Big Data Security Analytic for Smart Grid with Fog Nodes

Wenlin Han and Yang Xiao(✉)

Department of Computer Science, The University of Alabama,
342 H.M. Comer, Box 870290, Tuscaloosa, AL 35487-0290, USA
whan2@crimson.ua.edu, yangxiao@ieee.org

Abstract. Big data throws big security challenges in Smart Grid. Fog computing emerges as a novel technology to bring the ability of Cloud and big data analytic (BDA) to the edge of the networks. However, we lack practice on how to utilize Fog, Cloud, and BDA to address big data security challenges in Smart Grid. In this paper, we propose a Mapreduce-style algorithm, named Mapreduce-style Fast Non-Technical Loss Fraud Detection scheme (Mapreduce-style FNFD), to detect Non-Technical Loss fraud in Smart Grid. Compared to its original version, FNFD, Mapreduce-style FNFD can utilize the ability of Cloud, Fog, and BDA, and thus can solve big data security issues.

Keywords: Smart Grid security · Big data · Non-Technical Loss · Fog computing

1 Introduction

Smart Grid [1,2] employs smart devices, such as smart meters, to provide advanced and efficient power services. The generation, transmission, distribution, and redistribution processes in Smart Grid are under intensive monitoring and are interactive. The devices in Smart Grid are electric and programmable. The novel framework of Smart Grid introduces environmental friendly energy resources to our daily lives and makes better use of current energy resources. Various security schemes have been proposed to enhance Smart Grid security including intrusion detection-based schemes [3,4], such as SCADA [5] and Amilyzer [6], and other schemes targeting some special security issues in Smart Grid [7–10].

With the fast growth of the number of smart meters installed globally, the data in Smart Grid becomes tremendously big. Big data in Smart Grid throws various challenges to the utility including big data storage, big data processing, and big data security. Big data storage and processing are widely studied, but there are only a few research works on big data security, especially in Smart Grid [11–13].

Cloud and Big Data Analytic (BDA) are often employed to address the big data security challenge, which we call Big Data Security Analytic (BDSA). However, there are two three problems of applying BDSA to Smart Grid. The first

G. Wang et al. (Eds.): SpaCCS 2016, LNCS 10066, pp. 59–69, 2016.
DOI: 10.1007/978-3-319-49148-6_6

problem is availability. The limited resource of devices in Smart Grid often makes it difficult to access data in Cloud. The second problem is experience. There are only a few studies on BDSA in Smart Grid, and they only introduce BDSA at a high level. We lack experience and practice on how to analyze big data in Smart Grid for the security purpose. The last problem is feasibility. We have many traditional security algorithms but do not know whether they fit into BDSA and how.

Fog computing is an emerging technology that aims to bring Cloud and BDA closer to end-user devices. In this paper, we propose MapReduce-style FNFD, which is a BDSA algorithm for Smart Grid with the support of fog computing. MapReduce-style FNFD is built on a Non-Technical Loss (NTL) fraud [14] detector, FNFD [15]. FNFD is based on Recursive Least Square (RLS) [16]. We split the RLS problem into multiple sub-problems and parallel the problem-solving process, which is the basic idea behind MapReduce-style FNFD.

The main contributions of this paper include:

- We propose an algorithm for big data security analytic in Smart Grid;
- We study the feasibility of introducing fog computing into Smart Grid;
- Our study provides a concrete example on how to analyze big data in Smart Grid for the security purpose.

The rest of the paper is organized as follows: In Sect. 2, we introduce the background of this paper. In Sect. 3, we briefly introduce FNFD, a previously proposed NTL fraud detector. MapReduce-style FNFD is then presented in Sect. 4. We conclude the paper in Sect. 6.

2 Background

In this section, we will introduce the background of this paper, including big data, Cloud, Fog computing, big data security issues, and security analytic.

