Bio-inspired Think-and-Share Optimization for Big Data Provenance in Wireless Sensor Networks

Adel Alkhalil[1], Rabie Ramadan[2], Aakash Ahmad[3]
College of Computer Science and Engineering
University of Ha'il
Saudi Arabia

***Abstract*—Big data systems are being increasingly adopted by the enterprises exploiting big data applications to manage data-driven process, practices, and systems in an enterprise wide context. Specifically, big data systems and their underlying applications empower enterprises with analytical decision making (e.g., recommender/decision support systems) to optimize organizational productivity, competitiveness, and growth. Despite these benefits, big data applications face some challenges that include but not limited to security and privacy, authenticity, and reliability of critical data that may result in propagation of false information across systems. Data provenance as an approach and enabling mechanism (to identify the origin, manage the creation, and track the propagation of information etc.) can be a solution to above mentioned challenges for data management in an enterprise context. Data provenance solution(s) can help stakeholders and enterprises to assess the quality of data along with authenticity, reliability, and trust of information on the basis of identity, reproducibility and integrity of data. Considering the wide spread adoption of big data applications and the needs for data provenance, this paper focuses on (i) analyzing state-of-the-art for holistic presentation of provenance in big-data applications (ii) proposing a bio-inspired approach with underlying algorithm that exploits human thinking approach to support data provenance in Wireless Sensor Networks (WSNs). The proposed 'Think-and-Share Optimization' (TaSO) algorithms modularizes and automates data provenance in WSNs that are deployed and operated in enterprises. Evaluation of TaSO algorithm demonstrates its efficiency in terms of connectivity, closeness to the sink node, coverage, and execution time. The proposed research contextualizes bio-inspired computation to enable and optimize data provenance in WSNs. Future research aims to exploit machine learning techniques (with underlying algorithms) to automate data provenance for big data systems in networked environments.**

Keywords*—Big data systems; data provenance; fuzzy logic; bio-inspired computing

I. INTRODUCTION

Traditionally, the provenance of an object or data includes information about the ownership, source, transformation and evolution of data or object during their life span [1]. The term data provenance, as per the Encyclopedia of Database Systems, formally refers to 'a record trail that accounts for the origin of a piece of data (in a database, document or repository) together with an explanation of how and why it got to the present place [33].

Provenance of data have been proven as a useful technique–optimizing visibility and transparency of the data–to enhance traceability of errors back to their root cause(s) during data analytics and processing [34]. In big data systems, provenance information enhances the data trustworthiness (identification of data sources) and support data compliance (policies for data processing), to ensure accountability and compliance [2]. In data-centric systems and applications, preserving the security and privacy of data are of central importance to avoid the exploitation or misuse of critical data [25, 35]. Critical data or specifically security critical data can be diverse that can range from personal information of users [35] to security features of systems such as access control usage control decisions and data forensics [25].

The characteristics of big-data systems such as size and magnitude of data, distributed users, and heterogeneous platforms poses some serious challenges to data provenance. Further, big-data applications are usually dynamic and heterogeneous by nature. They involve many components provided by various vendors which must be integrated together to develop and operate the system. Therefore, tracing the provenance of an object in big data including the collection of evidences and data from various sources to determine the causes and effects is very difficult [3]. Many of big-data applications rely on Wireless Sensor Networks (WSN) for their common activities to process and transmit data among applications [7]. A WSN consists of many small and low-cost sensor nodes. The nodes in WSNs have different limitations such as energy, processing capabilities, small communication range, small sensing range, and storage capabilities. These characteristics make the data transmitted across WSNs exposed to various threats. This signifies the need for a mechanism that can enhance the reliability of data collected from WSNs by assessing their trustworthiness. This paper addresses the challenge of node trust in WSNs that transmit large amount of data in an enterprise context. The primary question to be answered by this research is: How nodes that process, generate, and consume large size of data in WSNs can be trusted to maintain the integrity of data?

This paper reviews the current situation with regards to big-data provenance. It focuses on how to apply data provenance in WSNs and gain nodes trust with minimum memory usage and reduced packets size. This paper also proposes a solution to the aforementioned problems by applying data provenance in WSN systems. Data provenance

can help support data integrity, ultimately enhancing data security in networked systems [34] to improve the security. However, enhanced security measures need high computation and storage resources that represent the challenge for big data systems. The proposed contributions of this research are:

- Exploiting bio-inspired approach to support data provenance in WSNs in the context of enterprise scale systems.
- Developing a novel algorithm named Think and Share Optimizations (TaSO) that modularizes and automates nodes' selection process inspired from human collaboration in education.

We have evaluated the proposed solution and its underlying algorithms bases on test-case strategy. Results of the evaluation demonstrate algorithmic efficiency in terms of connectivity, closeness to the sink node, coverage, and execution time.

