International Journal of Grid and Distributed Computing
Vol. 10, No. 1 (2017), pp.269-282
http://dx.doi.org/10.14257/ijgdc.2017.10.1.23

An Efficient and Secure Data Aggregation Scheme in Smart Grid

Yiwen Le* and Jinghan He

School of Electrical Engineering, Beijing Jiaotong University
leyiwen13117376@sina.com, jhhe@bjtu.edu.cn

Abstract

In order to deal with the massive, heterogeneous and distributed power data, it is urgent to use new information technology to process and manage power data in smart grid. In this paper, an efficient and secure data aggregation scheme in smart grid is proposed (ESDAS). This paper first models the distributed and hierarchical data aggregation architecture of smart grid by utilizing mathematical method. In order to find the optimal aggregator placement, a distributed cost update algorithm is also proposed. By introducing the expired timer parameter considering link states, the proposed algorithm can significantly reduce the communication cost of data aggregation. Finally, for the purpose of dealing with internal false data injection attacks, the paper presents an effective solution by establishing rule matching set of users' data. We show with simulation that ESDAS can improve the efficiency and security of smart grid.

Keywords*: Smart Grid; Data Aggregation; aggregator placement; rule matching set.*

1. Introduction

In recent years, with the advent of the Internet plus era, internet technologies and innovative thinking have been deeply integrated into various industries. As an important part of the energy internet, the smart grid is the focus of the Internet thinking and ideas in the energy field.

The typical characteristic of traditional power grid is the centralized one-way transmission (fixed transmission direction) and demand-driven response [1]. Compared with it, smart grid allows distributed two-way flows and introduces new technologies and solutions (e.g., Internet of things) for intelligent power generation, transmission, substation, distribution, utilization [2-3]. Consequently, the concept of smart grid has attracted more and more attention and been considered as the future development direction of the next generation of power grid [4-6].

The use of smart grid systems can bring significant benefits, including promoting energy efficiency and management reliability, reducing carbon emissions [7], improving the adaptability of wide-ranging conditions and unexpected events, etc. These benefits stem from the large-scale deployment of intelligent sensing devices, such as various sensors and smart meters. With intelligent sensing devices, smart grid has the ability to collect real-time operating parameters and data information. Thus, the intelligent sensing devices and their recorded data are essential in smart grid.

The power data in smart grid has the following crucial features:

- Mass-data processing. In order to capture the required data, sensors and other sensing devices need to be located in all aspects of power production and power use, which will produce a large amount of data information.
- Diversity of data type. According to different application environments and requirements, power systems will generate a wide variety of data types.

* Yiwen Le is the corresponding author.

ISSN: 2005-4262 IJGDC

- Distributed data acquisition. Power data are derived from power users, transmission and distribution systems, power plants, and growing distributed energy generation and storage systems, which means that the data in power system requires distributed processing.
- High accuracy and security requirements. Due to the significance of power systems to social production and human life, the power data should be preserved in high accuracy and security, which is mainly reflected in two aspects: on the one hand the acquisition of power data requires maintaining high accuracy and reliability, on the other hand, the power system should have capability to resist against malicious attacks, such as false data injection and system intrusion.

In this case, data aggregation (fusion or integration) technology is a natural choice to meet the demand of characteristics of power data [8]. However, most of data aggregation schemes in existing networks (e.g., wireless sensor networks, ad-hoc networks) are not applicable for smart grid, which is mainly reflected in three aspects. Firstly, as the architecture of some traditional data aggregation mechanism is centralized [9], it cannot satisfy the different requirements of distributed information flow in smart grid. Secondly, most of the existing data aggregation schemes focus on the complexity of the fusion algorithm, but not pay enough attention to the additional communication overhead caused by the process of data aggregation. Finally, not much work has studied the impact of network security (especially the false data injection attack) on data aggregation mechanism.

In order to solve the above problems in existing schemes, this paper mainly focuses on the issues of data aggregation of HANs (Home Area Networks) in smart grid. According to the limitations of existing data aggregation mechanism, we propose an efficient and secure data aggregation scheme (ESDAS) in mart grid. The main contributions of this paper are summarized as follows:

- Firstly, we describe the whole network architecture of smart grid. According to its characteristics, we also model the distributed data aggregation architecture of HANs in smart grid by utilizing the mathematical modeling called *semiring* [10].
- Secondly, based on the distributed data aggregation architecture, this paper analyzes the overall cost of data aggregation in a HAN, and presents a distributed cost update algorithm of ESDAS to find the optimal aggregator placement in smart grid. Compared with the existing typical schemes, our proposed algorithm can significantly reduce the communication overhead.
- Finally, by introducing the typical feature model of users' data and rule matching set, this paper gives a specific method to improve the ability to resist against internal false data injection attacks.

This paper is organized as follows. We first present some relevant works in Section II. Section III describes our system model which includes the network model and the security model. Section IV gives the distributed architecture and data aggregation strategy, followed by the aggregator placement algorithm in smart grid in Section V. Section VI discusses the countermeasure to the internal false data injection attack. The simulation results are presented in Section VII. Section VIII draws conclusions and discusses some future works.

