Model-based Assessment for Balancing Privacy Requirements and Operational Capabilities in the Smart Grid

Fabian Knirsch*, Dominik Engel*, Marc Frincu† and Viktor Prasanna†
*Josef Ressel Center for User-Centric Smart Grid Privacy, Security and Control
Salzburg University of Applied Sciences
Urstein Sued 1, A–5412 Puch/Salzburg, Austria
Email: {fabian.knirsch, dominik.engel}@en-trust.at
†Ming-Hsieh Department of Electrical Engineering
University of Southern California
Los Angeles, USA
Email: {frincu, prasanna}@usc.edu

***Abstract*—The smart grid changes the way energy is produced and distributed. In addition both, energy and information is exchanged bidirectionally among participating parties. Therefore heterogeneous systems have to cooperate effectively in order to achieve a common high-level use case, such as smart metering for billing or demand response for load curtailment. Furthermore, a substantial amount of personal data is often needed for achieving that goal. Capturing and processing personal data in the smart grid increases customer concerns about privacy and in addition, certain statutory and operational requirements regarding privacy aware data processing and storage have to be met. An increase of privacy constraints, however, often limits the operational capabilities of the system. In this paper, we present an approach that automates the process of finding an optimal balance between privacy requirements and operational requirements in a smart grid use case and application scenario. This is achieved by formally describing use cases in an abstract model and by finding an algorithm that determines the optimum balance by forward mapping privacy and operational impacts. For this optimal balancing algorithm both, a numeric approximation and – if feasible – an analytic assessment are presented and investigated. The system is evaluated by applying the tool to a real-world use case from the University of Southern California (USC) microgrid.**

I. Introduction

In a smart grid a number of systems have to cooperate effectively. For instance, in a demand response (DR) use case, data is captured by a smart meter, stored in a database and finally used by a prediction unit to forecast customer energy usage. Data is captured, exchanged and processed in order to achieve this high-level use case. Other examples of such use cases include smart metering for billing or automated electric vehicle charging. As these use cases rely to a great extent on personal data, security and privacy are current issues and subject to ongoing research [1], [2], [3]. Privacy aware data retrieval and processing is therefore crucial in order to meet statutory and customer requirements. However, when adding to many privacy constraints, the system's ability to perform the intended task may degrade. In this paper, use cases are investigated that need an optimum trade-off between privacy and operational capabilities. There are use cases where both, privacy and operational capabilities can be achieved fully at the same time, this is, however, not subject of this paper.

As a motivating example, imagine a simple demand response use case where future energy consumption of a particular customer at a certain point in the day (e.g., around noon) is predicted based on past behavior. This requires to have smart meter data from that customer in a sufficient resolution (e.g., one meter value each fifteen minutes). On the other hand, when providing data in such a granularity the customer might be subject to privacy threats, such as predicting when the customer is present at home or the intended or inadvertent release of fine grained meter data to the public. One of the challenges in system engineering in the smart grid is thus to find a good trade-off between protecting an individual's privacy and being able to provide useful services. In Section II work is presented that performs privacy and security assessments based on an operational description of the system. There is, however, currently no approach that focuses on the evaluation of entire systems in the smart grid in order to find the optimum balance between privacy requirements and operational capabilities. This paper therefore contributes (i) a model that formally describes use cases in the smart grid; (ii) an algorithm to find the optimum balance between privacy and operational capabilities based on that model; and (iii) an approach to assess the impact of privacy constraints on the system. The algorithm presented in this paper involves the analytic solving of an equation. If this is not feasible, a numeric approximation can be applied. For evaluation a specific real-world DR use case drawing on insights from the USC microgrid is investigated closely.