2.1 Big Data, Cloud and Fog Computing in Smart Grid

Smart Grid is the new generation of power grid, and it provides various advanced features including self-healing [17], real-time pricing and billing, distributed energy generation, real-time monitoring, renewable energy resources, smart home appliances, smart meters, etc. The communication flow is real time and two way in Smart Grid [18]. When smart home appliances connect with smart meters in a home, it is a Home Area Network (HAN). When homes in a community or neighborhood connect to each other, it is a Neighborhood Area Network (NAN) [19–22]. Factories have smart meters and other electric devices installed, and these form Industrial Area Networks (IAN). Vehicles are using electricity to replace gas and they can sell excess electricity back to the grid. The communication networks between these vehicles are Vehicle to Grid (V2G) networks [23,24].

However, with the growth of smart devices, applications, and networks in Smart Grid, the data are becoming tremendously big. Big data brings big challenges to the utility to deal with huge volume of data in Smart Grid. As a solution

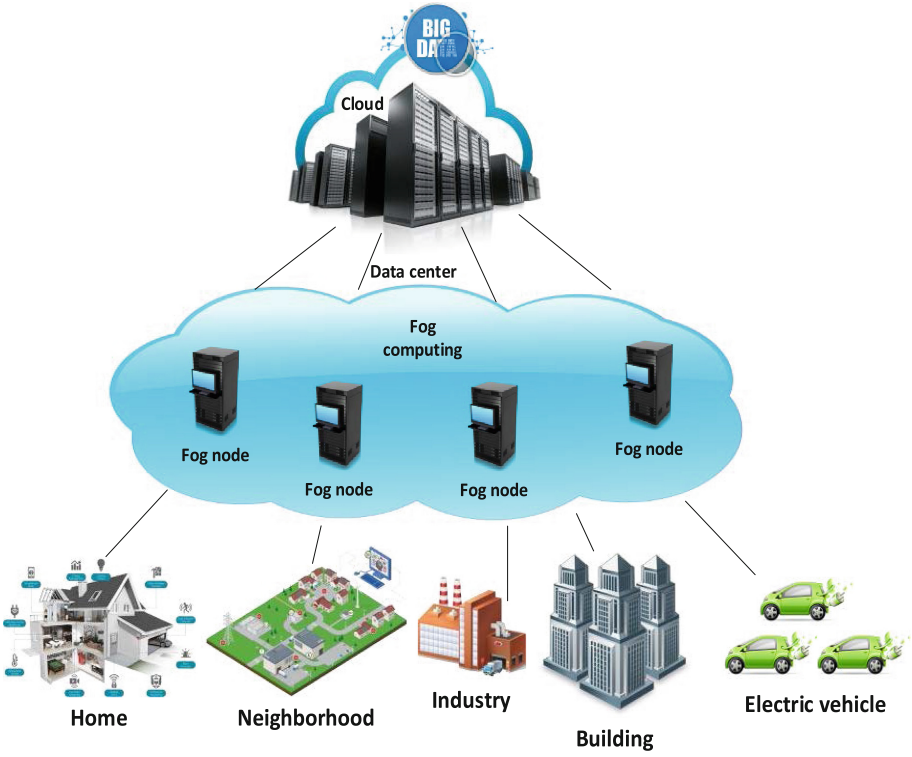

Fig. 1. The conceptual framework of Cloud and Fog computing in Smart Grid.

to big data problem, Cloud and big data analytic (BDA) often co-occur. Cloud utilizes the processing and storage capabilities of millions of servers to process or store data that a single server cannot handle. BDA is to employ Hadoop technologies, including MapReduce [25,26], Hadoop distributed file system (HDFS), Hive, and Pig, to process and store big data. The main problem of Cloud in Smart Grid is bandwidth. Cloud has to process and store all the data in the data center, and the data center is often far from the devices in Smart Grid. Some networks in Smart Grid are wireless and highly dynamic, such as V2G networks. The bandwidth in these networks is slow.

Fog computing is an emerging technology that aims to address the above problem in Cloud. Instead of processing and storing data in the data center, fog computing brings processing and storage capability closer to the end-user devices. By the collaboration of multiple near-user devices, which we call fog nodes, fog computing can handle big data that a single device cannot process and can provide faster responses than Cloud. As shown in Fig. 1, it is the conceptual framework of Cloud and Fog computing in Smart Grid.