The paper is structured as follow: Section 2 provides an overview of big data provenance. Section 3 summarizes the state-of-the-art with a holistic presentation of provenance in big data applications. Section 4 discusses our proposed solution for data provenance in WSNs. Section 5 presents the bio-inspired algorithm as a modular solution for data provenance in WSNs. Section 6 presents the experimental results of evaluation Section 7 presents some threats to the validity of proposed solution with conclusions and dimensions of future work.

II. Background and Related Work on Big-Data Provenance

Data provenance is a well-known research area within database and data mining. It considers the problem of identifying the origin, the creation, as well as the propagation processes of data [4]. It may be defined as the process of detecting the lineage and the derivation of data and data objects [5]. The execution environment of transformation such as the library versions, operating systems, and the nodes responsible for excusing the transformation may also be considered as provenance data [3].

Data provenance is an essential component in a various areas such as: database management systems, workflow management systems, distributed systems, and debugging ICT systems [4]. In the case of security violations, a system administrator should be able to identify the origination of the error, in addition to its causes and impacts [6].

In the literature (see for example [3, 5, and 7]), a number of advantages for applying provenance data can be obtained from different resources in several areas including:

- *Business Domain*. Data provenance can provide stakeholders an integrated vision of valuable historical data from multiple sources.
- *Data Quality*. Data provenance can be utilized to assess the reliability and trust of data.
- *Audit Trail*. It can be used to trace audit trails for error detections and debugging.
- *Reusability and Reproducibility*. It can be used to verify, compare, and repeat results.
- *Informational*. The provenance meta-data can be used to query data for discovering and browsing historical data.
- *Benchmarking*. It can be used to identify and analyze performance, compute performance metrics, and test the ability to exploit commonalities in data and processing.

Being an important area for data quality and authenticity, the application of data provenance has received an increasing attention in research [8]. It is a well-known topic in the fields of database management systems, workflow management systems, and distributed systems communities [9, 10]. However, applying data provenance in big data is a fairly new trend that involves continuing development for provenance-aware systems. Therefore, most of the existing studies on data provenance in big-data applications are exploratory [2, 3, 4, 5, and 7], descriptive [11], or case-based research [12]. For example, studies from [13], [14], and [15] focus on the general conceptualization and definition of data provenance, as well as its importance, challenges, and applications. Further, many previous studies explored the business benefits, such as [16, 17]. The problem of security and privacy encountered in provenance management also received an increasing attention [2, 18, 19].

Although big data provenance has been increasingly gaining attention, it is still in an early stage of maturity. Majority of the existing research on big data provenance focused on the application of data provenance on big data [3], exploring the challenges of capturing, analyzing, visualizing big data provenance, and identifying future research directions. Further, a number of studies are focused on capturing and modeling provenance, visualizing [20], and mining provenance data [13]. Despite the fact that those works provided a considerable contribution to the advancement of data provenance utilization onto big data, the size, heterogeneity, complexity, and overhead computation and storage remain significant challenges.

This section highlights some of the contributions in big-data provenance. In [4] the author highlighted the problem of collecting and analyzing provenance in big data. The study reviewed a number of studies in this area in an attempt to present the-state-of-the-art. This was followed by the provision of an overview of 14 research issues and challenges of provenance within big data in which the author highlighted potential future research directions. Similarly, Glavic [3] highlighted the need for provenance in big data. The author argued that without provenance information, it is impossible for a user to understand the relevance of data, estimate its quality, investigate unexpected or inaccurate results, and define meaningful access control policies. Glavic then examined how Big Data bench-marking could benefit from provenance information. Simmhan et al. [7] created a taxonomy of data provenance techniques, and applied a classification to existing studies in the area. It aimed to help in building scientific and business management systems in order to understand the designs of existing provenance systems. The taxonomy is focused on categorizing provenance studies based on their intentions, contexts, provenance storage and propagation.

III. Related Work on Data Provenance Applications

A number of studies focused on data provenance utilization within a specific context of big-data applications. This section explores some of the recent applications of data provenance including social media, E-science, cloud computing.

A. Social Media

In the today's ubiquitous influence of social-media's contents and activities, information credibility has increasingly become a major issue [13]. Correspondingly, the identification of false information and rumors circulation in social-media environments attracted a considerable recent research and interests [20]. A number of studies attempted to address the lack of information credibility in social media through data provenance. Although information diffusion in social media networks has received an increasing attention [21], there is limited works on the reverse process of information diffusion, information provenance [22].

In [13] the author developed a web-based tool for collecting a number of attributes of interest associated with a social-media user. It provides a technique to effectively combine different attributes from different social-media sites into a single user profile. The tool was evaluated using different types of Twitter and Facebook users. In the case of further provenance attributes are needed, the tool offers best possible URL (Uniform Resource Locator) to help for further findings. Although, the contribution of this work provides helpful user-profile metadata, it lacks other essential provenance information including paths and sources. The study in [23] considered seeking the provenance of information for a few known recipients. To achieve this there is a need to find the provenance paths first. Therefore, the study attempted first to explain propagation processes from the sources to the P-nodes. It also includes information of the nodes responsible for retransmitting information through intermediaries. In [24] a method was developed that integrate provenance data on two levels; a low-level through information cascades, and a high-level through similarity-based clustering.