The remainder of this paper is structured as follows: Section II provides an overview of related work in the domain of data flow analysis for security and privacy assessments. Further, state of the art assessment tools are discussed and it is shown how this work extends these tools with a holistic approach for optimization. Section III presents the abstract model for describing data flows and system dependencies by using graphs, transition functions and merging operators. Section IV discusses the two approaches for the optimal balancing algorithm, hence the analytic assessment and the numeric approximation. In Section V both approaches are evaluated by applying the tool to a real-world use case. Section VI

summarizes this work and provides an outlook to future work.

II. Related Work

This section presents related work in the domain of data flow analysis and state of the art assessment tools. A workflow-oriented security assessment tool using graphs is presented in [4]. The framework proposed by the authors is based on the evaluation of argument graphs. The system's input are security goal, workflow description, system description, attacker model and evidence. The assessment itself applies a discriminative set of graphs, containing the workflow goal, the actors involved and the messages exchanged. The result of the assessment process is quantitatively presented as an availability score and a confidentiality score. Both are plugged into the system by the evidence, which is based on (statistical) data about the devices. This tool is comprehensive for security analysis, however does not deal with the impact of security constraints on the operational capabilities. In the domain of the smart grid, McKenna et al. [5] discuss the issue of finding the optimum trade-off for smart metering frequencies between customer privacy and application feasibility. The authors illustrate some of the privacy impacts that are becoming evident with certain frequency intervals and investigate typical use cases, such as DR, and the need of data for the successful operation of these systems. The issue of balancing privacy requirements and operational capabilities is also addressed in other fields apart from the smart grid: Oliveira and Zaiane [6] present algorithms for balancing privacy constraints in data mining applications. Massaguer et al. [7] discuss a middleware for pervasive spaces. Their focus is on finding the trade-off between privacy and utility of such a middleware. While these approaches deal with balancing for data retrieval and processing, they do not propose a mathematical model to formally address the issue of balancing privacy and operational requirements.

III. Data Flow Model

An approach towards the modeling of use cases in the smart grid based on the European Smart Grid Reference Architecture [8] are *Data Flow Graphs* (DFG). Neureiter et al. [3], Dänekas et al. [9] and Knirsch et al. [10] thoroughly discuss the application of such directed graphs to privacy assessments in the smart grid. DFGs provide a detailed view of a system on multiple layers, ranging from high-level business goals to low level interactions of components. DFGs capture actors and information objects and support a wide range of attributes. These graphs provide a holistic view of a use case and are a powerful tool for interdisciplinary communication and detailed assessments. Based on the concept of representing data flows in the smart grid as directed graphs, we propose an abstraction of DFGs to a simplified *Data Flow Model* that only consists of nodes and directed edges and a minimum set of attributes, hence transition functions and a privacy requirements/operational requirement for each node. Reduced complexity makes numeric and analytic calculations feasible to be performed on this model.

Each use case is characterized by a set of actors, i.e., units (smart meter, DR prediction unit, ...), and by a set of information flows from one actor to another, i.e., data items. The model presented here is not limited to physical units, but also allows to be applied to more high-level concepts

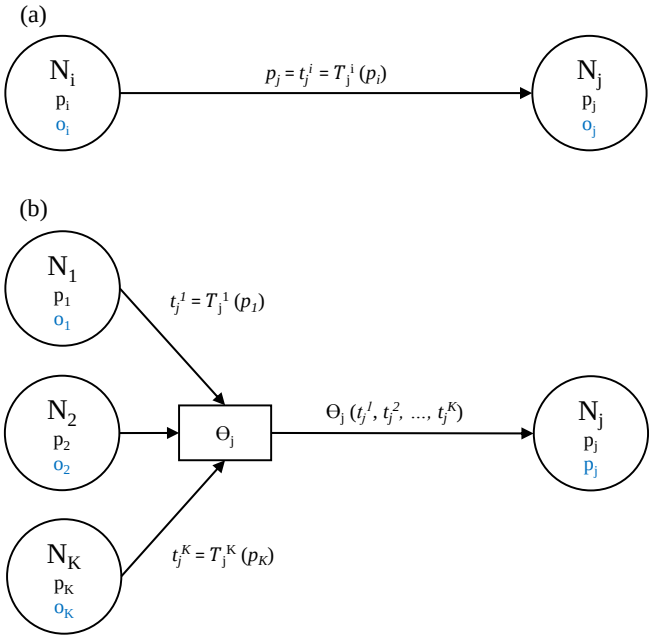

Fig. 1. Abstract data flow graph describing the model with nodes and edges.