2. Related Work

The existing data aggregation network architecture in smart grid can be divided into two categories: the centralized data aggregation architecture and the distributed data aggregation architecture [11]. The centralized data aggregation architecture requires that each node in the network should know the entire network topology and state information, and a large number of network topology maintenance overhead will limit its application in large-scale network [9].

Based on neural network algorithm, the thesis [12] presented a data aggregation technology for power network system, which can fuse the data generated by various sources in the power system. However, this scheme is too complex to be applied to distributed network environments. Chen *et al* proposed a novel multi-functional data fusion mechanism called MuDA [13]. By adopting MuDA, the data control center of smart grid can compute a variety of statistical functions of user data. But the mechanism mainly focuses on the data aggregation process in the operation and data center, and the data aggregation problem of HANs is not fully considered.

Bonfils *et al* proposed a distributed data aggregation node location algorithm by using the information discovery and adaptive mechanism of neighbor nodes [3]. The algorithm can improve the position of aggregators gradually, and realize the optimization of the network data aggregation structure. Similarly, in order to optimize the data aggregation process, a distributed algorithm based on tree structure was proposed in [14]. This algorithm can query the data aggregation nodes in the network through interactive information. But it requires realizing the whole network clock synchronization, which has strong application limits. Lu *et al* proposed an asynchronous algorithm for tree-structured data aggregation in smart grid [15]. They focused on the issue of communication overhead produced by ADV broadcasting packets in the asynchronous data aggregation process. However, they did not consider the network deployment cost [16]. Furthermore, the effect of link state on the scheme was also not considered.

In recent years, security problems in smart grid have always been the focus of the research [17-19]. A data fusion mechanism in smart grid with high efficiency and privacy protection was proposed in [1]. This mechanism uses a super-increasing sequence to structure multi-dimensional data, and encrypt the structured data by the homomorphic cryptosystem technique. Rottondi *et al* presented a secure architecture for data aggregation in smart meters, which provided the encrypted communications capability to prevent the gateway and the external entity from inferring user information from independent data [20]. However, the complex asymmetric encryption algorithm may bring additional computation and communication overhead to the intelligent sensing terminals (such as sensors and smart meter devices). In addition, the above mechanisms based on cryptographic primitives only aim to protect the confidentiality of the data and the user's privacy. Once the network is compromised, these mechanisms may fail to resist against the damage caused by internal false data injection attack. In other words, an attacker can inject a large number of false data into data aggregation process to achieve the purpose of destroying the normal operation of smart grid.

3. System Model

3.1. Network Model

As shown in Figure 1, we briefly consider smart grid consisting of power generation (power plant), power transmission, power distribution [21], and HANs. HAN is a core component of smart grid, which is comprised of various smart meters and sensors. In a HAN, wired and wireless communication is utilized to support the two-way information flow between the intelligent sensing devices. Each intelligent sensing device is in charge of collecting real-time measurements of energy consumption, production and/or storage, and then reports a huge amount of data to the operation and data center. As all the intelligent sensing devices are resource-constrained, data aggregation mechanisms will be applied in the process of data transmission. With the aggregated information, the operation and data center can automatically and timely monitor grid status, balance electricity load, maintain system operation, optimize energy consumption [22]. In this paper, we mainly focus on the data aggregation process in HANs.

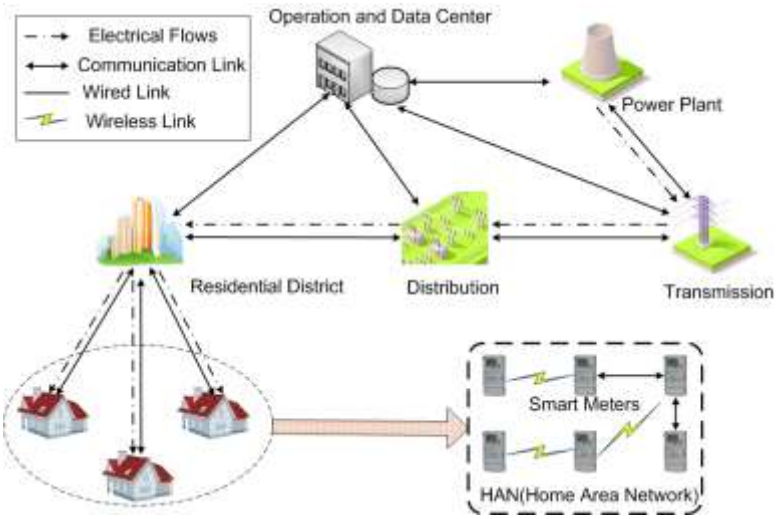

Figure 1. Network Architecture of Smart Grid

3.2. Security Model

As smart grid is largely dependent on the legacy systems which are not initially designed to be high security consideration, guaranteeing the security of smart grid is a challenging task. With the open and distributed deployment environment, we assume that smart meters and sensors in HANs are generally vulnerable to various attacks, such as blackhole attack, wormhole attack, sybil attack and false data injection attack [23]. Compared with the sensing devices, the operation and data center can be recognized as a trusted entity that can resist against common attacks. As described in chapter II, we mainly consider the countermeasures against the internal false data injection attack in this paper.

