Towards a Privacy-Aware Information-Sharing Framework for Advanced Metering Infrastructures

Final Project Report

Power Systems Engineering Research Center

Empowering Minds to Engineer
the Future Electric Energy System

Towards a Privacy-Aware Information-Sharing Framework for Advanced Metering Infrastructures

Final Project Report

Project Team

Vinod Namboodiri, Project Leader
Visvakumar Aravinthan, Murtuza Jadliwala
Wichita State University

Lalitha Sankar
Arizona State University

PSERC Publication 15-06

September 2015

For information about this project, contact

Vinod Namboodiri
Associate Professor
Electrical Engineering and Computer Science
Wichita State University
Email: vinod.namboodiri@wichita.edu
Phone: 316-978-3922

Power Systems Engineering Research Center

The Power Systems Engineering Research Center (PSERC) is a multi-university Center conducting research on challenges facing the electric power industry and educating the next generation of power engineers. More information about PSERC can be found at the Center's website: http://www.pserc.org.

For additional information, contact:

Power Systems Engineering Research Center
Arizona State University
527 Engineering Research Center
Tempe, Arizona 85287-5706
Phone: 480-965-1643
Fax: 480-965-0745

Notice Concerning Copyright Material

Acknowledgements

This is the final report for the Power Systems Engineering Research Center (PSERC) research project titled "Towards a Privacy-Aware Information-Sharing Framework for Advanced Metering Infrastructures" (project S-54). We express our appreciation for the support provided by PSERC's industry members and by the National Science Foundation under the Industry / University Cooperative Research Center program.

We wish to thank Mirrasoul Mousavi from ABB who was one of the industry advisors on the project. He participated in numerous conference calls and shared industry's perspective on our project's goals. His comments and suggestions for this final report were invaluable. We also would like to thank Debbie Brodt-Giles from the National Renewable Energy Laboratory (NREL), and Slobodan Matic from GE for serving as industry advisors for the full term, and Anant Venkateswaran (formerly with GE), Nivad Navid (formerly with the Midcontinent ISO), Rameet Kohli (formerly with GE), and Aaron Beach (formerly with NREL) for serving as advisors for a part of the project period.

This project also benefited greatly from the assistance of Dennis Ray, Deputy Director of PSERC. He kept us on track with constant reminders of deadlines and offered answer to any questions we had. We would also like to thank Ward Jewell from Wichita State University and Kory Hedman from Arizona State University for offering guidance about the PSERC process of funded projects.

Last, but not the least, we would like to thank all the students that worked on this project.

Executive Summary

The information communication and control layer of the smart grid brings about numerous advances, including the empowerment of customers to actively participate in the maintenance of the supply-demand balance around the clock and the resulting reliability improvement in electricity service. Such customer participation is made possible by the Advanced Metering Infrastructure (AMI) which has the capability to support many functions beyond billing. The success of this vision depends on the effective design and implementation of a reliable and economically sustainable information-sharing infrastructure.

Many AMI deployments by grid operators are emerging around the U.S, Europe, and Asia, each of them demonstrating different levels of maturity. These deployments are capable of data collection by employing various technologies, but have not been backed up by an effective framework to share data and information. The design of an information-sharing framework for the AMI and associated home area networks (HANs) to meet smart grid requirements is still an open research problem, with both further encumbered due to the stringent requirement of ensuring customer privacy. On one hand, utilities need to collect low-level customer data to improve operational planning and control. On the other hand, customers have privacy preferences which need to be met to encourage greater smart meter adoption rates that in turn benefits utilities. An ideal information-sharing framework will allow a customizable level of data collection to meet specific customer privacy requirements within the context of the AMI.

In this project we worked on solving some of these challenges towards building such an information-sharing framework for the AMI-enabled communication/control/information layer with special emphasis on customer privacy and its potential impact on smart meter adoption and utility operations. As part of the design of such a privacy-aware information-sharing framework, we studied the interactions between the data collection needs of the utility and the preservation of privacy.

The report and its main contributions have four parts.