such as *goals*, e.g., effective DR prediction. In an abstract notation this can be represented as a directed graph with a set of nodes N representing units or goals and a set of transitions T representing flows from a source node to a target node. A node N_k is described by a value p_k that defines the privacy requirement for that node and by a value o_k that defines the operational requirements for that node, so that $\alpha p_k + (1-\alpha)o_k = 1$ and therefore $\alpha \in [0,1]$. Thus requirements are represented as a numeric value in the range 0 to 1. The higher the value the more the requirement weighs. The above condition is introduced for normalization purposes.

An edge is described by a transition function T and a merging operator Θ, described in detail in Section III-A and Section III-B, respectively. The model for describing systems and information flows is in its simplest form as shown in Figure 1 (a). An edge connects a source node N_i with a target node N_j by $(p_j, o_j) = t_j^i = T_j^i\,(p_i, o_i)$; hence a function that maps the privacy and operational requirements of the source to the target. The other, general, case where a target node has more than one incoming edge is shown in Figure1 (b). In addition to the transition function a merge operation needs to be defined and the general form of an edge is given in Equation 1. The merging operator Θ maps a set of one or more input values p_k (the transition vector), each in the domain [0,1], from parent nodes N_k with $k = 1 \ldots K$ to one single output value in the same range. Note that this notation can be simplified in practice as the sum of privacy requirements and operational requirements in each node is defined to be 1 and thus only one of the parameters (either p or o) needs to be passed. Therefore, in the following only p is taken into account. For the sake of simplicity, recursive edges are not defined in the data flow model. Recursive edges would represent a system that sends data to itself and for the transition only the identity function would be feasible, since such a system has no practical privacy

$$p_i = \Theta_i \left(t_i^1 = T_i^1(p_1), \ldots, t_i^k = T_i^k(p_k) \right) \quad (1)$$

or operational impact on itself.

A. Transition Function

The transition function T is crucial as it immediately defines to what extent the destination system is able to perform its operations. The transition function is a case specific function that needs to satisfy the condition $T : [0,1] \rightarrow [0,1]$, so that the sum of the privacy requirements and operational requirements is one; and hence must not have a singularity in that interval, so that the model is not running in an undefined state. The transition function can be determined by practical observations or models, depending on the particular use case.

As an example for determining the transition function a (sub-)graph with two nodes is given, *smart meter* (N_a) and *DR prediction unit* (N_b). The transition function should represent the fact that the accuracy of DR prediction degrades if (for the particular use case) only data in low resolution is available, e.g., if DR prediction is used to forecast customer energy consumption on a hourly basis, one meter value per day is not sufficient. If we are further assuming that accuracy is following exponential behavior, the following transition function could be used: $p_b = T_b^a(p_a) = \frac{e^{p_a}-1}{e-1}$. The more the privacy is tuned up (thus lower frequency for metering), the less capable (thus less accurate) is the prediction unit.

B. Merging Operator

The merging operator Θ maps the transition vector which described incoming transitions to one single output value in the range 0 to 1. This operator can be determined by practical observations or models, depending on the particular use case or a generic approach can be found that equally incorporates each input value, e.g., by calculating the arithmetic mean.

C. Interpreting Results

Once proper values for p and o for the node of interest are found, these results must be interpreted accordingly to be applied to the system's characteristics in reality. The objective of interpreting results is therefore to map these normalized values to a property that impacts the privacy awareness or the operational capability of the system. This mapping is heavily dependent on individual characteristics and generic approaches provide only limited applicability. In our motivating example we discussed the impact of metering frequency on privacy and operation for subsequent systems. Hence, here we need to find a mapping from p_i with $N_i =$ "Smart Meter" to the meter frequency f_s. In the evaluation we discuss this issue thoroughly and we present such a mapping for the DR use case and in particular for the metering frequency in that scenario.