4. Distributed Architecture and Data Aggregation Strategy

In this section，we first describe the data aggregation architecture of HANs in smart grid by utilizing the mathematical modeling. For a weighted graph $G_H(V,E,\omega)$, G_H denotes the set of HANs, the set of vertices V stands for the sensing devices in HANs. $E \subseteq V \times V$ represents the communication links between sensing devices. The weighted label ω stands for the communication overhead of the corresponding links. Given the data aggregation tree of G_H as the directed graph $G_T(C,D,\eta)$, C denotes the set of communication dependences connecting the data aggregation services and D represents the set of data objects. For any data object d_k, $d_k \in D$. η is utilized for measuring the size of the data object.

In the data aggregation tree, as described in Figure 2, each aggregator can have one or more child nodes but only one parent node. The root node of the data aggregation tree is the operation and data center (n_1). Hence, the data objects sent by child nodes generate the data objects produced by parent nodes.

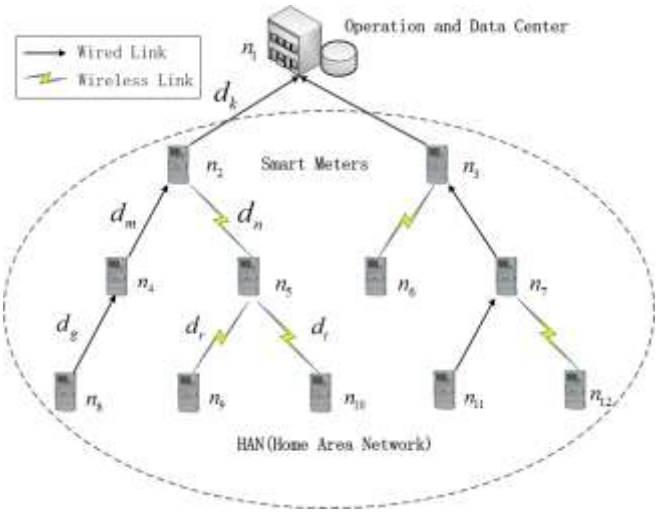

Figure 2. Diagram of Data Aggregation Tree

Based on the mathematical model, we then propose a distributed and privacy-preserving data aggregation strategy by introducing a mathematical theory called *semiring*. The rigorous mathematical proof of *semiring* was depicted in [10, 24].

Definition 1. A *semiring* is an algebraic structure $(Q,\oplus,\otimes,0,1,\prec)$, where Q is a set. $\oplus$, $\otimes$, and $\prec$ are operators with the following properties.

$\oplus$ is commutative and associative. For $\oplus$, 0 is a neutral element.

$$\begin{cases} a\oplus b=b\oplus a, \\ (a\oplus b)\oplus c=a\oplus(b\oplus c), \\ a\oplus 0=a. \end{cases} \tag{1}$$

$\otimes$ is associative. For $\otimes$, 0 is an absorbing element and 1 is a neutral element.

$$\begin{cases} (a\otimes b)\otimes c=a\otimes(b\otimes c), \\ a\otimes 0=0, \\ a\otimes 1=a. \end{cases} \tag{2}$$

$\prec$ is an order relation with respect to the operators.

$$\begin{cases} \forall a,b\in Q: a\prec b, c\prec d \Rightarrow a\oplus c\prec b\oplus d \\ \forall a,b\in Q: a\prec b, c\prec d \Rightarrow a\otimes c\prec b\otimes d \\ \forall a,b\in Q: a\prec b \Leftrightarrow \exists c\in Q: a\oplus c=b \end{cases} \tag{3}$$

By utilizing the theory of *semiring*, we can easily describe the data aggregation process. For example, the aggregated data object d_k in Figure 2 can be expressed as follows:

$$\begin{aligned} d_k &= [(d_g\otimes LS(n_8,n_4))\oplus(0\otimes LS(\varnothing,n_4))]\otimes LS(n_4,n_2) \\ &\quad\oplus[(d_r\otimes LS(n_9,n_5))\oplus(d_t\otimes LS(n_{10},n_5))]\otimes LS(n_5,n_2) \\ d_g &\in \Omega(d_m), d_r, d_t\in\Omega(d_n), d_m, d_n\in\Omega(d_k) \end{aligned} \tag{4}$$

Where Ω denotes the relative children data objects, LS denotes the local link impact factors, such as the asymmetric encryption or packet loss and so on. $\otimes$ represents an operator to concatenate the transmission of data objects along a communication path, while $\oplus$ stands for the data aggregation operations.

More specifically, the electricity usage data $(D_i = d_{i1,} d_{i2}, d_{i3}, \cdots, d_{ik})$ collected by intelligent sensing devices is first encrypted by user's private key pk_i, that is, $pk_i \cdot H(\mathrm{D}_i \| \mathrm{G}_T \| \mathrm{U}_i \| \mathrm{T}_s)$, where U_i stands for user i in a HAN and T_s is the timestamp. When receiving the encrypted data, the aggregators can verify its validity, and then perform the aggregation operations. Finally, all the aggregated data should be reported to the operation and data center to support the decision-making and the performance of smart grid. The discussion of encryption algorithm and details of data aggregation operations are out of the scope of this paper.

5. Aggregator Placement in Smart Grid

In ESDAS，the location selection of aggregators is a crucial issue to be solved. This section aims to find the optimal aggregator placement in smart grid with the minimum cost of data aggregation. In consideration of link states, we also try to propose a method to minimize the communication overhead as much as possible in the interaction process of data aggregation.

