Security and Privacy for Smart Meters: A Data-Driven Mapping Study

Ioannis Antoniadis
Electrical and Computer Engineering
Aristotle University of Thessaloniki
Thessaloniki, Greece
giannis.antoniadis@issel.ee.auth.gr

Kyriakos Chatzidimitriou
Electrical and Computer Engineering
Aristotle University of Thessaloniki
Thessaloniki, Greece
kyrcha@issel.ee.auth.gr

Andreas Symeonidis
Electrical and Computer Engineering
Aristotle University of Thessaloniki
Thessaloniki, Greece
asymeon@eng.auth.gr

***Abstract*—Smart metering systems have been gaining popularity as a vital part of the general smart grid paradigm. Naturally, as new technologies arise to cover this emerging field, so do security and privacy related issues regarding the energy consumer's personal data. These challenges impose the need for the development of new methods through a better understanding of the state-of-the-art. This paper aims at identifying the main categories of security and privacy techniques utilized in smart metering systems from a three-point perspective: i) a field research survey, ii) EU initiatives and findings towards the same direction and iii) a data-driven analysis of the state-of-the-art and the identification of its main topics (or themes) using topic modeling techniques. Detailed quantitative results of this analysis, such as semantic interpretation of the identified topics and a graph representation of the topic trends over time, are presented.**

***Index Terms*—privacy, security, smart grid, smart meters, topic modeling**

I. Introduction

Since the first official definition of the term smart grid in the Energy Independence and Security Act of 2007 [1], there have been many efforts, worldwide, towards the migration from the legacy power grid to an advanced energy transmission and distribution infrastructure. The main characteristics of the new smart grid infrastructure were set as follows: i) use of information and communication technologies, ii) optimization of grid operations and resources, iii) deployment of renewable energy resources and smart devices for metering energy consumption and iv) integration of smart appliances, accompanied by respective standards and regulations with the goal to improve the reliability, security, interoperability and efficiency of the electric grid [2].

Smart metering systems play a major role in this technological shift since they provide a two-way communication channel between consumers and distribution system operators, energy suppliers and other third parties [3]. The establishment of an Advanced Metering Infrastructure (AMI) increases the frequency and accuracy of the measured consumption data to near real time. This, in turn, improves the quality of services by assisting the optimization of electric energy distribution and management. The adoption of smart metering systems could also help energy consumers better understand the nature of their consumption and find motives and intuitive ways to reduce it. Such a development could subsequently result in lower-priced energy bills and an eco-friendlier behavior.

The proper function of smart metering systems relies heavily on communication technologies and the Internet to support the two-way communication channel. This fact poses new security and privacy threats due to the size of the infrastructure and its several structural layers. Smart metering systems display vulnerabilities that can be exploited in a number of cyber-attacks [4] and eventually hinder their fast and smooth adoption. Therefore, dedicated protection measures and techniques need to be developed, ensuring the security and privacy of the data being transferred through the grid.

In that context, of major significance are initiatives issued by worldwide organizations such as the National Institute for Standards and Technology (NIST) and the European Commission (EC). Their actions have resulted in the production of standards and recommendations for the deployment of secure and privacy-aware smart metering systems. The aforementioned standardized work is used as a baseline by the field specialists, who follow and further extend it in order to create dedicated and high quality solutions [5].

This paper aims at contributing to a comprehensive interpretation of the field by identifying the main categories of utilized techniques. The rest of the document is structured as follows. In Section II, a survey of the state-of-the-art in the field of smart metering security and privacy is provided, followed by relevant actions and findings issued by dedicated EU instruments. Section III presents the methodology that was followed in the context of the data-driven analysis, while, in Section IV, detailed quantitative results of this analysis are presented. Finally, Section V concludes the paper by highlighting its most significant points and addressing future research prospects.