IV. Optimal Balancing

Once the model is constructed and all nodes and edges including the transition functions and the merging operators are defined, it is possible to calculate the optimal balancing between privacy and operational requirements. The optimal balancing is given by the solution of an equation. If solving this equation is not feasible, a numeric approach for approximating the result can be applied. This section discussed both approaches in detail. The objective of the optimal balancing is to perform the following: (i) automatically find the best trade-off between privacy requirements and operational capabilities for a system that is under development; or (ii) assess to what extent an existing system meets given privacy or operational requirements.

$$\bar{p} = \sum_{i=1}^{N} p_i \quad (2)$$

$$p_1 = \bar{p}N - \sum_{i=2}^{N} p_i \quad (3)$$

A. Analytic Assessment

The optimal balancing algorithm is performed on the entire system. The analytic assessment thus involves the solving of Equation 3, given an arbitrary $\bar{p}$ in the interval $[0,1]$, e.g., $\frac{1}{2}$ for the optimal balance. Again, the equality condition can be replaced by a greater equal or less equal condition. The equation yields a solution for each p_i for each node. In practice it is sufficient to specify the solution for p_1 or, in case each function T and each operator Θ has a well defined inverse function, to find the solution for an arbitrary p_i and then apply the given functions or the inverse functions in order to calculate the values for the node of interest. By doing so the model can be used to assess the impact of privacy/operational requirements in a particular node for other nodes. For complex systems consisting of many nodes, transitions and merging operations, solving the equation might not be possible or feasible. In the following section we therefore present a numeric algorithm.

B. Numeric Approximation

For approximating the result, a numeric approach can be applied. The algorithm for this approach is given as follows: (1) vary the values for p_1 and o_1, respectively, in the very first node and in the allowed interval, hence from 0 to 1 in a given step size Δ (e.g., 0.01); (2) compute T and Θ for each subsequent transition to get according values for each node; (3) for each variation, summarize and normalize the values for p and o for each node by $\bar{p} = \frac{1}{N}\sum_{i=1}^{N} p_i$ where N is the total number of nodes; and (4) find the variation where $\bar{p} = \frac{1}{2}$. It can be shown that the variation that satisfies the above condition yields the optimal balance between privacy requirements and operational requirements for the system as a whole.

If not a balanced system is intended, but a system that is either privacy aware to a certain extend or able to perform operations to a certain extent, the equality condition in $\bar{p} = \frac{1}{2}$ can be replaced by a more general condition involving a threshold s, such as $\bar{p} \geq s$ or $\bar{p} \leq s$. The remaining variations that satisfy this condition may then be subject to closer investigation.

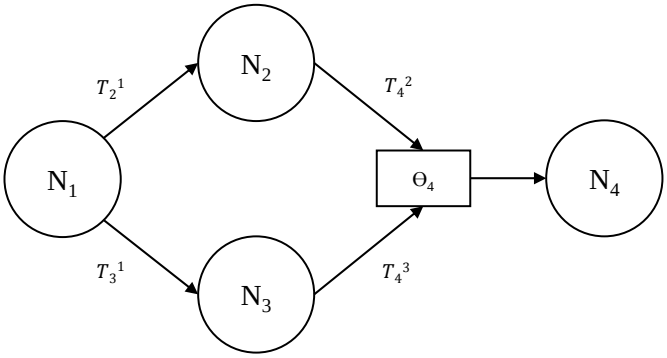

Fig. 2. Data flow model of the demand response use case.