The overall cost of data aggregation in a HAN can be divided into three parts: communication cost, computation cost, and deployment cost. The communication cost of a HAN can be denoted as: $\omega(P, G_H, G_T)$, where P denotes the placement of the aggregators. The computation cost of a HAN can be denoted as $\Upsilon(P, G_H, G_T)$. The deployment cost is mainly from the wired communication links. In this paper, we assume that all wired communication links use optical fiber cables. Hence, the cost of connecting node n_i with node n_j by optical fibers is u_{ij}, which is related to their distance s_{ij}, that is, $u_{ij} = \Lambda(\mathrm{s}_{ij})$, where $u_{ij} = 0$ indicates that there is no wired communication link between node n_i and node n_j. Finally, the overall cost of data aggregation in a HAN can be denoted as:

$$f_{\mathrm{cost}}(P, G_H, G_T) = \omega(P, G_H, G_T) + \Upsilon(P, G_H, G_T) + \sum_{n_i, \mathrm{n}_j \in V, i \neq j} \Lambda(\mathrm{s}_{ij}) \tag{5}$$

Thus, to sum up the problem becomes to find the optimal P^* with the minimum cost of data aggregation in a HAN. In this section, we give a distributed cost update algorithm to solve the optimization problem inspired by Ying *et al* [14, 15].

In our algorithm, each node in HANs should maintain a local minimum cost list of data aggregation, which can be updated by the interactions of advertisement (ADV) packets. In the local minimum cost list, $\mathrm{L}(d_k, n_i)$ stands for the current lowest cost of acquiring data object d_k at node n_i. $P(d_k, n_i) = 1$ represents the data object d_k is generated at node n_i. Otherwise, $P(d_k, n_i) = 0$. The algorithm of ESDAS is shown in Algorithm 1.

At the initialization stage of ESDAS, for any $d_k \in D, n_i \in V$, all $\mathrm{L}(d_k, n_i)$ should be

set to ∞ (step 1 - 2). If $P(d_k, n_i)$ is equal to 1, node n_i computes and updates the cost value if it can get a lower acquiring cost (step 3 - 4). Furthermore, if the initial $\mathrm{L}(d_k, n_i)$ is equal to ∞, node n_i should broadcast the ADV packet containing the data object and the acquiring cost immediately after updating the cost value (step 5 - 7).

When node n_i receives the ADV packet from node n_j for data object d_k, it computes the acquiring cost and updates the cost value if the latest cost is the smaller one (step 10 - 13). Then, our algorithm also updates the acquiring cost for the ancestors of data object d_k (step 16 - 19).

By utilizing the algorithm above, we can obtain the optimal placement P^* of aggregators when there are no ADV packets exchanged in the network. At the same time we can also get the optimal data aggregation tree G_T^* consisting of all the optimal placement of aggregators.

Algorithm 1. Distributed Cost Update Algorithm

1: Process Initialization
2: $\forall d_k \in D, n_i \in V, L(d_k, n_i) = \infty$
3: **if** $P(d_k, n_i) == 1$ **then**
4: $L'(d_k, n_i) = \sum_{d_m \in \Omega(d_k)} L(d_m, n_i) + \Upsilon(d_k, n_i)$
5: **if** $L'(d_k, n_i) < L(d_k, n_i) \& \& L(d_k, n_i) == \infty$ **then**
6: $L(d_k, n_i) = L'(d_k, n_i)$
7: Broadcast the ADV packet containing $L(d_k, n_i)$ immediately
8: **end if**
9: **end if**
10: Node n_i receives the ADV packet from node n_j for the data object d_k
11: $L'(d_k, n_i) = L(d_k, n_j) + \omega(n_i, n_j) + \sum_{n_i, n_j \in V, i \neq j} \Lambda(S_{ij})$
12: **if** $L'(d_k, n_i) < L(d_k, n_i)$ **then**
13: $L(d_k, n_i) = L'(d_k, n_i)$
14: Set the timer of d_k to T_e
15: **end if**
16: **for** Each ancestor d'_k of d_k **do**
17: $L'(d'_k, n_i) = \sum_{d'_m \in \Omega(d'_k)} L(d'_m, n_i) + \Upsilon(d'_k, n_i)$
18: **if** $L'(d'_k, n_i) < L(d'_k, n_i) \& \& L(d'_k, n_i) \neq \infty$ **then**
19: $L(d'_k, n_i) = L'(d'_k, n_i)$
20: Set the timer of d'_k to T_e
21: **end if**
22: **end for**
23: **while** The timer of node n_i for the data object d_k expires **do**
24: Broadcast the ADV packet containing $L(d_k, n_i)$
25: **end while**
26: END Process

However, the overhead of asynchronous algorithm mainly depends on the broadcasting ADV packets. If the nodes broadcast ADV packets immediately after obtaining the lower acquiring cost, it will produce much communication overhead. In order to mitigate this problem, Lu etc. set a delay time to be proportional to the transmission cost [17]. But they did not consider the packet loss issue caused by different link quality. In ESDAS, the

expired timer T_e is given by:

$$T_e = \frac{\lambda}{1-[1-P_s(\mathrm{n}_i, \mathrm{n}_j)]^N} \cdot w(\mathrm{n}_i, \mathrm{n}_j) \tag{6}$$

Where $P_s(\mathrm{n}_i, \mathrm{n}_j)$ is the probability of packet success transmission in $e(\mathrm{n}_i, \mathrm{n}_j)$. Thus, $1-[1-P_s(\mathrm{n}_i, \mathrm{n}_j)]^N$ represents the probability of at least one packet can be successfully transmitted in N times. In Algorithm 1, we set the expired timer of the data object to T_e. When the timer expires, the node should broadcast the ADV packet immediately, otherwise, it should keep silence.