II. Background

A. Related Work

The need for privacy-aware and secure solutions in the smart metering infrastructure arose shortly after the introduction of this relatively new type of energy metering systems. This need was amplified due to the rapid growth of big data and machine learning techniques that managed to extract patterns from unstructured data collections. The majority of them used

unsupervised learning to detect patterns in consumption data [6], [7], while others utilized analytics methods in order to collect large volumes of energy measurements and extract useful insights by analyzing them, based on criteria such as temporal constraints and aggregation thresholds [8], [9].

The first research publications towards secure and privacy-aware smart metering systems involved techniques such as data aggregation combined with homomorphic encryption[1] [10], differential privacy[2] and noise addition [11], [12], combinations of data aggregation and secret sharing schemes[3] [13], [14] and appliance load modification with the assistance of rechargeable batteries [15], [16]. A number of publications tackled the subject from an adversarial point of view, trying to expose the vulnerabilities of the existing smart metering infrastructure [17], [18]. Another stream of work used hard-coded values in the hardware layer in order to separate the privacy-sensitive information and anonymize it in subsequent stages [19]. Finally, survey papers attempted to present the existing state of the field and relevant techniques [20], [21].

As the field gained more traction, knowledge from various research areas was utilized to the forging of new solutions that meet the evolving security and privacy requirements of the smart metering systems. Advanced network topologies, cryptography techniques and anonymization schemes were used and often combined to build threat models [22], [23], as well as to anonymize energy consumption data before their transmission through the grid [24]–[26]. Mapping the aforementioned research publications to a single category of techniques (e.g. networking, cryptography, anonymization) is a rather difficult task since the majority of them use a combination of these.

Several research efforts address privacy preservation by the introduction of models that include Storage Unit Devices (SUD) and Alternative Energy Sources (AES) at the consumer's premises [27], [28]. The main benefit of their utilization is the capability to conceal energy consumption patterns by partially covering the energy load requirements from these sources, instead of the distribution network.

On top of the above-presented work, the following publications are equally worth mentioning. In [29], data sanitization is applied to critical measurements that are utilized to identify the consumer's habits, while [30] summarizes security and privacy concerns from the consumer's perspective. Finally, [31] proposes a gamification model that compensates consumers in the context of utility-privacy trade-off, according to their level of sharing their personal data with third party systems. Finally, a considerable amount of field surveys that cover security and privacy from various standpoints on the smart grid infrastructure were issued as well [32]–[35].

[1]an encryption type that allows cipher text computation. This process generates an encrypted result that, when decrypted, matches the operations result as if they had been applied on the plain text.

[2]a constraint - on methods that expose aggregate statistics - which ensures the privacy of the individual participating data.

[3]distribution of a secret to a group of participating members, each of which possesses a different share of the secret. An adequate number of shares must be combined for the secret's successful reconstruction.

B. EU Initiatives

The European Commission (EC) has set up a dedicated unit in 2009, namely the Smart Grids Task Force, to advise on issues with regards to smart grids development and deployment. The SGTF's Expert Group 2 (EG2) is responsible for the investigation of security, privacy and data protection issues in the smart grid environment. In this context, the group conducted a two-year assessment of the Best Available Techniques (BATs) for security and privacy in smart metering systems (2014-2016) with the assistance of external stakeholders. Their collective efforts resulted in the creation of a recommendations package [36] in line with the General Data Protection Regulation (GDPR) [37].

According to the Best Available Techniques Reference Document [36] issued by EG2, the identified techniques can be clustered in a number of categories reflecting the type/domain of application: access control, communication/transport, reading/tariffing, cryptography, monitoring, security architecture, time synchronization, privacy and hardware security. A detailed table of techniques per type/domain is provided on the paper's companion Github repository[4].

The sets of techniques were evaluated across four main dimensions: i) cyber-security, ii) privacy and data protection, iii) maturity and upgradeability and iv) impact towards architecture [36]. One of the main conclusions of the EG2 assessment was that the participating stakeholders have differing views regarding the utilized solutions. This might be due to differing perceptions of the underlying security threats, and, also due to the diverse deployment architectures featured in the individual EU Member States.

III. Methodology of the Data-Driven Analysis

The focus of this paper is the identification of underlying topics (or themes) that characterize the field of smart metering security and privacy. This section presents the data-driven methodology that was followed in that context, including a description of the utilized topic modeling algorithm, Latent Dirichlet Allocation (LDA).