Part I: Scalable Meter Data Collection in Smart Grids through Message Concatenation

This part addresses the looming issue of how to communicate and handle consumer data collected by electric utilities and manage limited communication network resources. This part of the project studied the smart meter message concatenation (SMMC) problem of how to efficiently concatenate multiple small smart metering messages arriving at data concentrator units (DCUs) in order to reduce protocol overhead and thus network utilization. This work provides hardness results for the SMMC problem, proposes six heuristics, and evaluates them to gain a better understanding of the best data volume reduction policies that can be applied at data concentrators of AMI infrastructures.

Our results indicate that the proposed heuristic-based concatenation algorithms can reduce data volume in the range of 10-25% for typical backhaul technologies used, with greater benefits seen for scenarios with higher data traffic rates. These benefits are obtained operating only on packet headers without compressing or aggregating the

underlying information in messages. Our results are also shown to hold up well under various practical issues such as network and processing delays, tighter application deadlines, and lossy backhaul links.

Part II: Impacts of Communication and Control on Distribution System

The purpose of this part of the project was to understand the impacts of aggregation on distribution system. The primary focus of this work revolved around two directions.

a) Impact of control frequency on demand management and consumer comfort

This work focused on evaluating the impacts of price incentive-based load control on distribution level transformers and consumer comfort. The results obtained show that both consumer level impacts (e.g., deviation from set temperature) and grid level impacts (violation in terms of power and energy) depend on the control interval used.

b) Estimation error analysis due to aggregation interval

The focus of this part of the work was in determining the link between the data and power network layers. A methodology to quantify the relationship between the data aggregation interval and the prediction/estimation accuracy was developed using the design of experiments technique. The results showed that for a particular distribution system topology, the data aggregation interval has a significant effect on predicting the number of tap changes or the power loss. Finally, a polynomial function was developed to determine the estimation error.

Part III: Preserving Privacy of Advanced Metering Data using Efficient Aggregation and Prediction Techniques

In this part of the project, we propose novel and efficient techniques for privacy-preserving data aggregation in smart grid communication networks.

a) AgSec: Secure and Efficient CDMA-based Aggregation for Smart Metering Systems

Most existing security mechanisms utilize cryptographic techniques that are computationally expensive and bandwidth intensive. However, aggregating the large outputs of these cryptographic algorithms has not been considered thoroughly. Smart Grid Networks (SGN) generally has limitations on bandwidth, network capacity, and energy. Hence, utilizing data aggregation algorithms, the limited bandwidth can be efficiently utilized. In this work a CDMA-based data aggregation method is proposed that provides access to all the data of all the smart meters in the root node, which in this case is the utility control center, while keeping the smart metering data secure. The efficiency of the proposed method is confirmed by mathematical analysis.

b) Seer Grid: Privacy and Utility Implications of Two-Level Energy Load Prediction in Smart Grids

Energy consumption signatures present in the data reported by smart meters to the control center can pose privacy risks for customers. A popular solution in the research and academic literature to overcome these privacy threats is to perturb the actual energy usage data before sharing it with the control. The degree of correlation between the actual energy usage data and the perturbed data produced by the perturbation technique

typically characterizes the trade-off between the privacy requirement (of the customer) and data utility or data usefulness requirement (of the control center). A larger tradeoff implies one requirement (privacy or data-utility) is given more preference over the other. The main goal of this research is to propose a mechanism to minimize this trade-off, i.e., provide both reasonable levels of privacy protection as well as data-utility. A two-level prediction mechanism is proposed to preserve the correlation between the predicted and actual energy consumption patterns at the cluster or neighborhood level and removes this correlation in the predicted data communicated by each smart meter to the control center. The two-level prediction mechanism was evaluated using real smart meter data and showed our proposed mechanism to be successful in hiding private consumption patterns at the household-level while still being able to accurately predict energy consumption at the neighborhood-level.