V. Evaluation

For evaluating the system we apply privacy constraints and operational capabilities to a real-world use case that draws on insights from the USC microgrid[1]. First, a data flow graph for this use case is defined and all transition functions and merge operations are set in accordance to practical experiences. Second, we compare the results of the numeric approximation and the analytic assessment in order to find an optimal balance for that use case. The resulting value is finally validated with experiences for that use case gained in practical applications.

A. Use Case Outline

The system that is modeled as a data flow graph is a typical DR use case as described by Simmhan et al. in [11]. The purpose of DR is to curtail load during peak periods by requesting customers to reduce their energy demand for a certain period of time of a certain amount, e.g., by turning off or adjusting the HVAC units. In order to determine which action at which customer is most effective, a prediction model based on current and past energy usage is applied. This model, however, needs meter data of a customers' energy usage at a certain frequency, high enough for accurate predictions. Currently data granularity is one value each fifteen minutes, however, if necessary resolutions up to one value each minute are feasible. In practice the former is used in order to avoid fluctuations in data.

This setting implies two major privacy issues for customers, also addressed by Wicker and Schrader in [12]: (i) if the metering frequency is to high, information about the customer is revealed in (almost) real-time, e.g., if the customer presence at home can be predicted with high accuracy or even which devices are turned on; and (ii) metered data is stored in a database and it is therefore possible to maintain detailed profiles over time. Such information can be released to the public and may immediately affect the customer.

A graph representing this use case is depicted in Figure 2. N_1 represents a smart meter capturing data at a certain frequency, N_2 represents a database storing that data, N_3 represents a DR prediction unit and N_4 represents the goal *effective load curtailment*. The transitions and merging operations are defined as follows:

[1]http://smartgrid.usc.edu/

$$\bar{p} = \frac{1}{4}\left(2p_1 + \frac{e^{p_1}-1}{e-1} + \frac{1}{2}\left(\left(\frac{e^{p_1}-1}{e-1}\right)^3 + p_1\right)\right) \quad (4)$$

- T_2^1, the metering frequency has no operational impact for data storage in the data base. We are assuming a scalable database which can handle an arbitrary number of streams from meters at any frequency. This transition is therefore the identity function $p_2 = T_2^1(p_1) = p_1$.
- T_3^1, effective DR prediction heavily relies on a metering frequency that is close to real-time. A low frequency therefore reduces the operational capabilities of the prediction unit. This transition is therefore defined as $p_3 = T_3^1(p_1) = \frac{e^{p_1}-1}{e-1}$.
- T_4^2, there are no operational impacts for the overall goal of load curtailment on this path. This transition is therefore again the identity function $t_4^2 = T_4^2(p_2) = p_2$.
- T_4^3; if the operational capabilities of the DR prediction unit are low, the goal of load curtailment can not be achieved sufficiently. This transition therefore reduces the operational capabilities or increases the privacy: $t_4^3 = T_4^3(p_3) = (p_3)^3$.
- $\Theta_{2_{1,2,3}}$, for the sake of simplicity the merging operation is defined as the arithmetic mean by $\Theta_4(T_4^2, T_4^3) = \frac{1}{2}(T_4^2 + T_4^3)$.

All functions are bound in the interval $[0, 1]$, hence any value lower than 0 is mapped to 0 and any value greater 1 is mapped to 1.

B. Assessment

For the analytic assessment Equation 3 is applied to the above definitions. This yields Equation 4. Solving this equation for p_1 gives $p_1 \approx 0.59$.

Fig. 3 shows the results for the numeric approximation. Evaluation is performed with a step size $\Delta = 0.01$ for $\bar{p} \geq \frac{1}{2}$ and implemented in Matlab R2010b. The top plot shows the sum of the privacy requirements for each step, the middle plot shows the sum of the operational requirements for each step and the bottom plot shows the overlap of figures, indicating the intersection of the curves where the condition for S is first met. The greater equal condition is preferred over an equality condition in order to deal with numeric inaccuracies (the exact value of $\bar{p}$ might not be reached). Values for p_1 where the condition is met are indicated with a dotted line. The condition is first met at $p_1 \approx 0.59$ and therefore identical to the expected analytic result.