6. Countermeasure to False Data Injection Attacks

As described in chapter II, existing security mechanisms have considered many attacks on smart grid. However, there is no effective method to deal with the false data injection attack launched by the malicious nodes in the network. Such attack obtains benefits (such as modifying the use data of electricity) or destroys the stable operation of smart grid (such as providing the wrong parameters) by tampering with the data.

In order to solve the problem, we enhanced the ability to resist against the internal false data injection attacks by establishing the typical feature model of users' data and rule matching set. The typical feature model of users' data can be given as follows.

$$\mathrm{F}(D_i) = \{f_1(D_i), f_2(D_i), \cdots, f_x(D_i)\} \tag{7}$$

Assume that the data of user i contains x different behavior characteristics which is denoted by $f_1, f_2, \cdots, f_x$. Accordingly, our scheme defines a rule matching set which are derived from the characteristics of a series of normal data.

$$R(D_i) = \{r_1(D_i), r_2(D_i), \cdots, r_x(D_i)\} \tag{8}$$

Where $r_1(D_i), r_2(D_i), \cdots, r_x(D_i)$ map the rules of normal behavior characteristics for user i. In our scheme, the decision procedure of false data injection attack can be expressed as:

$$\hbar(D_i) = \begin{cases} 0, & \text{if} \quad \mathrm{M}(f_1(D_i), f_2(D_i), \cdots, f_x(D_i)) \wedge \mathrm{R}(\mathrm{D}_i) \\ 1, & \textit{otherwise} \end{cases} \tag{9}$$

$\hbar(D_{ik})$ stands for the decision result of false data injection attack for the data of user i. The result 0 indicates that the data is normal, that is, all the characteristics of user's data meet the rules matching set. Otherwise the result is 1. $\mathrm{M}(\cdot)$ represents the mathematical mapping function, such as weighted average function, weighted variance function, one-way ANOVA calculation [12], etc.. To sum up, by utilizing the mechanism above, we can effectively detect the false injection data and adopt corresponding measures (such as filter the false data) to protect the secure operation of smart grid.

7. Simulation Results and Performance Evaluation

In this section, we evaluate our proposed data aggregation scheme in smart grid (ESDAS) by using NS-2 simulator [25] and MATLAB, which is used for network

performance analysis and algorithm implementation, respectively. All the default simulation parameters that we have chosen are summarized in Table 1.

The simulations can be divided into three parts. First, we investigate the selection of the expired timer with different probability of packet success transmission. Second, compared with the sole asynchronous scheme and MCFA scheme [15], the performance merits and defects of ESDAS are analyzed. At last, we verify the ability of ESDAS to resist against the internal false data injection attacks.

Table 1. Simulation Parameters

Parameters	Default Values
Deployment area	1000×1000 m^2
Number of nodes	100
Communication range	200 m
Data packet interval	5s
Length of data packet	100 bytes

7.1. Selection of Expired Timer

The expired timer T_e is the key parameter in ESDAS. In the simulation, we model a network consisting of 100 sensing devices distributed in a 1000 × 1000m^2 area. Then, as is depicted in Equation (6), we list common value λ to analyze the performance of data aggregation process with various probability of packet success transmission P_s. As shown in Figure 3, regardless of P_s, the number of ADV packets decreases as the value λ increase. This is because the high value λ we set, the longer expired timer of ADV packets it is. The phenomenon that a sensing device broadcasts an ADV packet more than once will reduce. However, when we change P_s of communication links, we find that the drop of P_s will lead to a significant increase in the number of ADV packets. This is because packet loss will result in retransmission of packets.

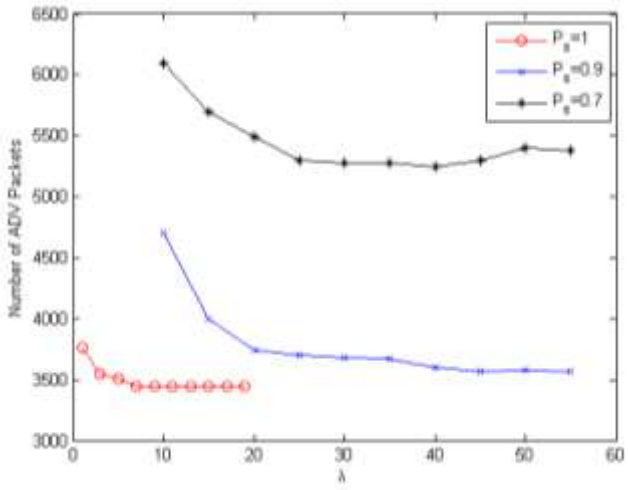

Figure 3. Impact of λ on ADV Packets with Various P_s

As shown in Figure 4, with the increasing value λ, the aggregator placement setup time decreases at first, and then increases quickly after reaching the minimum value.