B. Cloud Computing

Cloud computing is still at an early level of maturity and it is still evolving in which provenance is yet to be fully implemented. A recent attention has been devoted to this aspect. They explored the challenges of implementing provenance in the cloud environment as in [25 and 26]. In [25] the authors highlighted the challenge of collecting and merging provenance information from a dynamic environment in which resources are independence; for example, different domains (scientific, business, database), different platforms (windows, Linux), and various applications. Further, computation and storage overhead also presents another challenge [26]. In [27] a framework named CloudProv was developed. It aimed to integrate, model and monitor data provenance in cloud computing environments. CloudProv was based on a method that allows users to model the collected provenance information to continuously obtain and monitor such information to be utilized for real-time applications. Oruta [28] developed a privacy-preserving auditing mechanism. It supports data sharing in untrusted cloud environments. The proposed mechanism makes use of homomorphism authenticators that allows third parties auditors to check the integrity of the shared data from a certain user group, without the need for accessing all data.

C. E-Science

Provenance in e-Science is studied comprehensively see for example [7, 29, 30]. The research in e-Science provenance has focused on the process of capturing, modeling, and storing provenance information. It uses provenance for various purposes within the scientific domains. Publications and Digital Object Identifiers (DOIs) are common examples of provenance [7]. The Geographic information system (GIS) standards advise that metadata about the quality elements of a dataset should include a description of its lineage. This can help users to decide if the dataset meets their requirements [2].

D. Provenance in Wireless Sensor Networks

Typically, all nodes have sensing, limited data processing, and communicating capabilities. Unlike wired computer networks, the nodes of WSNs are often resource-tightened and deployed in an unprotected environment [31]. Further, communications in WSNs depend upon multi-hop wireless signal relays. These unique characteristics make data transmitted across WSNs exposed to threats. Therefore, there is a need for a mechanism that can enhance the reliability of the data collected from WSNs by assessing their trustworthiness. Provenance is a mechanism of trust and reputation evaluation which can enhance the security of networks [32]. Network information is crucial for seeking provenance information [1]. There are two main approaches: First, the use of available information to search the provenance directly. It is suitable in cases that all recipients are known for particular information. The second approach is to detect the flow of information propagation from origins to all known recipients. It can be suitable in cases of small number of recipients [1].

IV. WSN Model and Problem Statement

WSNs consist of a set of nodes S forming a connected network that can be represented by a graph G (V, E), where V is the nodes and E are the links in the graph. E represents the connectivity of each node, communication range s_r, in the graph G. In addition, a node might be able to sense its surroundings with a sensing range s_s. the sensing model in this context is considered as a binary model where any target falls within the sensing range of a node will be detected (1) and otherwise the node will report undetected (0). A node could work on one of two modes sensing or routing. During the sensing mode, a node is responsible for sensing the environment and sending its findings to its neighbor towards a sink node. In this case, the node is considered as a source node in the network. The sink node is considered as one of the powerful nodes that could handle the received messages from all of the source nodes. The second mode of a node is the routing mode where a node works only as a router that forwards the received nodes from its neighbors.

We assume a WSN with a set of source nodes S distributed in the networks and another set of routing nodes S_r. Nodes,

either sources or routers; have different limitations such as energy, processing capabilities, small communication range, small sensing range, and storage capabilities. At the same time, they are usually deployed in unattended areas in which it is exposed to multiple threats such as packet injection or node capturing. Therefore, data provenance is a must especially for critical WSNs such as the ones used for battle field or healthcare monitoring. However, it is not appropriate that nodes keep track of each and every packet with accumulated route, especially in a large scale WSN. Thus, the focus of this paper is on how to apply data provenance in such networks and gain nodes trust with minimum memory usage and reduced packets size.

The second problem that this paper handles is the nodes trust. As stated before, nodes in a WSN is prone to failure as well as attacks; so, the question is how nodes in such network can be trusted? We consider three parameters that could help in nodes trust which are nodes availability, nodes neighbors' opinion, and node's drop rate. Although these parameters could form a satisfying trust model for a sensor node, this information could be uncertain. Therefore, another issue that needs to be considered is the parameters uncertainty.

V. Data Provenance Solution Approach

In this section, we show how data provenance problem in WSN could be solved. The solution considers limited node resources in terms of energy, memory, and communication range. In addition, we propose a trust model based on fuzzy logic considering the previously mentioned parameters: nodes availability, nodes neighbors' opinion, and node's drop rate.