A. Latent Dirichlet Allocation

LDA is a generative mixture model that can be applied in collections of discrete data in order to identify latent topic structures [38]. In the case of text modeling, LDA is applied over a set of documents and produces: (i) a distribution over K topics for each of the D documents, $\theta[D][K]$, and ii) a distribution over V words (V is the size of the vocabulary) for each of the K topics, $\phi[K][V]$.

The generative process of the model, given a Dirichlet distribution with parameter vector α, of length K and a Dirichlet distribution with parameter vector β, of length V, is as follows:

1) For each topic, k, draw a word distribution, i.e. a multinomial with parameter vector ϕ_k according to β: $\phi \sim Dirichlet(\beta)$

[4]https://github.com/AuthEceSoftEng/eeris-lda

2) For each document, d, draw a topic distribution, i.e. a multinomial with parameter vector θ according to α: $\theta \sim Dirichlet(\alpha)$
3) For each word, w, of the document d:
   - draw a topic z according to θ: $z \sim Multinomial(\theta)$
   - draw a word, w, according to ϕ_z: $w \sim Multinomial(\phi_z)$

The only observed variables of the model are the words of the documents, while the values of θ and ϕ need to be inferred. A commonly used technique for LDA inference, which is utilized in this work as well, is Gibbs sampling [38]. For the evaluation of the LDA model, the log-likelihood metric is employed to verify the model's convergence during the inference phase [39]. Likewise, the perplexity metric is utilized to evaluate the estimated model.

B. Implementation of the Methodology

To feed the LDA model with input data, a systematic research for relevant publications was conducted in Elsevier Scopus database[5]. A corpus of publications was retrieved and pre-processed, before its main processing by the LDA module. More precisely, the below procedural steps were followed:

(i) **Query formulation**: a single search query was formulated by applying a number of search terms to Scopus engine and reviewing the result hits in terms of the addressed topic and total number of hits. This process resulted in the selection of the query *"smart meter data privacy"*.

(ii) **Documents retrieval**: the selected query was applied to Scopus engine combined with the requirement that the result hits should be either conference papers or journal publications. The search returned 340 hits, 90 of which could not be downloaded due to paywall restrictions or broken links. This resulted in a number of 250 downloaded PDF documents, along with their meta-info.

(iii) **Documents overview**: the 250 documents were manually inspected in terms of content similarity and duplicates detection. 5 duplicate documents were found and removed from the corpus, resulting in a number of 245 documents.

(iv) **Documents indexing and pre-processing**: each document in the corpus was parsed and indexed to a NoSQL database storage (Elasticsearch) along with its meta-info. During this step, each document's text body was pre-processed by means of the below sequential steps:
   - Removal of any text before the abstract section and any text from the references section and after.
   - Stripping of all symbols and non-ASCII characters.
   - Conversion of all characters to lowercase.
   - Removal of extra spaces.
   - Join collocating terms "smart meter" to "smart_meter" and "smart grid" to "smart_grid".
   - Removal of strings with less than 4 characters.
   - Tokenization.
   - Removal of common English stop-words via a stop-words filter.
   - Reduction of inflectional forms of words via a stemming algorithm.

   During the pre-processing step, one document could not be parsed correctly, so it was removed from the corpus. As a result, a final number of 244 documents were used as input to the LDA model.

(v) **Model application and tuning**: the LDA model was applied to the pre-processed document corpus in the context of several experiments, the results of which are presented in the next section.

The implementation of steps (i), (ii) and (iii) above was done manually, while the respective implementation of steps (iv) and (v) was developed in Java. A link to the implementation's Github repository is provided in footnote 4.

IV. Results

To assess the convergence of the LDA inference, 5 individual runs were executed with parameters: $\alpha = 0.5$, $\beta = 0.1$ and $K = \{5, 10, 20, 30, 50\}$. As shown in Fig. 1, where the log-likelihood curves of the multiple runs are superimposed, the values stabilize after a couple hundred iterations and the algorithm converges.

