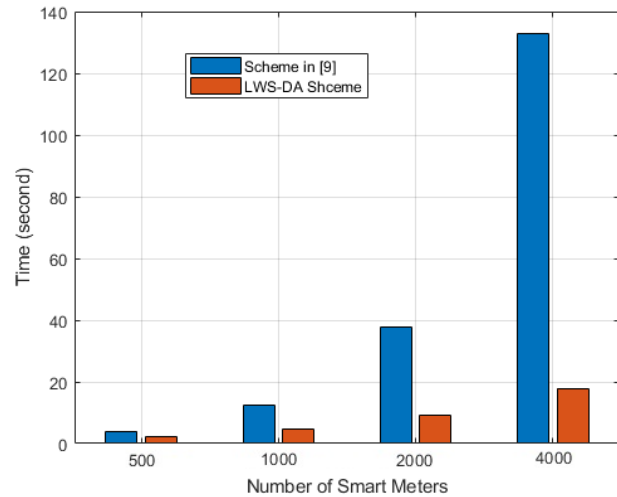


Figure 8. Aggregation time in terms of the number of SMs

7. Conclusion

In this paper, an improved light-weight secure scheme for data aggregation in large-scale IoT-based smart grids is proposed. Firstly, the security of the scheme is discussed and then the effectiveness of computation and communication overheads are demonstrated. The performance analysis shows that the proposed scheme becomes computationally and communicationally more

lightweight than the scheme in the literature, especially when the number of SMs grows in the network. Therefore, the proposed scheme can be used in large-scale smart grids in order to aggregate the data consumption of SMs.