The data provenance problem in WSNs is transferred into an optimization problem where instead of keeping all nodes busy with data provenance, some of these nodes will be selected as data provenance nodes (DP). For a node to be a DP node, it has to satisfy the following conditions:

1) Closed to a source node $s \in S_s$

2) Each Source $s \in S_s$ is covered by at least one provenance node.

3) Has enough energy to handle the data provenance requirements.

4) Has enough memory to handle the data provenance requirements.

5) Trusted node in terms of availability, neighbors' votes, and Message Drop Rate (MDR).

6) The DP nodes have to be connected to each other.

These conditions show that finding DPs in this case is a hard problem where optimal solutions might not be able to find nodes using these constraints. Therefore, greedy algorithms could be the best solution. In the following subsections, our proposal for node trust and the proposed node selection solutions are described.

A. Fuzzy based Trust Model

Looking at the nodes properties and their role in WSNs, there are some parameters that can be used to trust a node. For instance, non-availability, neighbors voting, and message drop rate. These parameters do not have to be captured during the initialization phase; however, in the subsequent phases, these parameters are captured for trust evaluation.

The non-availability parameter means how much time a node is turned off compared to the total time that it supposed to be on. Neighbor's vote is another parameter that it is used to evaluate nodes' trust. One more parameter is used for nodes trust which is message drop rate which could be computed by the sink node.

A fuzzy logic controller is used to handle nodes trust. The trust will be computed for all of the nodes during the setup phase with rough estimate; however, in the following rounds, the trust becomes more realistic due to the actual operation of the nodes. The input membership functions are based on three linguistic variables as shown in Fig. 1(a), (b), and (c). Linguistic variables are assumed to be low, medium, and high. The output membership function with two linguistic variables, trusted and not trusted, is presented in Fig. 2.

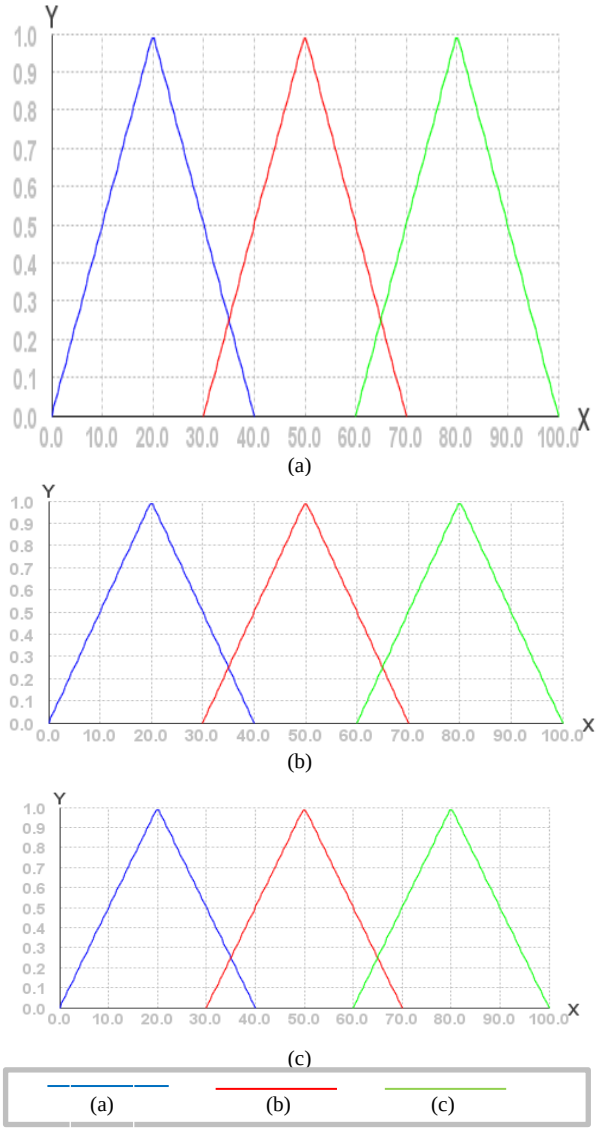

Fig. 1. Fuzzy Input Linguistics, (a) Availability ,(b) Neighbors Votes , (c) Messge Drop Rate.

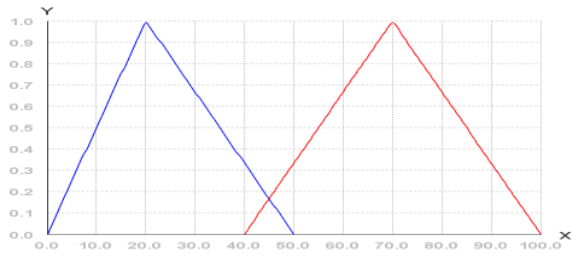

Fig. 2. Fuzzy Output Lingusitcs.

B. *Think-and-Share Optimization (TaSO) Algorithm*

In the following, we now present the algorithm as a modular solution for data provenance. The algorithm is presented in terms of the required inputs, outputs and processing steps along with the description of the algorithm and its underlying steps. TaSO algorithm consists of four different phases which are *Think*, *Pair*, *Share*, and *Evaluate* as detailed below.