2.2 Big Data Security Analytic in Smart Grid

Big data throws big challenges to security in Smart Grid, including key management, trust management, privacy preservation, fraud protection, identity management, etc. Big security data includes single big security dataset, large amount of security datasets and big heterogeneous security data [27]. In Smart Grid, the main type of big security data is large amount of security datasets. The size of a single security dataset may be small, but the volume of the data is big.

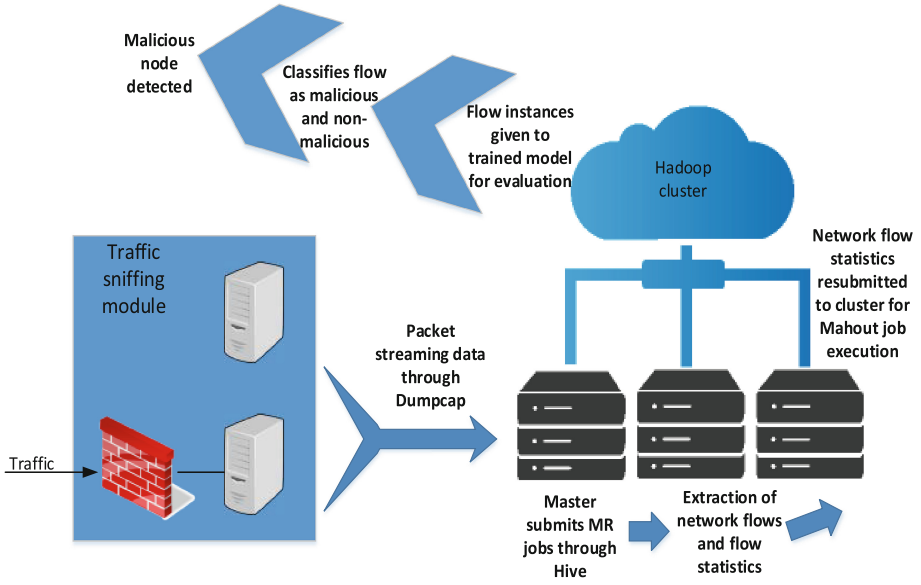

Fig. 2. A typical BDSA framework that uses a hadoop cluster to analyze network traffic and detect malicious nodes [27].

Big data security analytic (BDSA) is to use BDA tools or technologies to analyze security related data and make security related decisions. The traditional security solutions can handle these small datasets, however the speed is slow. Instead of analyzing the datasets one by one, BDSA parallelizes the process by using MapReduce-style algorithms. As shown in Fig. 2, it is a typical BDSA framework that uses a hadoop cluster to analyze network traffic and detect malicious nodes.

We propose MapReduce-style FNFD, which is a BDSA algorithm for Smart Grid. MapReduce-style FNFD is based on FNFD. FNFD is a algorithm that runs only at a single machine. MapReduce-style FNFD is a paralleled algorithm that runs collaboratively at fog nodes. Before introducing MapReduce-style FNFD, we briefly introduce FNFD and the problem it aims to solve in the following section.

3 Non-technical Loss Fraud and FNFD

In this section, we will briefly introduce FNFD and the problem it aims to solve.

3.1 Non-technical Loss Fraud

NTL fraud is a problem that harasses the utility for long. NTL fraud is where a fraudster tampers a smart meter so that the meter reports less amount of electricity than the home/business consumed. The fraudster could be the customers who tampered their own meters or the attackers who hacked the meters remotely. The utility is the victim who suffers from economic loss.