When the value λ is small, a large number of broadcasting ADV packets may cause network congestion. This problem gradually disappears as the value λ increases. When the value λ is greater than a certain threshold, the setup time is proportional to λ. Considering the results of Figure 3 and Figure 4, we choose the threshold as the optimal value λ (when $P_s = 1$, $\lambda = 7$, when $P_s = 0.9$, $\lambda = 20$, when $P_s = 0.7$, $\lambda = 26$).

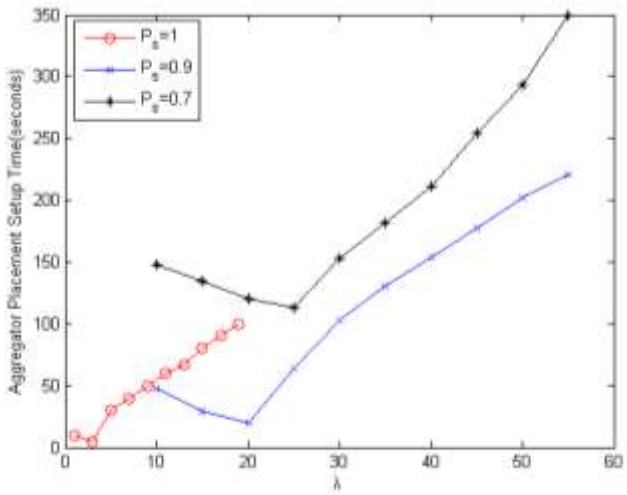

Figure 4. Impact of λ on Setup Time with Various P_s

7.2. Performance Comparison and Analysis

In order to conduct performance analysis, we compare our proposed scheme ESDAS with the sole asynchronous scheme and MCFA. In the simulations, we use the optimal λ which is chosen in the last subsection.

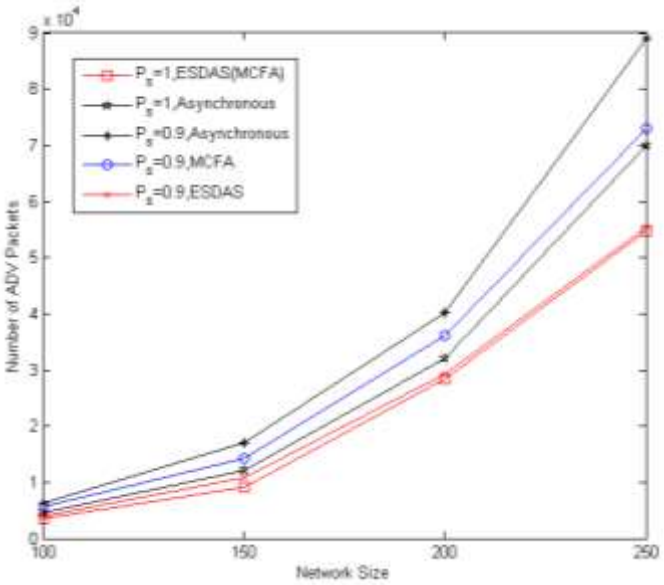

Figure 5. Comparison of Number of ADV Packets

As is illustrated in Figure 5, with the increasing scale of the network, the ADV packets produced by any kind of data aggregation method will increase significantly. When

$P_s = 1$, the ADV packets produced by ESDAS is the same as the ones produced by MCFA, because these two methods use the same setting of expired timer in this case. As the sole asynchronous scheme does not consider the expired timer parameter, it will produce much more communication overhead. When $P_s = 0.9$, the ADV packets produced by MCFA and the sole asynchronous scheme increase significantly (always more than 30%) as they do not consider the effect of link state and packet loss. We can also find that the extra network overhead produced by ESDAS can almost be ignored (always less than 5%) in this case.

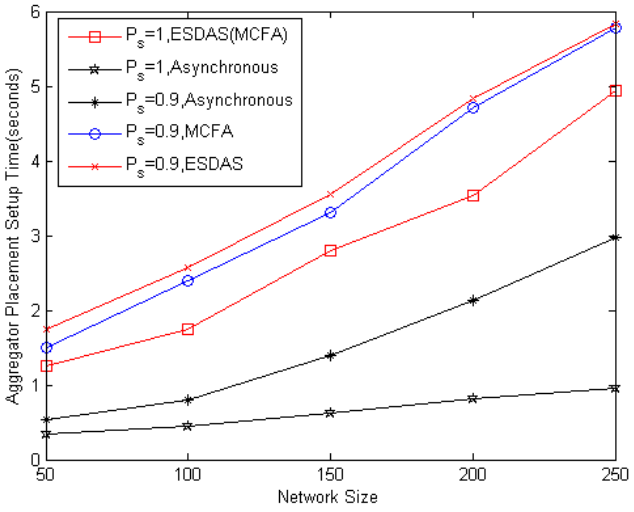

Figure 6. Comparison of Setup Time

As shown in Figure 6, the setup time of any kind of method will grow with the increasing network size. The sole asynchronous scheme does not consider the problem of reducing the network overhead. Once the cost of acquiring data object updates, the node in the network will broadcast the ADV packets. Consequently, it has the shortest setup time. On the other hand, ESDAS and MCFA have similar setup time values as they both adopt the expired timer mechanism.