Phase 1. Think Phase: represents the initial phase in which the system generates initial solutions for the problems under consideration.

Phase 2. Pair Phase: focuses on generating some agents and solutions (from Phase 1) are assigned/distributed to the individual agents. Each of the generated agents aims to find the optimal solution from the set of solutions that have been assigned to it. The agents can work iteratively until the most optimal solution is found.

Phase 3. Share Phase: represents a phase where the agents share the optimal solutions with each others to proceed for building the final solution. During the share phase, the algorithm may jump back to the Paring Phase (Phase 2) for some rounds.

Phase 4. Evaluation Phase: is the terminal phase that aims to evaluate the best possible solution. **From a technical perspective**, the most optimal solution(s) is passed to the Think phase (Phase 1) to assess, evaluate, and select the optimal solution that have been shared by the agents.

The TaSO could be summarized in steps as follows:

TaSO Algorithm:

Input
problem description,
number of internal Iterations (I),
number of external Iterations (E),
[required performance],
icounter = 0, xcounter=0 , n, m

Phase 1 - Think: the thinker generates n random solutions, S1 S1, S2, ….., Sn.

Pair:
1. Generate m agents A1,……, Am
2. Divide the n solutions on the m agents
3. Evaluate the solutions
4. Generate new solutions based on the old ones
5. icounter ++;
6. if (icounter < I) go to step 3

Phase 2 - Share:
1. xcounter++
2. Agents share their best solutions to each other
3. Agents evaluate their current solutions
4. Agents remove solutions with minimum performance
5. if (xcounter < E) go to Pair step 3

Phase 3 - Evalution:
1. Agents return their best solutions to the thinker
2. The thinker evaluates the returned solutions and selects the best solution(s) as final solution to the problem.
3. Terminate

Given a set of nodes, some of them are source nodes, deployed in a specified area. Regarding to our problem, some random solutions are initially generated; at the same time, a set of agents are generated to work on the solutions. Agents try to find the best solutions and pass them to the main agent. Solutions are evaluated based on the following criteria:

- Closeness the sink node in terms of number of hops. To simplify the computation, we assume that the max number of hop is equal to (n-1) where n is the number of nodes in the network, considering the network is a linear network.
- Percentage of covered source nodes out of the total number of source nodes.
- For energy, the node has to have more than 50% of the max energy.
- For memory, the node has to have enough memory for at least half of the nodes IDs.
- The node is trusted as output of the fuzzy logic controller.
- Connectivity percentage is measured by the percentage of the selected nodes out of the total number of selected provenance nodes.

VI. Evaluation and Experimental Results

In this section, we will test the think and share algorithm compared to Brute Force algorithm for solving the data provenance in WSNs [35]. Comparison criteria involve running time, coverage to the source node, closeness to the sink node, and connectivity percentage. The simulation is built using java program running in Intel Core i7 with 8GB RAM. The deployment area is assumed to be a square of 2000 m X 2000 m. 100 nodes are initially deployed in this area with 100 m communication range. 20 nodes are considered source nodes and they are randomly distributed in the deployment area. The number of internal and external iterations of TaSO are 100 and 10, respectively. Agents may use crossover and mutation techniques to find alternatives to the solution they have. The used mutation is 25% of the solution length and the crossover technique is to cross 50% of the solution. The following are the test cases for the TaSO algorithm.

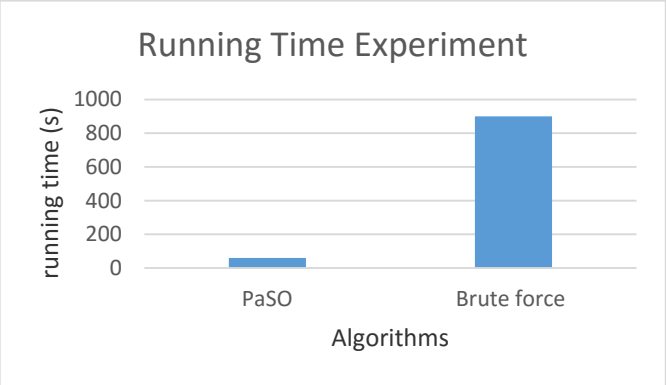

Fig. 3. Running Time Comparison.

A. Running Time Comparison

In this test case, a set of 10 experiments are averages to compare the running time of the TaSO with the Brute force algorithm. As can be seen in Fig. 3, although the number of iterations is 1000 iterations, the difference in time between TaSO and the Brute force trying every combination is very large. In addition, 100 nodes in WSNs are considered as a small network; therefore, it is expected that Brute force running time for large scale network to be huge.