3.2 FNFD

FNFD is a detection scheme that aims to detect NTL fraud in Smart Grid. The criterion that used to differentiate a normal meter and a tampered meter is the relationship between a coefficient a_i and two thresholds α_{min} and α_{max}. The coefficient a_i represents Meter i. If $a_i > \alpha_{max}$, the meter is tampered. If $\alpha_{min} \leq a_{ij} \leq \alpha_{max}$, it is a normal meter. A is the vector of all coefficients, denoted as:

$$A = (a_1, a_2, \ldots, a_n). \tag{1}$$

A is obtained from the function:

$$E = AX, \tag{2}$$

where X is the vector of the amount of electricity reported by the meters, and E is the vector of the amount of consumed electricity.

The problem is solved based on RLS. The basic idea is to get an estimation of A which satisfies:

$$\min_{\forall A} \|E - AX\|^2 \tag{3}$$

4 MapReduce-style FNFD

The basic idea of MapReduce-style FNFD is to divide the global problem, shown in Eq. 3, into several sub-problems, assign sub-problem to processes in fog nodes, and finally solve the global problem via local solutions of sub-problems. At each iteration, the processes communicate, exchange local solutions, and update their own local solutions for the next iteration. Since the sub-problems can be solved simultaneously, it saves a lot of time dealing with big data.

We partition matrix X into p blocks by columns, say:

$$X = (X_1, X_2, \ldots, X_p). \tag{4}$$

We have

$$AX = \sum_{i=1}^{p} A_i X_i. \tag{5}$$

Let's define $b_i(A)$ as:

$$b_i(A) = E - \sum_{j \neq i} A_j X_j \tag{6}$$

The global problem shown in Eq. 3 can be divided into p sub-problems, that is:

$$\min_{\forall y} \|b_i(A) - yX_i\|^2, \tag{7}$$

where $1 \leq i \leq p$.

Each sub-problem is also a RLS problem, thus it is a paralleled RLS problem [28]. Let's define the global solution at iteration k as:

$$A^k = (A_1^k, A_2^k, \ldots, A_p^k). \tag{8}$$

The global solution at iteration $k+1$, A^{k+1}, can be obtained by solving a local problem:

$$\min_{\forall y^{k+1}} \left\| b_i(A^k) - y^{k+1} X_i \right\|, \tag{9}$$

where $1 \leq i \leq p$.

Let's define $\tilde{A}^{k+1}$ as the updated local solution at iteration $k+1$, and $\tilde{A}^{k+1}$ can be obtained from local solutions of previous iterations, denoted as

$$\tilde{A}^{k+1} = (A_1^k, A_2^k, \ldots, A_{i-1}^k, y_i^{k+1}, A_{i+1}^k, \ldots, A_p^k). \tag{10}$$

The global solution at iteration $k+1$ is given by:

$$A^{k+1} = \sum_{i=1}^{p} \sigma_i^{k+1} \tilde{A}_i^{k+1}, \tag{11}$$

where $\sum_{i=1}^{p} \sigma_i^{k+1} = 1$.

The local solution at iteration $k+1$ is given by:

$$\begin{aligned} A_i^{k+1} &= (\sigma_i^{k+1} \tilde{A}_i^{k+1})_i + (\sum_{\substack{j=1 \\ j \neq i}}^{p} \sigma_j^{k+1} \tilde{A}_j^{k+1})_i \\ &= \sigma_i^{k+1} y_i^{k+1} + \sum_{\substack{j=1 \\ j \neq i}}^{p} \sigma_j^{k+1} A_i^k \\ &= \sigma_i^{k+1} y_i^{k+1} + (1 - \sigma_i^{k+1}) A_i^k \\ &= A_i^k + \sigma_i^{k+1} (y_i^{k+1} - A_i^k) \\ &= A_i^k + \sigma_i^{k+1} \xi_i^{k+1}, \end{aligned} \tag{12}$$

where $\xi_i^{k+1} = y_i^{k+1} - A_i^k$.