C. Interpretation

Once the assessment is performed, the resulting value, hence $p_1 \approx 0.4$, needs to be mapped to practical meaning. While this is heavily depending on the use case at hand, we propose the following approach for this scenario.

Electricity usage is continuous and digital (smart) metering is sampling that continuous signal at a certain frequency f_s. Following the Nyquist-Shannon sampling theorem [13], f_s needs to be at least twice as high as the highest frequency f_{max} in the signal in order to keep all the information of the original

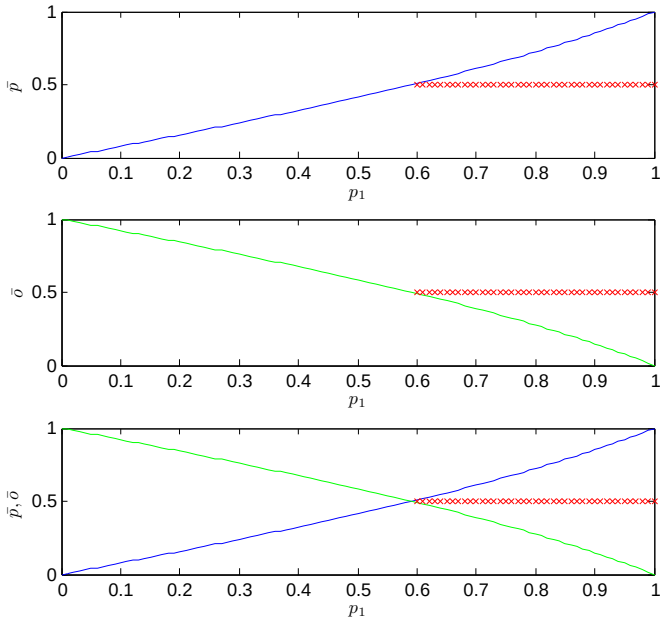

Fig. 3. Plot of the results of the numeric approximation for the use case applied for evaluation.

signal and to restore it losslessly. In practice a metering signal will consist of high frequencies due to peaks in the time domain, e.g., when switching on the light. Full operational capability is therefore given with f_s close to infinity and hence not feasible. If f_s approaches zero, by contrast, privacy is at its maximum. In practice, the upper bound for a meter frequency is given by physical limitations in data capturing and processing by e.g., $f_s = \frac{1}{5}$, hence one value each five seconds.

By describing this with a linear function yielding the privacy impact dependent on the frequency, with $p_1 = -5x + 1$ the intended outcome is achieved. Solving this equation for f_s and by replacing p_1 with 0.59 we get $\frac{0.59-1}{-5} = 0.082$ and thus a meter value approximately every 12.2 seconds. This metering frequency is the one – that based on the model – describes the optimum trade-off between the privacy requirements of the user and the designated goal *effective DR prediction*. Optionally, for a given metering frequency the impact on the goal can be determined, e.g., if f_s is given by $\frac{1}{10}$, this yields $p_1 = 0.5$ and by applying the transition functions and the merging operation $p_4 \approx 0.27$.

VI. Conclusion and Future Work

In this paper an approach has been presented that allows to assess the trade-off between privacy requirements and operational capabilities. Therefore a use case in the smart grid is modeled as a directed graph with nodes and edges. For edges transition functions and merging operations are defined. Based on that graph, an algorithm can be applied for finding the optimum balancing. This can be achieved by either solving an equation or – if this is not feasible – by using a numeric approximation. Finally, we proposed a mapping of the resulting values back to real-world applicability. For evaluation, a demand response use case from the USC microgrid was assessed and discussed.

Future work will focus on integrating this model into existing privacy assessment tools. This allows such systems to provide a more holistic assessment also taking into account the operational capabilities.