Part IV: Incentivizing Privacy-Guaranteed Data-Sharing by Consumers using AMI

The goals of this part of the project were three-fold: (a) understanding whether the AMI data collected by utility companies or third parties can lead to privacy violations such as making inferences about specific consumer behavior; (b) determining incentive strategies that utility companies can use to ensure large-scale adoption of AMI technologies (even as a large number of end-users in the grid adopt renewable energies, e.g., PV); and (c) evaluating the effect of cyber-attacks on data integrated from end-user AMIs into the transmission system. The following tasks summarize these efforts.

a) Incentivizing consumers with access to renewables

In this task, the objective was to determine incentives that utility companies can use to encourage households that have access to renewables to consume a minimal amount of energy directly from the grid. The main contribution of this task was to propose a novel approach using non-cooperative game theory for N-person strategic games to study the tradeoff between privacy and energy cost minimization under the assumption that the utility company offers incentives to multiple households to encourage data sharing through energy consumption. In this respect, we formulate a non-cooperative game to model interactions between households and the utility company.

b) The tradeoff between privacy leakages due to inferences from AMI data and consumer benefits from sharing

In this task, we asked the question: can consumers use alternate energy sources to both gain in costs and achieve privacy? To be more specific, we studied the tradeoff from using battery/PV to achieve energy cost savings versus using them specifically for retaining a certain measure of privacy. Based on the meter data, Neyman-Pearson hypothesis testing was performed to estimate and provide guarantees against inference for the private feature.

c) Modeling AMI cyber-attacks as restrictions on information access at the transmission level

In addition to privacy, a natural question arises on the security of AMI data. We assume that AMI data will be encrypted; however, the sophistication of cyber-attacks suggests that the data can be compromised. However, modeling the effect of a large scale man-in-

the-middle (MitM) attack (changing data en route to the data center) on AMI data is not straightforward. To this end, we model this as an MitM attack for transmission network – the model assumes that an attacker changes data enough to change topology shared between two areas – this is a very coarse abstraction of a possible consequence of a large scale attack on AMI data, but it is a start. Our results then looked at the consequences on power systems operations.

Overarching Conclusions

This project has led to the contribution of effective techniques (as part of a broader information-sharing framework) to manage the tradeoffs between (i) data volume and communications network capacity, and (ii) data needs of utilities and customer privacy, and (iii) application accuracy/quality and data intervals. Utilities can adopt some or all of these techniques for their information-sharing framework to manage the tradeoffs they encounter.

Based on our research, we offer several recommendations for electric utilities.

1) It is best to design and allow for a tunable data collection framework that can adapt to communication network constraints dynamically while preserving application quality. Current communication networks of utilities are rarely dedicated and well thought out in terms of future data carrying capacities. Electric utilities will need to dedicate more resources to the planning and design of communication networks in the future.
2) Any data collection from customers should employ some of the proposed techniques in this project to be able to maximize benefits from data collected while alleviating customer privacy concerns. Customer privacy does not seem to figure in a lot of the design and planning from the electric utility side; however, including it in earlier than later will allow greater flexibilities and targeted data collection to meet application needs of the future.
3) Control frequency and aggregation interval need to be carefully thought out in terms of their power system impacts as these were shown to impact application quality and accuracy. Current intervals do not have a range that can be tuned to adapt to application needs or communications network constraints.

Though this project solves some challenges in the area of data management in smart grids, there remain other challenges. One such challenge is that of determining what data is needed where, and in what granularity. Solving this challenge will help alleviate some data volume concerns within communications networks while ensuring applications at various locations have the information they need. The second challenge is that of designing more resilient cyber and power networks; this work only looked at the AMI scenario and there remain many other areas (such as energy management systems) where more work needs to be done especially in determining how AMI data can be integrated to make better real-time control and decisions.