7.3. Network Security Analysis

In order to analyze the ability of ESDAS to resist against the internal false data injection attacks, we introduce the malicious sensing devices that always launch false data injection attacks in the simulated scenarios. For simplifying the analysis process, we choose the weighted average function and weighted variance function as the mapping function. And then we can obtain the rule matching threshold.

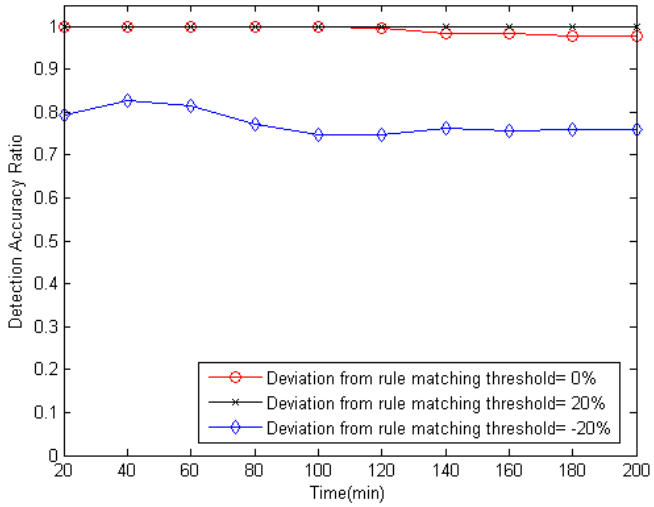

Figure 7. Detection Accuracy Ratio Analysis

As is illustrated in Figure 7 and Figure 8, when the percentage of the deviation from rule matching threshold is 20% (choose normal data features as the positive direction), the detection accuracy ratio of false data injection attacks is close to 100%, but the false alarm ratio is as high as 15%. This is because when the rule matching threshold trend to the direction of the normal data characteristics, more normal data will be adjudged to the abnormal one. In contrast, when the percentage of the deviation from rule matching threshold is -20%, the false alarm ratio of false data injection attacks is close to 0%, but the detection accuracy ratio is about 75%. Thus, the simulation results show that ESDAS has a high ability to resist against the internal false data injection attacks as it has a relatively high detection accuracy ratio and low false alarm ratio.

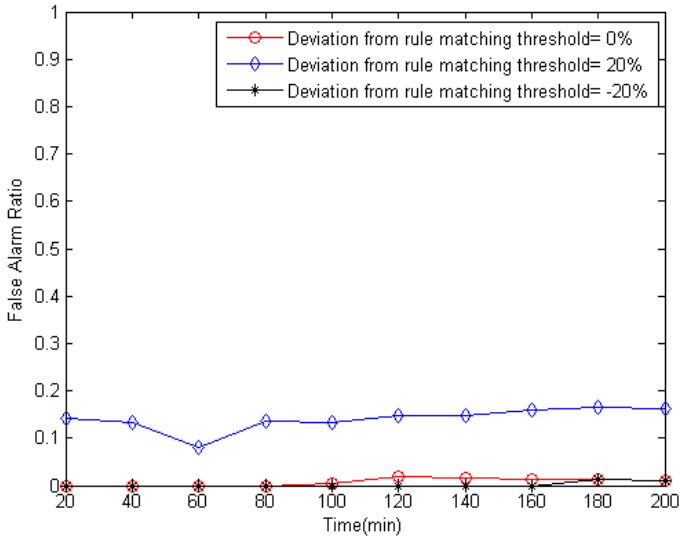

Figure 8. False Alarm Ratio Analysis

8. Conclusion and Future Work

In this paper, we proposed a novel data aggregation scheme in smart grid. First, we modeled the distributed data aggregation architecture of HANs in smart grid. Based on the data aggregation architecture, we then presented a distributed cost update algorithm to

find the optimal aggregator placement in smart grid. By introducing the mathematical theory called *semiring* and the expired timer considering link state and packet loss, the scheme can significantly reduce the network communication overhead produced by data aggregation. Finally, the countermeasure to false data injection attacks was proposed. It can provide a high ability to resist against internal false data injection attacks by utilizing the rule matching set. The simulation results show that our scheme can satisfy requirements of efficiency and security in smart grid. In the future, we will focus other security issues of data aggregation in smart grid, and how to use the sensing data to improve the stable and efficient operations of smart grid.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (NSFC) [grant number 51277009].