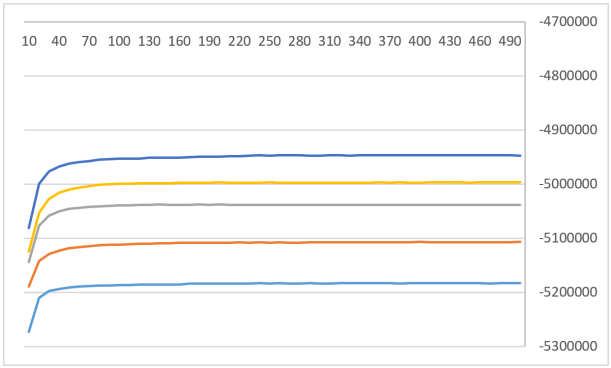

Fig. 1. Log-Likelihood vs. number of iterations

The final selection of K was done via human judgment, a commonly used method in the literature [40], [41]. The process included reviewing multiple models with varying K and choosing the one that produced the most sensible output in terms of topic coherence and presence in the dataset, according to the inferred θ distribution values. Checks were operated on samples of documents to assess the level of topic presence and if -and to what extend- they actually deal with the topics indicated by θ. As a result, the value $K = 7$ was chosen.

In an effort to optimize the model's hyper-parameters α and β, multiple runs were executed with $K = 7$ and varying values of α and β, and the perplexity values were calculated. Model's perplexity is minimized for $\{\alpha, \beta\} = \{0.1, 0.1\}$, therefore these are the selected values for the final model. The 10 top-words (stemmed terms) for each of the 7 identified topics of the selected model are presented in Table I, while the topic proportions in the input dataset are shown in Fig. 2.

[5]https://www.scopus.com/

TABLE I
10 MOST PROBABLE WORDS FOR EACH TOPIC

Topic 0	privacy	data	smart_meter	aggreg	valu
	consumpt	measur	provid	custom	nois
Topic 1	data	scheme	aggreg	encrypt	comput
	secur	user	propos	privaci	time
Topic 2	data	energi	time	load	applianc
	power	inform	consumpt	model	batteri
Topic 3	data	energi	smart_meter	consum	privaci
	system	electr	protect	inform	home
Topic 4	comput	meter	user	protocol	input
	proof	send	bill	game	read
Topic 5	secur	data	system	smart_grid	meter
	attack	communic	network	servic	control
Topic 6	node	network	data	aggreg	gateway
	protocol	messag	meter	share	communic

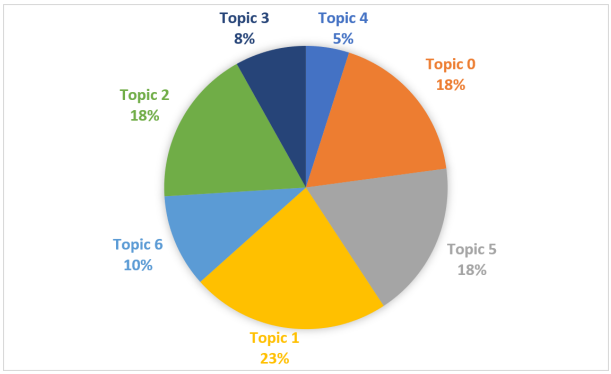

Fig. 2. Topic proportions in the dataset

The following observations regarding each topic's semantics can be made:

- **Topic 0**: indicative of privacy methods such as aggregation of consumption data values and noise addition.
- **Topic 1**: related to a combination of security and privacy techniques, such as aggregation schemes and encryption.
- **Topic 2**: refers to rechargeable battery models for appliance load modification (privacy preserving methods).
- **Topic 3**: references general notions of smart metering data privacy and protection at the consumer's side.
- **Topic 4**: related to security protocols and proofs in the context of user billing according to meter readings.
- **Topic 5**: conceptually close to smart grid security at the network/communications layer.
- **Topic 6**: refers to privacy preserving methods at the network/communications layer such as data aggregation at different nodes and data sharing/messaging protocols.