Acknowledgment

The financial support of the Josef Ressel Center by the Austrian Federal Ministry of Economy, Family and Youth and the Austrian National Foundation for Research, Technology and Development is gratefully acknowledged. Funding by the Austrian Marshall Plan Foundation is gratefully acknowledged. This material is based upon work supported by the United States Department of Energy under Award Number number DE-OE0000192, and the Los Angeles Department of Water and Power (LA DWP). The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof, the LA DWP, nor any of their employees.

References

[1] P. McDaniel and S. McLaughlin, "Security and privacy challenges in the smart grid," *Security Privacy, IEEE*, vol. 7, no. 3, pp. 75–77, May 2009.

[2] A. Cavoukian, J. Polonetsky, and C. Wolf, "Smartprivacy for the smart grid: embedding privacy into the design of electricity conservation," *Identity in the Information Society*, vol. 3, no. 2, pp. 275–294, 2010.

[3] C. Neureiter, G. Eibl, A. Veichtlbauer, and D. Engel, "Towards a framework for engineering smart-grid-speficic privacy requirements," in *Proc. IEEE IECON 2013, Special Session on Energy Informatics*. Vienna, Austria: IEEE, November 2013.

[4] B. Chen, Z. Kalbarczyk, D. Nicol, W. Sanders, R. Tan, W. Temple, N. Tippenhauer, A. Vu, and D. Yau, "Go with the flow: Toward workflow-oriented security assessment," in *Proceedings of New Security Paradigm Workshop (NSPW)*, Banff, Canada, September 2013.

[5] E. McKenna, I. Richardson, and M. Thomson, "Smart meter data: Balancing consumer privacy concerns with legitimate applications," *Energy Policy*, vol. 41, pp. 807–814, 2012, modeling Transport (Energy) Demand and Policies.

[6] S. Oliveira and O. Zaiane, "Algorithms for balancing privacy and knowledge discovery in association rule mining," in *Database Engineering and Applications Symposium, 2003. Proceedings. Seventh International*, July 2003, pp. 54–63.

[7] D. Massaguer, B. Hore, M. Diallo, S. Mehrotra, and N. Venkatasubramanian, "Middleware for pervasive spaces: Balancing privacy and utility," in *Middleware 2009*, ser. Lecture Notes in Computer Science, J. Bacon and B. Cooper, Eds. Springer Berlin Heidelberg, 2009, vol. 5896, pp. 247–267.

[8] CEN, Cenelec, and ETSI, "Smart Grid Reference Architecture," CEN/Cenelec/ETSI Smart Grid Coordination Group Std., Tech. Rep., November 2012.

[9] C. Dänekas, C. Neureiter, S. Rohjans, M. Uslar, and D. Engel, "Towards a model-driven-architecture process for smart grid projects," in *Digital Enterprise Design & Management*, ser. Advances in Intelligent Systems and Computing, P.-J. Benghozi, D. Krob, A. Lonjon, and H. Panetto, Eds. Springer International Publishing, 2014, vol. 261, pp. 47–58.

[10] F. Knirsch, D. Engel, C. Neureiter, M. Frincu, and V. Prasanna, "Model-driven Privacy Assessment in the Smart Grid," Josef Ressel Center for User-Centric Smart Grid Privacy, Security and Control, Tech. Rep., January 2014.

[11] Y. Simmhan, Q. Zhou, and V. Prasanna, "Semantic information integration for smart grid applications," in *Green IT: Technologies and Applications*, J. H. Kim and M. J. Lee, Eds. Berlin Heidelberg, Germany: Springer, 2011, pp. 361–380.

[12] S. Wicker and D. Schrader, "Privacy-aware design principles for information networks," *Proceedings of the IEEE*, vol. 99, no. 2, pp. 330–350, Feb 2011.

[13] H. Nyquist, "Certain topics in telegraph transmission theory," *Proceedings of the IEEE*, vol. 90, no. 2, pp. 280–305, Feb 2002.