The local value $b_i(A)$ at iteration k can be obtained by:

$$b_i(A^{k+1}) = b_i(A^k) - \sum_{\substack{j=1 \\ j \neq i}}^{p} \sigma_j^{k+1} X_j \xi_j^{k+1}. \tag{13}$$

Let's define $B_j^{k+1} = X_j \xi_j^{k+1}$, and we get:

$$b_i(A^{k+1}) = b_i(A^k) - \sum_{\substack{j=1 \\ j \neq i}}^{p} \sigma_j^{k+1} B_j^{k+1}. \tag{14}$$

After getting the value of the vector A, we compare the value to α_{max} and α_{min}. If $a_i > \alpha_{max}$, the meter is tampered. If $\alpha_{min} \leq a_{ij} \leq \alpha_{max}$, it is a normal meter.

Algorithm 1. MapReduce-style FNFD algorithm at a slave node

Initiation: For all slave nodes i, $1 \leq i \leq p$, test linear independent of X_i. $b_i(A^{k+1}) = b$, $y_i^0 = A_i^0$, $k = 0$.
repeat
  calculate $B_i^{k+1} = X_i \xi_i^{k+1}$;
  global communication: B_i^{k+1} to all slave nodes;
  calculate $b_i(A^{k+1}) = b_i(A^k) - \sum_{\substack{j=1 \\ j \neq i}}^{p} \sigma_j^{k+1} B_j^{k+1}$;
  solve local least square $\min_{\forall y^{k+1}} \left\| b_i(A^k) - y^{k+1} X_i \right\|$;
  $\xi_i^{k+1} = y_i^{k+1} - A_i^k$;
  $A_i^{k+1} = A_i^k + \sigma \xi_i^{k+1}$;
  test local convergence;
  global communication: convergence results;
  $k = k + 1$.
until converged
for each a_i in $\mathbf{A_i}$ **do**
  if $a_i > \alpha_{max}$ **then**
    identified as tampered
  if $\alpha_{min} \leq a_i \leq \alpha_{max}$ **then**
    identified as normal
  else
    report error

The algorithm of MapReduce-style FNFD can be divided into the master algorithm and the slave algorithm. The master algorithm runs at the master node. The main function is to split metrics A and X into p blocks by column. The slave algorithm runs at every slave node, shown in Algorithm 1. The matrix X has to be linear independent to have a solution. Testing liner independent can be carried out either at the master node or every slave node. Here, we let every slave node test liner independent individually.

Table 1. Registered electricity consumption (kWh) in the experiment: 3 m and 1 observer meter

Meter 1	Meter 2	Meter 3	Observer meter
3	4	1	10
2	3	1	8
1	2	3	12

Local convergence means that a given error, e, is satisfied when solving a local least square problem. When all the slave nodes get converged results, the global convergence is achieved and the detection process obtains a result.

5 A Case Study

In this section, we will use a simple case study to show the effectiveness of MapReduce-style FNFD.

As shown in Table 1 is the electricity consumption data of three smart meters and one observe meter. The data is collected every 15 min. The observer records the total unit supplied to these three meters. By simply adding these numbers, we can see that the total billed unit is less than the total unit supplied. Thus, some meter(s) are tampered.

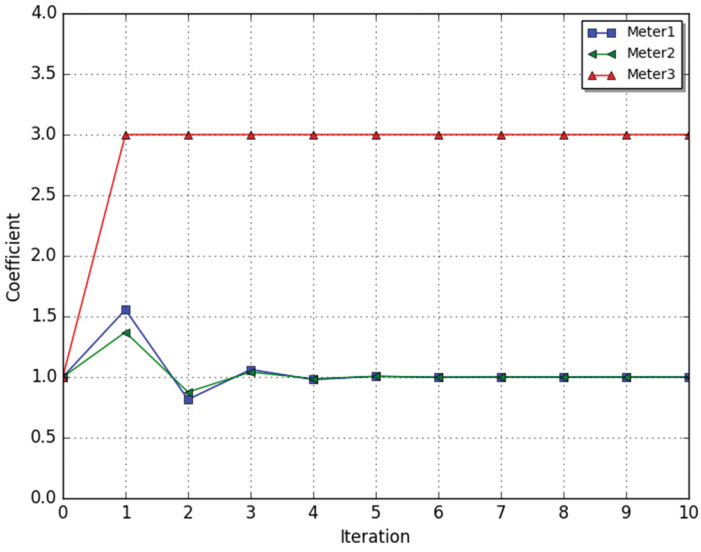

Fig. 3. The detection process of MapReduce-style FNFD. The coefficient of Meter 3 converges to 3 while the coefficients of other meters converge to 1. It shows that Meter 3 is tampered and Meter 1 and Meter 2 are normal.