Fig. 3 presents the plot of topic trends over time. To achieve that, each document, d, was classified to a single topic according to the highest probability of its corresponding $\theta[d][k]$ values. The horizontal axis represents the publication year while the vertical axis shows the total number of documents. Since this research was carried out in the first quarter of 2019, the retrieved publications of this year are only 8. Therefore, it was decided that they are excluded from the plot to avoid a negative impact on the validity of this analysis.

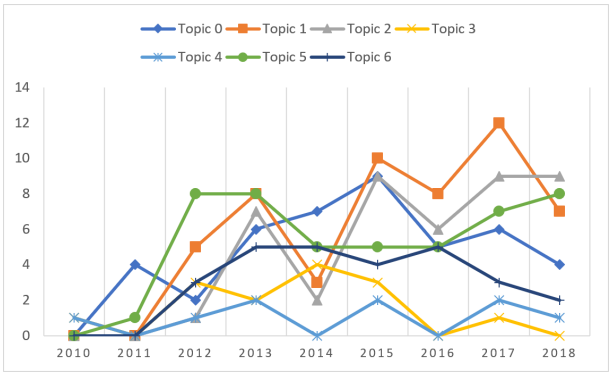

Fig. 3. Topic trends over time

An observation of the graphs shows that topics 2 and 5 maintain a dominant place in the dataset during the last few years, while topic 1, which used to be the highest ranked topic from 2015 to 2017, fell to third place in 2018. On the other hand, topics 3 and 6 display a constant decrease during the same period. Topic 0 was high-ranked in the years 2013-2015, however it shows a drop from 2016 and on. Finally, topic 4, which seems to be the weakest of all, follows a non-evaluable regression pattern throughout the years.

V. CONCLUSION

The design of secure and privacy-aware techniques in the field of smart metering requires a solid understanding of the existing state-of-the-art methods. This paper provides: (i) a relevant state-of-the-art survey, (ii) a presentation of EU initiatives and their results, operated by a dedicated SGTF group (EG2) and (iii) a data-driven analysis against a corpus of related publications in order to identify underlying topics (or themes) that characterize the field.

All three approaches managed to identify a number of techniques categories, such as encryption methods and networking protocols for data security and consumer billing, aggregation schemes and noise addition methods for data privacy preservation, rechargeable battery models and energy storage devices for appliance load modification. More specifically, the data-driven analysis identified trending topics such as appliance load modification techniques, general smart grid security methods at the network layer, as well as combinations of security and privacy solutions.

Finally, a direction for future work would be towards the improvement of the data-driven analysis. More precisely, the publications extraction procedure could be extended in order to increase the dataset's size, which is expected to increase the analysis' field coverage and reliability. The increase of the input dataset could be complemented by an increase in the number of topics, K, to be identified by the LDA model. By increasing K, the model is expected to generate more coherent topics that could result to a more comprehensive interpretation. A final clustering step could be added in order to group the K topics into sets of similar semantic content.

ACKNOWLEDGMENT

This research has been co-financed by the European Regional Development Fund of the European Union and Greek national funds through the Operational Program Competitiveness, Entrepreneurship and Innovation, under the call RESEARCH-CREATE-INNOVATE (project code: T1EDK-04045).

REFERENCES

[1] NIST, "Nist special publication 1108r2 nist framework and roadmap for smart grid interoperability standards," National Institute of Standards and Technology, Tech. Rep., 2012.

[2] O. H. T. C. of the United States of America, "Energy independence and security act of 2007," 2007.

[3] S. Finster and I. Baumgart, "Privacy-aware smart metering: A survey," *IEEE Communications Surveys Tutorials*, vol. 17, no. 2, pp. 1088–1101, Secondquarter 2015.

[4] S. Shapsough, F. Qatan, R. Aburukba, F. Aloul, and A. R. Al Ali, "Smart grid cyber security: Challenges and solutions," in *2015 International Conference on Smart Grid and Clean Energy Technologies (ICSGCE)*, Oct 2015, pp. 170–175.