B. Coverage to the Source Node

In this subsection, the coverage percentage of the source nodes from the selected provenance node is examined. As can be seen in Fig. 4, TaSO is covering only 70% of the source nodes; however, with increasing the number of iterations, this percentage has been increased to reach 85% which is good percentage for a greedy algorithm.

C. Closeness to the Sink Node

Here, 10 experiments are averaged showing the average distance of all of the selected provenance nodes to the sink node. The sink node in this case is located at point (0, 0) which is the left top corner of the deployment area. As shown in Fig. 5, compared to the Brute force algorithm, TaSO still satisfying almost 90% of what Brute force algorithm is gaining.

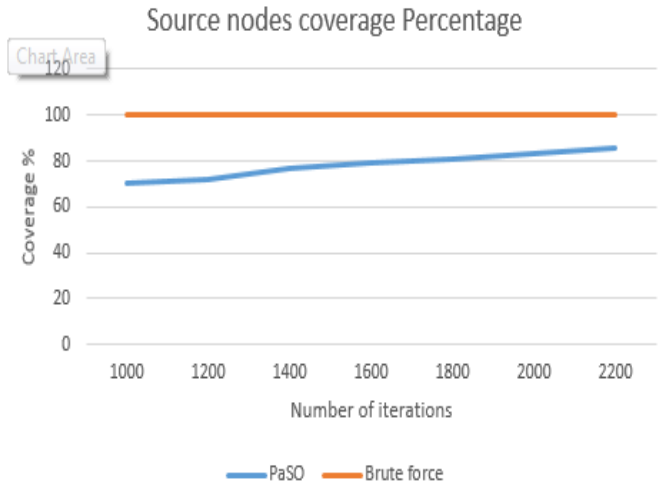

Fig. 4. Source Node Coverage Percentage.

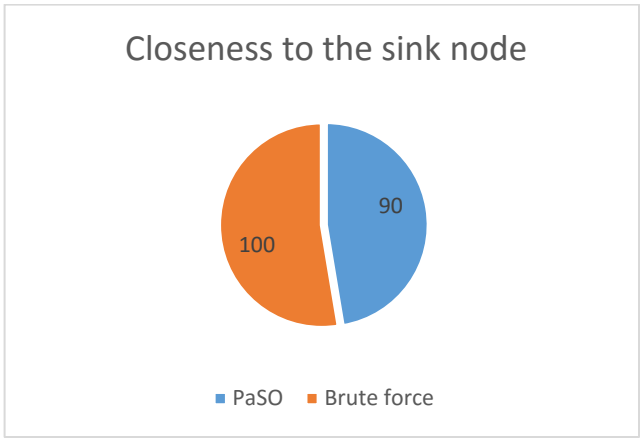

Fig. 5. Closeness to the Sink Node.

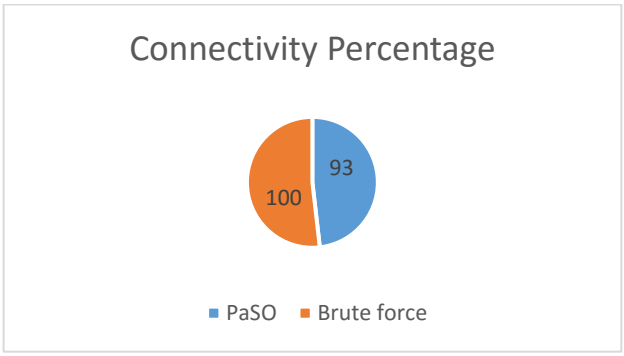

Fig. 6. Connectivity Percentage.

D. Connectivity Percentage

The connectivity percentage, in this context, means the degree of connectivity between the selected provenance nodes out of 100%. This percentage will have a great impact on the connectivity of the overall network. The average of the connectivity percentage, as shown in Fig. 6, of TaSO turns to be 93% with a drop of 7% of the elected nodes with 2200 iterations.

VII. Conclusions and Future Research

We now conclude the paper with a discussion of the potential threats to the validity of research and possible dimensions of the futuristic research. In this paper, we present a bio-inspired technique and it's under think and share optimization algorithm for big data provenance in the context of WSNs. In addition, major applications of data provenance are stated. One of the important examples of big data provenance with many restrictions on the used devices such as memory, storage, and processing capabilities is WSNs. Therefore, this paper focused on the problem of selecting the best nodes to serve as data provenance. Different parameters are considered during the selection process including the connectivity of the selected nodes, within the range of a source node, nodes memory, nodes energy, and trust. The paper proposed a novel algorithm (TaSO) for the nodes' selection process that inspired from human collaboration in education. Moreover, we proposed a technique to solve the node's trust using fuzzy logic with three membership functions which are availability, neighbors' votes, and

message drop rate. The initial results show that the algorithm is efficient in terms of connectivity, closeness to the sink node, coverage, and running time.