In the previous work, we showed how to use FNFD, which runs at a single machine, to detect the tampered meter(s). Here, we split the global problem into three sub-problems and use four machines to detect the tampered meter(s), among which three machines are used as slave nodes. Thus, the value of p is 3. A_i^0 is set to a vector of 1 initially. The given error e is set to 10^{-5}.

The coefficient convergence process is shown in Fig. 3. The coefficient of Meter 3 converges to 3 while the coefficients of other meters converge to 1. It means that Meter 1 and Meter 2 are normal meters while Meter 3 is tampered. The simple case is only to illustrate how MapReduce-style FNFD works. It can work on large data sets and the performance is affected by coefficients p, e, σ and the size of the data sets, which will be introduced in the long and journal version of this paper.

6 Conclusion

In this paper, a big data analytic algorithm was proposed to address security issues in Smart Grid, which is named MapReduce-style FNFD. We introduced big data security challenges in Smart Grid and how Cloud, fog computing, and big data analytic can help to address these challenges. MapReduce-style FNFD is an algorithm built on an existing NTL fraud detector. Our study provides real practice of introducing big data security analytic into Smart Grid. As a future work, we will further test the performance and convergence of MapReduce-style FNFD in various aspects.

Acknowledgments. This work was supported in part by the National Natural Science Foundation of China under the Grant 61374200, and the National Science Foundation (NSF) under Grant CNS-1059265.

References

1. Liu, J., Xiao, Y., Li, S., Liang, W., Chen, C.L.P.: Cyber security, privacy issues in smart grids. IEEE Commun. Surv. Tutorials **14**(4), 981–997 (2012)
2. Kundur, D., Feng, X., Mashayekh, S., Liu, S., Zourntosand, T., Butler-Purry, K.L.: Towards modelling the impact of cyber attacks on a smart grid. Int. J. Secur. Netw. **6**, 2–13 (2011)
3. Han, W., Xiong, W., Xiao, Y., Ellabidy, M., Vasilakos, A.V., Xiong, N.: A class of non-statistical traffic anomaly detection in complex network systems. In: Proceedings of the 32nd International Conference on Distributed Computing Systems Workshops (ICDCSW 2012), pp. 640–646, June 2012
4. Kalogridis, G., Denic, S.Z., Lewis, T., Cepeda, R.: Privacy protection system and metrics for hiding electrical events. Int. J. Secur. Netw. **6**, 14–27 (2011)
5. Gao, J., Liu, J., Rajan, B., Nori, R., Fu, B., Xiao, Y., Liang, W., Chen, C.L.P.: SCADA communication and security issues. Secur. Commun. Netw. **7**(1), 175–194 (2014)
6. Berthier, R., Sanders, W.H.: Monitoring advanced metering infrastructures with amilyzer. In: Proceedings of C&ESAR: The Computer & Electronics Security Applications Rendez-vous, Cyber-security of SCADA & Industrial Control Systems, Rennes, France, 19-21 Nov. 2013, pp. 130–142 (2013)