References

- [1] IoT Analytics, "Why the internet of things is called internet of things: definition, history, disambiguation," <https://iot-analytics.com/internetof-things-definition>, 2014.
- [2] M. A. Ferrag, L. A. Maglaras, H. Janicke, J. Jiang, L. Shu, "Authentication protocols for internet of things: A Comprehensive Survey", Hindawi, Security and Communication Networks, vol. 2017, Article ID 6562953.
- [3] G. Gharepatyan, M. Shahidehpour and B. Zaker, Smart Grids and Microgrids, Tehran: Amirkabir University of Technology, 2019.
- [4] N. Saxena, B. J. Choi, and R. Lu, "Authentication and authorization scheme for various user-roles and devices in smart grid," IEEE Trans. on Information Forensics and Security, vol. 11, pp. 907-921, May 2016
- [5] E. Kabalci, Y. Kabalci, Smart Grids and Their Communication Systems, Singapore: Springer, 2019
- [6] J. Zhang, Y. Zhao, J. Wu, and B. Chen, "A lightweight privacy-preserving data aggregation scheme for edge computing," in 15th International Conference on Mobile Ad-hoc and Sensor Systems, October 2018.
- [7] H. Shen, M. Zhang and J. Shen, "Efficient Privacy-Preserving Cube-Data Aggregation Scheme for Smart Grids," IEEE Trans. on Information Forensics and Security, vol. 12, pp. 1369-1381, 2017.
- [8] Z. He, Sh. Pan, D. Lin, "PMDA: privacy-preserving multi-functional data aggregation without TTP in smart grid," in 17th IEEE International Conference on Trust, Security and Privacy In Computing and Communications, 2018
- [9] R. Lu, X. Liang, X. Li and X. Shen, "EPPA: an efficient privacy-preserving aggregation scheme for secure smart grid communications," IEEE Trans. on Parallel Distribution Systems, vol. 23, pp. 1621-1631, September 2012.
- [10] S. Li, K. Xue, Q. Yang and H. Peilin, "PPMA: Privacy-Preserving Multisubset Data Aggregation in Smart Grid," IEEE Trans. on Industrial Informatics, vol. 14, pp. 462-471, 2018.
- [11] Y. Chen, J. Ortega, P. Castillejo and L. Lopez, "A Homomorphic-Based Multiple Data Aggregation Scheme for Smart Grid," IEEE Sensors Journal, vol. 19, pp. 3921-3929, 2019.
- [12] J. H. Jo, S. I. Kim and H. D. Lee, "Efficient and Privacy-Preserving Metering Protocols for Smart Grid Systems," IEEE Trans. on Smart Grids, vol. 7, pp. 1732-1742, 2016.
- [13] A. Abdallah and X. Shen, "A Lightweight Lattice-Based Homomorphic Privacy-Preserving Data Aggregation Scheme for Smart grid," IEEE Trans. on Smart Grid, vol. 9, pp. 396-405, January 2018.
- [14] M. M. Fouda, Z. M. Fadlullah, N. Kato, R. Lu and X. Shen, "A lightweight message authentication scheme for smart grid communication," IEEE Trans. on Smart Grid, vol. 2, pp. 675-685, December 2011.
- [15] H. J. Jo, I. S. Kim and D. H. Lee, "Efficient and privacy-preserving metering protocols for smart grid," IEEE Trans. on Smart Grid, vol. 7, pp. 1732-174, May 2016.
- [16] E. Vahedi, M. Bayat, M. Pakravan and M. Aref, "Secure ECC-based privacy preserving data aggregation scheme for smart grids," in Computer Networks, vol. 129, no. 1, pp. 28-36, 2017
- [17] Y. Liu, W. Guo, C. Fan, L. Chang, and C. Cheng. "A practical privacy-preserving data aggregation (3PDA) scheme for smart grid," IEEE Trans. on Industrial Informatics, vol. 15, pp. 1767-1774, March 2019.
- [18] D. He, N. Kumar, Sh. Zeadally, A. Vinel, L. T. Yang, "Efficient and privacy-preserving data aggregation scheme for smart grid against internal adversaries", IEEE Trans. on Smart Grids, vol. 8, pp. 2411-2419, 2017
- [19] A. Abdallah and X. Shen, "Lightweight security and privacy preserving scheme for smart grid customer-side networks," IEEE Trans. on Smart Grids, vol. 8, pp. 1064-1074, May 2017
- [20] M. A. Mustafa, S. Cleemput, A. Aly and A. Abidin, "A Secure and Privacy-Preserving Protocol for Smart Metering Operational Data Collection," IEEE Trans. on Smart Grids, vol. 10, pp. 6481-6490, 2019.
- [21] F. Knirsch, G. Eibl and D. Engel, "Error-Resilient Masking Approaches for Privacy Preserving Data Aggregation," IEEE Trans. on Smart Grid, vol. 9, pp. 3351-3361, 2018.
- [22] P. Gope and B. Sikdar, "Lightweight and privacy-friendly spatial data aggregation for secure power supply and demand management in smart grids," IEEE Trans. on Information Forensics and Security, vol. 14, pp. 1554-1566, June 2019.
- [23] F. Maqsood, M. Ahmed, M. M. Ali and M. A. Shah, "Cryptography: a comparative analysis for modern techniques", International Journal of Advanced Computer Science and Applications, vol. 8, Issue 6, 2017
- [24] H. Gilbert and H. Handschus, "Security analysis of SHA-256 and sisters", International Workshop on Selected Areas in Cryptography, Springer, SAC 2003, pp. 175-193.
- [25] J. So, C. He, C. Yang, S. Li, Q. Yu, R. E. Ali, B. Guler and S. Avestimehr, "LightSecAgg: a lightweight and versatile design for secure aggregation in federated learning", in Proceedings of Machine Learning and Systems, vol. 4, pp. 694-720, 2022.
- [26] M. Zhang, Y. Li, Y. Ding, and B. Yang, "A Lightweight and Robust Multi-Dimensional Data Aggregation Scheme for IoT", IEEE Internet of Things Journal, vol. 1, no. 1, pp. 1-1, 2023
- [27] J. Qian, Z. Cao, X. Dong, J. Shen, Z. Liu, and Y. Ye, "Two Secure and Efficient Lightweight Data Aggregation Schemes for Smart Grid," IEEE Trans. on Smart Grid, vol. 12, no. 3, pp. 2625-2637, May 2021
- [28] Junhua Wu, Zhuqing Xu, Guangshun Li, Cang Fan, Zhenyu Jin, Yuanwang Zheng, "E-LPDAE: An Edge-Assisted Lightweight Power Data Aggregation and Encryption Scheme", Security and Communication Networks, vol. 2022, Article ID 6218094, 12 pages, 2022

CONTENTS

A Data Replication Algorithm for Improving Server Efficiency in Cloud Computing Using PSO and Fuzzy Systems	Mostafa Sabzekar - Ehsan Mansouri Arash Deldari	1
An Efficient Ramp Secret Sharing Scheme Based on Zigzag-Decodable Codes	Saeideh Kabirirad - Sorour Sheidani Ziba Eslami	15
Analysis of the Impact of Wireless Three-User Multiple Access Channel Coefficients Correlation on Outage Probability: A Copula-Based Approach	Mona Sadat Mohsenzadeh Ghosheh Abed Hodtani	25
Optimization of FlexiTP Energy-Aware Algorithm in Wireless Sensor Networks	Hamid Mirvaziri	31
Dynamic Security Risk Management Considering Systems Structural and Probabilistic Attributes	Masoud Khosravi-Farmad Abbas Ghaemi-Bafghi	41
A Lightweighted Secure Scheme for Data Aggregation in Large-Scale IoT-Based Smart Grids	Mohammad J. Abdolmaleki Amanj Khorramian Mohammad Fathi	57