[5] R. Leszczyna, "Cybersecurity and privacy in standards for smart grids - a comprehensive survey," *Computer Standards & Interfaces*, vol. 56, pp. 62–73, 2018.

[6] V. Ford and A. Siraj, "Clustering of smart meter data for disaggregation," in *2013 IEEE Global Conference on Signal and Information Processing*, Dec 2013, pp. 507–510.

[7] K. Okada, K. Matsui, J. Haase, and H. Nishi, "Privacy-preserving data collection for demand response using self-organizing map," in *2015 IEEE 13th International Conference on Industrial Informatics (INDIN)*, July 2015, pp. 652–657.

[8] O. V. Livingston, T. C. Pulsipher, D. Anderson, A. Vlachokostas, and N. Wang, "An analysis of utility meter data aggregation and tenant privacy to support energy use disclosure in commercial buildings," *Energy*, vol. 159, 06 2018.

[9] R. Razavi, A. Gharipour, M. Fleury, and I.-J. Akpan, "Occupancy detection of residential buildings using smart meter data: A large-scale study," *Energy and Buildings*, vol. 183, 11 2018.

[10] F. Li, B. Luo, and P. Liu, "Secure information aggregation for smart grids using homomorphic encryption," in *2010 First IEEE International Conference on Smart Grid Communications*, Oct 2010, pp. 327–332.

[11] G. Ács and C. Castelluccia, "I have a dream! (differentially private smart metering)," in *Information Hiding*, T. Filler, T. Pevný, S. Craver, and A. Ker, Eds. Springer Berlin Heidelberg, 2011, pp. 118–132.

[12] P. Barbosa, A. Brito, H. Almeida, and S. Clauß, "Lightweight privacy for smart metering data by adding noise," in *Proceedings of the 29th Annual ACM Symposium on Applied Computing*, ser. SAC '14. ACM, 2014, pp. 531–538.

[13] C. Rottondi, G. Mauri, and G. Verticale, "A data pseudonymization protocol for smart grids," in *2012 IEEE Online Conference on Green Communications (GreenCom)*, Sep. 2012, pp. 68–73.

[14] A. Barletta, C. Callegari, S. Giordano, M. Pagano, and G. Procissi, "Privacy preserving smart grid communications by verifiable secret key sharing," in *2015 International Conference on Computing and Network Communications (CoCoNet)*, Dec 2015, pp. 199–204.

[15] D. Varodayan and A. Khisti, "Smart meter privacy using a rechargeable battery: Minimizing the rate of information leakage," in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2011, pp. 1932–1935.

[16] S. McLaughlin, P. McDaniel, and W. Aiello, "Protecting consumer privacy from electric load monitoring," in *CCS'11 - Proceedings of the 18th ACM Conference on Computer and Communications Security*, ser. Proceedings of the ACM Conference on Computer and Communications Security, 11 2011, pp. 87–98.

[17] M. Jawurek, M. Johns, and K. Rieck, "Smart metering de-pseudonymization," in *Proceedings of the 27th Annual Computer Security Applications Conference*, ser. ACSAC '11. ACM, 2011, pp. 227–236.

[18] E. Buchmann, K. Böhm, T. Burghardt, and S. Kessler, "Re-identification of smart meter data," *Personal and Ubiquitous Computing*, vol. 17, 04 2013.

[19] T. W. Chim, S. M. Yiu, L. C. K. Hui, and V. O. K. Li, "Pass: Privacy-preserving authentication scheme for smart grid network," in *2011 IEEE International Conference on Smart Grid Communications (SmartGridComm)*, Oct 2011, pp. 196–201.

[20] F. Skopik, "Security is not enough! on privacy challenges in smart grids," *International Journal of Smart Grid and Clean Energy*, pp. 7–14, 01 2012.

[21] F. Siddiqui, S. Zeadally, C. Alcaraz, and S. Galvao, "Smart grid privacy: Issues and solutions," in *2012 21st International Conference on Computer Communications and Networks (ICCCN)*, July 2012, pp. 1–5.