Project Publications

Journal Articles

1. Karimi, B., Namboodiri, V., Jadliwala, M., "Scalable Meter Data Collection in Smart Grids through Message Concatenation," *IEEE Transactions on Smart Grids*, vol. 6, pp. 1697–1706, 2015.
2. Namboodiri, V.; Aravinthan, V.; Mohapatra, S.N.; Karimi, B.; Jewell, W., "Toward a Secure Wireless-Based Home Area Network for Metering in Smart Grids*," IEEE Systems Journal*, vol.8, no.2, pp.509-520, June 2014.
3. Karimi, B.; Namboodiri, V., "On the Capacity of a Wireless Backhaul for the Distribution Level of the Smart Grid," *IEEE Systems Journal*, vol.8, no.2, pp.521-532, June 2014.
4. Zhang, J.; Sankar, L., "Consequences of a Cyber-Physical Topology Attack on Power System Operations," submitted, *IEEE Trans. Smart Grid*, July 2015.

Conference Proceedings

1. V. C. Dev, U. Das, V. Namboodiri, S. Chakraborty, V. Aravinthan, Y. Guo, and A. Srivastava, "Towards application-aware data concentration schemes for advanced metering infrastructures," accepted, 2015 *IEEE SmartGridComm*, November 2015.
2. Alamatsaz, N., Boustani, A., Jadliwala, M. and Namboodiri, V., "AgSec: Secure and Efficient CDMA-based Aggregation for Smart Metering Systems," *in proceedings of IEEE CCNC*, 2014.
3. Karimi, B., Namboodiri, V., Jadliwala, M. "On the Scalable Collection of Metering Data in Smart Grids through Message Concatenation," *in proceedings of the IEEE SmartGridComm*, 2013.
4. Hu, J., and Sankar, L. "Cluster-and-Connect: A More Realistic Model for the Electric Power Network Topology," accepted, *IEEE SmartGridComm*, Nov. 2015.
5. Zhang, J., and Sankar, L., "Consequences of Man-in-the-Middle Topology Attacks on Power Systems Operations,", in *Proceedings CIGRE Grid of the Future Symposium*, Houston, TX, Oct 19-21, 2014.

Working Papers (latest copy can be obtained by email request)

1. Boustani, A., Jazi, S.Y., Maiti, A., Jadliwala, M. and Namboodiri, V., Seer Grid: Privacy and Utility Implications of Two-Level Energy Load Prediction in Smart Grids, working paper
2. Khan, Z., Jadliwala, M., Sinha, K. and Salari, E., A Unified Framework for Evaluating Smart Meter Data Perturbation Mechanisms, working paper
3. Huang, C., and Sankar, L., "Tiered pricing models to incentivize consumers with access to renewables to consume energy from the grid", working paper.

4. Huang, C., and Sankar, L., "Tiered pricing models to incentivize consumers with access to renewables to consume energy from the grid", working paper.
5. Huang, C., and Sankar, L., "Guarantees on inference attacks on AMI data for consumers access to renewables", working paper.

Student Theses

1. Babak Karimi. Ph.D. Dissertation. Capacity Analysis and Data Concentration for Smart Grid Communication Networks at the Power Distribution Level, Wichita State University, August 2014. Advisor: Vinod Namboodiri.
2. Arash Boustani. Ph.D. Dissertation Improving security, capacity and efficiency of wireless communications in modern cyber-physical systems, Wichita State University, Expected completion date summer 2016. Advisor: Murtuza Jadliwala.
3. C. Huang, Decisions for Privacy Sensitive Interactions, Ph.D qualifying exam Report, Arizona State University, Sep. 2015. Advisor: Lalitha Sankar, Arizona State University.
4. Navid Alamatsaz M.S. Thesis. Towards an Analytical Framework for Privacy-Preserving Aggregation in Smart Grid, Wichita State University. May 2014. Advisor: Murtuza Jadliwala.
5. Muhammad Usman Khan. M.S. Thesis. Impact of Control Frequency on Transformer Level Demand Management and Consumer Comfort, Wichita State University, May 2015. Advisor: Visvakumar Aravinthan.
6. Suvagata Chakraborty. M.S. Thesis. Estimation Error Analysis due to Aggregation Interval in Smart Distribution Feeders, Wichita State University, May 2015. Advisor: Visvakumar Aravinthan.
7. J. Zhang, "Unobservable Topology Attacks and Consequences on Power System Operations," Masters Thesis, Arizona State University, Jul. 2015. Advisor: Lalitha Sankar, Arizona State University.
8. Zoya Khan, "A Unified Framework for Evaluating Smart Meter Data Perturbation Mechanisms" (Expected completion date: Spring 2016).