7. Han, W., Xiao, Y., Combating, T.: Non-technical loss fraud targeting time-based pricing in smart grid. In: The 2nd International Conference on Cloud Computing and Security (ICCCS 2016), July 2016
8. Han, W., Xiao, Y.: NFD: a practical scheme to detect non-technical loss fraud in smart grid. In: Proceedings of the 2014 International Conference on Communications (ICC 2014), pp. 605–609, June 2014
9. Han, W., Xiao, Y.: CNFD: a novel scheme to detect colluded non-technical loss fraud in smart grid. In: Yang, Q., Yu, W., Challal, Y. (eds.) WASA 2016. LNCS, vol. 9798, pp. 47–55. Springer, Heidelberg (2016). doi:10.1007/978-3-319-42836-9_5
10. Zhang, J., Gunter, C.A.: Application-aware secure multicast for power grid communications. Int. J. Secur. Netw. **6**, 40–52 (2011)
11. IBM: Managing big data for smart grids and smart meters (2015). http://www-935.ibm.com/services/multimedia/Managing__big_data_for_smart_grids_and_smart_meters.pdf
12. Ray, P., Reed C., Gray, J., Agarwal, A., Seth, S.: Improving roi on big data through formal security and efficiency risk management for interoperating ot and it systems (2012). http://www.gridwiseac.org/pdfs/forum_papers12/ray_reed_gray_agarwal_seth_paper_gi12.pdf
13. Li, F., Luo, B., Liu, P.: Secure and privacy-preserving information aggregation for smart grids. Int. J. Secur. Netw. **6**, 28–39 (2011)
14. Han, W., Xiao, Y.: Non-technical loss fraud in advanced metering infrastructure in smart grid. In: The 2nd International Conference on Cloud Computing and Security (ICCCS 2016), July 2016
15. Han, W., Xiao, Y.: FNFD: a fast scheme to detect and verify non-technical loss fraud in smart grid. In: Proceedings of the International Workshop on Traffic Measurements for Cybersecurity (WTMC 2016), May–June 2016
16. Hayes, M.H.: Recursive least squares. In: Statistical Digital Signal Processing and Modeling, chap. 9.4, p. 154. Wiley (1996)
17. Mu, J., Song, W., Wang, W., Zhang, B.: Self-healing hierarchical architecture for ZigBee network in smart grid application. Int. J. Sens. Netw. **17**, 130–137 (2015)
18. Gao, J., Xiao, Y., Liu, J., Liang, W., Chen, C.L.P.: A survey of communication/networking in smart grids. Future Gener. Comput. Syst. **28**, 391–404 (2012). (Elsevier)
19. Xiao, Z., Xiao, Y., Du, D.: Exploring malicious meter inspection in neighborhood area smart grids. IEEE Trans. Smart Grid **4**(1), 214–226 (2013)
20. Xiao, Z., Xiao, Y., Du, D.: Non-repudiation in neighborhood area networks for smart grid. IEEE Commun. Mag. **51**(1), 18–26 (2013)
21. Xia, X., Liang, W., Xiao, Y., Zheng, M., Xiao, Z.: Difference-comparison-based approach for malicious meter inspection in neighborhood area smart grids. In: Proceedings of the 2015 International Conference on Communications (ICC 2015), pp. 802–807, June 2015
22. Xia, X., Liang, W., Xiao, Y., Zheng, M.: BCGI: a fast approach to detect malicious meters in neighborhood area smart grid. In: Proceedings of the 2015 International Conference on Communications (ICC 2015), pp. 7228–7233, June 2015
23. Han, W., Xiao, Y.: IP^{2}DM for V2G networks in smart grid. In: Proceedings of the 2015 International Conference on Communications (ICC 2015), pp. 782–787, June 2015
24. Han, W., Xiao, Y.: Privacy preserving for V2G networks in smart grid: a survey. Comput. Commun. **91–92**, 17–28 (2016)

25. Xiao, Z., Xiao, Y.: Accountable MapReduce in cloud computing. In: Proceedings of 2011 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS), pp. 1082–1087 (2011)
26. Xiao, Z., Xiao, Y.: Achieving accountable mapreduce in cloud computing. Future Gener. Comput. Syst. **30**, 1–13 (2014). (Elsevier)
27. Han, W., Xiao, Y.: Cybersecurity in internet of things - big data analytics. In: Big Data Analytics for Cybersecurity. Taylor & Francis Group (2016, in press)
28. Renaut, R.A.: A parallel multisplitting solution of the least squares problem. Numer. Linear Algeba Appl. **5**, 11–31 (1998)