7) Potential threats to the validity: Validity threats refer to cases or scenarios in terms of assumptions or constraints that represent some limitations for the proposed solutions. For example, in the proposed solution, limited number of trials for experimental results and validations can be considered as a potential threat. This means that based on more trials with divers case studies and inputs different results may be obtained. If such threats are not highlighted and addressed (as part of extension of future work), they can pose some limitations to proposed solution(s). We have identified following two types of threats:

a) Threat I–Availability of Diverse Data Set for Evaluation: In the proposed solution, we have performed experimental analysis based on comparatively limited amount of available data. Our proposed Futuristic research is also focused on incorporating more data for further validations of the solution. The primary challenge is the acquisition of more data from different systems to measure the quality attributes such as scalability and performance of the systems when new data is provided.

b) Threat II–Data Provenance in an Industrial Context: The second threat is about the applicability and customization of the solution in the context of industry scale data provenance. With an emergence of the Internet of Things (IoTs) and their role in industrial IoTs, can be a significant challenge for data management and provenance. The primary challenge is about the customization and scalability of the proposed solution at industry scale. The proposed algorithm will require appropriate parameterization (for customization) and user intervention (human decision support) for the to develop a solution that can address data provenance issues in larger and practical systems.

8) Dimensions of futuristic research: As part of the future research, we aim to extend the proposed bio-inspired algorithm with applied machine learning approach. The incorporation of machine learning and intelligence can optimize the algorithm (as part of autonomic computing) for efficient data provenance in networked and cyber-physical systems. We are currently in process for analyzing the existing challenges and solution for machine learning in the context of big data provenance. The ultimate solution can be built as an extension to the previous solution with specific focus on provenance intelligence in large, data-intensive systems.

ACKNOWLEDGMENT

This research is funded by the University of Ha'il through the Deanship of Scientific Research under the grant number 'BA- 1513'.

REFERENCES

[1] Taxidou, Io, Sven Lieber, Peter M. Fischer, Tom De Nies, and Ruben Verborgh. "Web-scale provenance reconstruction of implicit information diffusion on social media." Distributed and Parallel Databases 36, no. 1 (2018): 47-79.

[2] Pasquier, Thomas, Jatinder Singh, Julia Powles, David Eyers, Margo Seltzer, and Jean Bacon. "Data provenance to audit compliance with privacy policy in the Internet of Things." Personal and Ubiquitous Computing 22, no. 2 (2018): 333-344.

[3] Hu, Die, Dan Feng, Yulai Xie, Gongming Xu, Xinrui Gu, and Darrell Long. "Efficient Provenance Management via Clustering and Hybrid Storage in Big Data Environments." IEEE Transactions on Big Data (2019).Cuzzocrea, A., "Big Data Provenance: State-Of-The-Art Analysis and Emerging Research Challenges," In EDBT/ICDT Workshops, 2016.

[4] Cuzzocrea, A., "Provenance Research Issues and Challenges in the Big Data Era," In Computer Software and Applications Conference (COMPSAC), 2015 IEEE 39th Annual, Vol. 3, pp. 684-686.

[5] Abbadi, Imad M., and John Lyle. "Challenges for Provenance in Cloud Computing." In TaPP. 2011.

[6] Simmhan, Y.L., Plale, B. and Gannon, D., "A survey of data provenance in e-science," ACM Sigmod Record, 34(3), 2005, pp.31-36.

[7] Zeng, Yu, Xing Zhang, Rizwan Akhtar, and Changda Wang. "A Blockchain-Based Scheme for Secure Data Provenance in Wireless Sensor Networks." In 2018 14th International Conference on Mobile Ad-Hoc and Sensor Networks (MSN), pp. 13-18. IEEE, 2019.

[8] Crawl, D., Wang, J., and Altintas. I., "Provenance for mapreduce-based data-intensive workflows," In WORKS'11, Proceedings of the 6th Workshop on Workflows in Support of Large-Scale Science, co-located with , SC11, Seattle, WA, USA, 2011, pp. 21–30.

[9] Bhardwaj, V., & Johari, R., "Big data analysis: Issues and challenges," International Conference on Electrical, Electronics, Signals, Communication and Optimization (EESCO), 2015 pp. 1-6

[10] Arshad, B., "NeuroProv-A visualisation system to enhance the utility of provenance Data for neuroimaging analysis" (Doctoral dissertation, University of the West of England), 2015.

[11] Hammad, R. and Wu, C., "Provenance as a service: A data-centric approach for real-time monitoring," In 2014 IEEE International Congress on Big Data, Anchorage, AK, USA, 2014, pp. 258–265.

[12] Gundecha, P., Ranganath, S., Feng, Z., & Liu, H., "A tool for collecting provenance data in social media," In Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, 2013, pp. 1462-1465.

[13] Chen, P., & Plale, B. A., "Big data provenance analysis and visualization," International Symposium on Cluster, Cloud and Grid Computing (CCGrid), 15th IEEE/ACM, 2015 (pp. 797-800).