References

[1] R. Lu, X. Liang and X. Li, "EPPA: An efficient and privacy-preserving aggregation scheme for secure smart grid communications", IEEE Transactions on Parallel and Distributed Systems, vol. 23, no. 9, **(2012)**, pp. 1621-1631.
[2] M. Donohoe, B. Jennings and S. Balasubramaniam, "Context-awareness and the smart grid: Requirements and challenges", Computer Networks, vol. 79, no. 1, **(2015)**, pp. 263-282.
[3] L. Chen, R. Lu and Z. Cao, "MuDA: Multifunctional data aggregation in privacy-preserving smart grid communications", Peer-to-Peer Networking and Applications, vol. 2014, no. 1, **(2014)**, pp. 1-16.
[4] K. Moslehi and R. Kumar, "A reliability perspective of the smart grid", IEEE Transactions on Smart Grid, vol. 1, no. 1, **(2010)**, pp. 57-64.
[5] Z. M. Fadlullah, M. M. Fouda and N. Kato, "Toward intelligent machine-to-machine communications in smart grid", IEEE Communications Magazine, vol. 49, no. 4, **(2010)**, pp. 60-65.
[6] H. Liang, B. J. Choi and W. Zhuang, "Towards optimal energy store-carry-and-deliver for PHEVs via V2G system", Proceedings of the IEEE INFOCOM, **(2012)**; Orlando, FL, USA.
[7] K. Yagnik, S. Vadhva and R. Tatro, "California Smart Grid Attributes: California Public Utility Commission Metrics", Proceedings of the IEEE Green Technologies Conference, **(2011)**; Baton Rouge, Louisiana, USA.
[8] E. Ancillotti, R. Bruno and M. Conti, "The role of communication systems in smart grids: Architectures, technical solutions and research challenges", Computer Communications, vol. 36, no. 17, **(2013)**, pp. 1665-1697.
[9] A. Nechifor, M. Albu and R. Hair, "A flexible platform for synchronized measurements, data aggregation and information retrieval", Electric Power Systems Research, vol. 120, no. 1, **(2015)**, pp. 20-31.
[10] C. Zhang, X. Zhu, Y. Song and Y. Fang, "A formal study of trust based routing in wireless ad hoc networks", Proceedings of the IEEE INFOCOM, **(2010)**; San Diego, California, USA.
[11] J. Zhou, R. Q. Hu and Y. Qian, "Scalable distributed communication architectures to support advanced metering infrastructure in smart grid", IEEE Transactions on Parallel and Distributed Systems, vol. 23, no. 9, **(2012)**, pp. 1632-1642.
[12] X. Cheng, Y. Ni and Y. Wang, "The Application of Data Fusion Technology Based on Neural Network in the Dynamic Risk Assessment", Physics Procedia, vol. 25, no. 1, **(2012)**, pp. 1696-1700.
[13] B. Bonfils and P. Bonnet, "Adaptive and decentralized operator placement for in-network query processing", Proceedings of the 2nd Int. Conf. Inf. Process.Sensor Netw., **(2003)**.
[14] L. Ying, Z. Liu and D. Towsley, "Distributed operator placement and data caching in large-scale sensor networks", Proceedings of the IEEE INFOCOM, **(2008)**; Phoenix, AZ, USA.
[15] Z. Lu and Y. Wen, "Distributed Algorithm for Tree-Structured Data Aggregation Service Placement in Smart Grid", IEEE Systems Journal, vol. 8, no. 2, **(2014)**, pp. 553-561.
[16] X. Huang and S. Wang, "Aggregation Points Planning in Smart Grid Communication System", IEEE Communications Letters, vol. 19, no. 8, **(2015)**, pp. 1315-1318.
[17] L. Ting, S. Yanan, L. Yang and Y. Gui, "Abnormal traffic-indexed state estimation: A cyber–physical fusion approach for Smart Grid attack detection", Future Generation Computer Systems, vol. 49, no. 8, **(2015)**, pp. 94-103.
[18] H. Li, X. Lin and H. Yang, "EPPDR: an efficient privacy-preserving demand response scheme with adaptive key evolution in smart grid", IEEE Transactions on Parallel and Distributed Systems, vol. 25, no. 8, **(2014)**, pp. 2053-2064.
[19] C. I. Fan, S. Y. Huang and Y. L. Lai, "Privacy-enhanced data aggregation scheme against internal attackers in smart grid", IEEE Transactions on Industrial Informatics, vol. 10, no. 1, **(2014)**, pp. 666-675.

[20] C. Rottondi., G. Verticale and C. Kraus, "Secure distributed data aggregation in the automatic metering infrastructure of smart grids", Proceedings of the IEEE International Conference on Communications, **(2013)**; Budapest, Hungary.
[21] T. Flick and J. Morehouse, "Securing the smart grid: next generation power grid security", Elsevier, **(2010)**.
[22] W. Meng, R. Ma and H. H. Chen, "Smart grid neighborhood area networks: a survey", Network, vol. 28, no. 1, **(2014)**, pp. 24-32.
[23] Y. Yu, K. Li, W. Zhou and P. Li, "Trust mechanisms in wireless sensor networks: Attack analysis and countermeasures", Journal of Network and Computer Applications, vol. 35, no. 3, **(2012)**, pp. 867-880.
[24] G. Theodorakopoulos and J. S. Baras, "On trust models and trust evaluation metrics for ad hoc networks", IEEE Journal on Selected Areas in Communications, vol. 24, no. 2, **(2006)**, pp. 318–328.
[25] K. Fall and K. Varadhan, "The NS Manual," The VINT Project, vol. 1, **(2002)**.

Authors

Yiwen Le, He is a Ph.D. student with school of Electrical Engineering, Beijing Jiaotong University. His research interests include modeling, data aggregation and power prediction technologies in smart grid.

Jinghan He, She is the dean and a professor (full) now with school of Electrical Engineering, Beijing Jiaotong University. Her major research areas are on-line monitoring, protection and control of power system, power quality, new energy and intelligent power grid, and electrical rail transportation.