[22] V. Tudor, M. Almgren, and M. Papatriantafilou, "Analysis of the impact of data granularity on privacy for the smart grid," in *Proceedings of the 12th ACM Workshop on Workshop on Privacy in the Electronic Society*, ser. WPES '13. ACM, 2013, pp. 61–70.

[23] H. Qu, P. Shang, X. J. Lin, and L. Sun, "Cryptanalysis of a privacy-preserving smart metering scheme using linkable anonymous credential," *IACR Cryptology ePrint Archive*, vol. 2015, p. 1066, 2015.

[24] S. Finster and I. Baumgart, "Elderberry: A peer-to-peer, privacy-aware smart metering protocol," in *2013 Proceedings IEEE INFOCOM*, April 2013, pp. 3411–3416.

[25] C. Rottondi, G. Verticale, and A. Capone, "Privacy-preserving smart metering with multiple data consumers," *Comput. Netw.*, vol. 57, no. 7, pp. 1699–1713, May 2013.

[26] C. Rottondi, G. Mauri, and G. Verticale, "A protocol for metering data pseudonymization in smart grids," *Trans. Emerging Telecommunications Technologies*, vol. 26, pp. 876–892, 2015.

[27] Y. Sun, L. Lampe, and V. W. S. Wong, "Smart meter privacy: Exploiting the potential of household energy storage units," *IEEE Internet of Things Journal*, vol. 5, no. 1, pp. 69–78, Feb 2018.

[28] G. Giaconi, D. Gündüz, and H. V. Poor, "Smart meter privacy with renewable energy and an energy storage device," *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 1, pp. 129–142, Jan 2018.

[29] L. Yang, H. Xue, and F. Li, "Privacy-preserving data sharing in smart grid systems," in *2014 IEEE International Conference on Smart Grid Communications (SmartGridComm)*, Nov 2014, pp. 878–883.

[30] P. Diamantoulakis, V. M. Kapinas, and G. Karagiannidis, "Big data analytics for dynamic energy management in smart grids," *Big Data Research*, vol. 5, 04 2015.

[31] A. Yassine, A. A. Nazari Shirehjini, and S. Shirmohammadi, "Smart meters big data: Game theoretic model for fair data sharing in deregulated smart grids," *IEEE Access*, vol. 3, pp. 2743–2754, 2015.

[32] S. Desai, R. Alhadad, N. Chilamkurti, and A. Mahmood, "A survey of privacy preserving schemes in ioe enabled smart grid advanced metering infrastructure," *Cluster Computing*, 07 2018.

[33] H. Souri, A. Dhraief, S. Tlili, K. Drira, and A. Belghith, "Smart metering privacy-preserving techniques in a nutshell," *Procedia Computer Science*, vol. 32, pp. 1087 – 1094, 2014.

[34] N. Komninos, E. Philippou, and A. Pitsillides, "Survey in smart grid and smart home security: Issues, challenges and countermeasures," *Communications Surveys & Tutorials, IEEE*, vol. 16, pp. 1933–1954, 04 2014.

[35] W. Wang and Z. Lu, "Cyber security in the smart grid: Survey and challenges," *Computer Networks*, vol. 57, p. 1344–1371, 04 2013.

[36] S.-G. T. F. S. Forum, "Best available techniques reference document for the cyber-security and privacy of the 10 minimum functional requirements of the smart metering systems," European Commission, Tech. Rep., 11 2016.

[37] E. Parliament and E. C. of the European Union, "General data protection regulation (gdpr)," European Union, Tech. Rep., 4 2016.

[38] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Mar. 2003.

[39] T. L. Griffiths and M. Steyvers, "Finding scientific topics." *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101 Suppl 1, pp. 5228–35, 2004.

[40] X. Wei and W. B. Croft, "Lda-based document models for ad-hoc retrieval," in *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '06. ACM, 2006, pp. 178–185.

[41] D. Hall, D. Jurafsky, and C. D. Manning, "Studying the history of ideas using topic models," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, ser. EMNLP '08, 2008, pp. 363–371.