[14] Liu, Y. and Xu, S., "Detecting rumors through modeling information propagation networks in a social media environment," IEEE Transactions on Computational Social Systems, 3(2), pp.46-62, 2016.

[15] Bertino, E., Ghinita, G., Kantarcioglu, M., Nguyen, D., Park, J., Sandhu, R., & Xu, S., "A roadmap for privacy-enhanced secure data provenance," Journal of Intelligent Information Systems, 43(3), 481-501, 2014.

[16] Karvounarakis, G., Ives, Z. G., & Tannen, V., "Querying data provenance," In Proceedings of the 2010 ACM SIGMOD International Conference on Management of data ACM, 2010, pp. 951-962.

[17] McDaniel, P., Butler, K., McLaughlin, S., Sion, R., Zadok, E., Winslett, M. (2010). In 2nd USENIX workshop on the theory and practice of provenance.

[18] Lyle, J., & Martin, A., In 2nd USENIX workshop on the theory and practice of provenance, 2010.

[19] Qazvinian, V., Rosengren, E., Radev, D.R. and Mei, Q., "Rumor has it: Identifying misinformation in microblogs." In Proceedings of the Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2011, pp. 1589-1599.

[20] Taxidou, I., De Nies, T., Verborgh, R., Fischer, P.M., Mannens, E. and Van de Walle, R., "Modeling information diffusion in social media as provenance with W3C PROV," In Proceedings of the 24th International Conference on World Wide Web, ACM, 2015, pp. 819-824.

[21] De Nies, T., Taxidou, I., Dimou, A., Verborgh, R., Fischer, P.M., Mannens, E. and Van de Walle, R., "Towards multi-level provenance reconstruction of information diffusion on social media," In Proceedings

of the 24th ACM International on Conference on Information and Knowledge Management ACM, 2015, pp. 1823-1826.

[22] Gundecha, P., Ranganath, S., Feng, Z. and Liu, H., "A tool for collecting provenance data in social media," In Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, 2013, pp. 1462-1465.

[23] De Nies, Tom, et al, "Towards multi-level provenance reconstruction of information diffusion on social media," Proceedings of the 24th ACM International on Conference on Information and Knowledge Management, ACM, 2015.

[24] Katilu, Victoria M., Virginia NL Franqueira, and Olga Angelopoulou, "Challenges of Data Provenance for Cloud Forensic Investigations." Availability, Reliability and Security (ARES), 2015 10th International Conference on. IEEE, 2015.

[25] Muhammad Imran and Helmut Hlavacs, "Provenance Framework for the Cloud Infrastructure: Why and How?" International Journal on Advances in Intelligent Systems, vol 6 no 1 & 2, 2013.

[26] Hammad, R. and Wu, C., "Provenance as a service: A data-centric approach for real-time monitoring," In 2014 IEEE International Congress on Big Data, Anchorage, AK, USA, 2014, pp. 258–265.

[27] Wang, B., and Li, H., Oruta, "Privacy-preserving public auditingfor shared data in the cloud," IEEE T. Cloud Computing, 2014, 2(1):43–56.

[28] Davidson, S. B. and J. Freire, "Provenance and scientific workflows: challenges and opportunities," In Proceedings of the ACM SIGMOD International Conference on Management of Data, SIGMOD 2008, Vancouver, BC, Canada, 10-12, 2008, pp. 1345–1350.

[29] Muniswamy-Reddy, K., Macko, P., and Seltzer, M. I., "Provenance for the cloud," In USENIX Conference on File and Storage Technologies, 2010, (FAST), volume 10, pp. 15–14.

[30] Wang, Changda, Wenyi Zheng, and Elisa Bertino. "Provenance for Wireless Sensor Networks: A Survey," Data Science and Engineering 1.3, 2016, pp. 189-200.

[31] Priyanka, 2M.Devika, 2016, "A Survey of provenance management in wireless sensor network," nt. Journal of Engineering Research and Applications www.ijera.com ISSN: 2248-9622, Vol. 6, Issue 1, (Part - 5) 2016, pp.91-93.

[32] Liu, L. (2009). Encyclopedia of database systems (Vol. 6). M. T. Özsu (Ed.). New York, NY, USA:: Springer.

[33] Spiekermann, R., Jolly, B., Herzig, A., Burleigh, T., & Medyckyj-Scott, D. (2019). Implementations of fine-grained automated data provenance to support transparent environmental modelling. Environmental Modelling & Software.

[34] Sajjad, Maryam, Aakash Ahmad Abbasi, Asad Malik, Ahmed B. Altamimi, and Ibrahim Mohammed Alseadoon. "Classification and mapping of adaptive security for mobile computing." IEEE Transactions on Emerging Topics in Computing , 2018.

[35] A. Shadi, M. B. Yassein. A Resource-efficient Encryption Algorithm for Multimedia Big Data. In Multimedia Tools and Applications vol. 76, no. 21, pp: 22703-22724, 2017.