Part I

Scalable Meter Data Collection in Smart Grids through Message Concatenation

Vinod Namboodiri and Murtuza Jadliwala

Babak Karimi and Vishnu Cherusola Dev
Graduate Students

Wichita State University

For information about Part I, contact:

Vinod Namboodiri
Associate Professor
Electrical Engineering and Computer Science
Wichita State University
Email: vinod.namboodiri@wichita.edu
Phone: 316-978-3922

Power Systems Engineering Research Center

The Power Systems Engineering Research Center (PSERC) is a multi-university Center conducting research on challenges facing the electric power industry and educating the next generation of power engineers. More information about PSERC can be found at the Center's website: http://www.pserc.org.

For additional information, contact:

Power Systems Engineering Research Center
Arizona State University
527 Engineering Research Center
Tempe, Arizona 85287-5706
Phone: 480-965-1643
Fax: 480-965-0745

Notice Concerning Copyright Material

Contents

List of Tables

List of Figures

1 Introduction

The information communication and control layer of the smart grid brings about numerous advances, including the empowerment of customers to actively participate in the maintenance of the supply-demand balance around the clock and the resulting reliability improvement in electricity service. There are many benefits to grid operators, consumers, and society as a whole from adopting advanced metering infrastructure (AMI) technologies [1]. With the introduction of AMI technology, two-way communication between a "smart" meter and the grid operator's control center, as well as between the smart meter and consumer appliances, would be facilitated for various applications [2]. Besides AMI, there are many other applications that will be enabled by information flow across the electric power grid. These include distributed generation, state estimation of the power distribution system, demand-side management, to name a few.

A big challenge for smart grid application scenarios, and the information-sharing framework that enables them, will be handling the massive amount of data that is expected to be collected from data generators and sent through the communication backhaul to the grid operator. For example, by current standards, each smart meter sends a few kilobytes of data every 15-60 minutes to grid operators [3, 4]. When this is scaled up to many thousands, existing communication architectures will find it difficult to handle the data traffic due to the limited network capacities, especially in limited bandwidth last mile networks [5, 6]. Future applications may require data to be collected at a finer granularity, thus adding to the challenge [7]. Network capacity is a precious resource for electric utilities because they are either leasing such networks from third-party providers [8], or building infrastructure themselves and leasing bandwidth out (especially at the backhaul) to recuperate investment costs [9]. In either case, it is in the interest of electric utilities to reduce the volume of information transported through these networks for smart grid applications while ensuring QoS requirements are met.

One approach to reduce data volume given some application sampling rate is to concatenate multiple messages into a larger packet to reduce protocol overhead due to packet headers. This approach has the potential to reduce network capacity requirements significantly (quantified later in this part of of the report) due to the small size of messages sent in smart metering networks, with packet headers possibly being of a comparable size to the underlying message to be sent. Such concatenation of messages can be done by each smart meter itself. However, each meter may not generate messages frequently enough to be able to have the chance to concatenate enough packets to reduce overheads significantly and also meet their stated application deadlines. Each meter is also expected to be relatively constrained (compared to a concentrator) in terms of data storage capabilities to keep a large window of packets from which to aggregate. Thus, a better approach is to concatenate messages at an intermediate point upstream from individual meters.

Such an intermediate point where message concatenation can be done is at data concentrator units (DCUs) (or some similar entity, sometimes also called a data aggregator) that collect data from many smart meters and forward them upstream. Figure 1.1 depicts this concept and shows the DCU's role at the power-distribution level of the power grid. Data concentrators or aggregators can play an important role in reducing network capacity requirements by reducing packet protocol overhead through message concatenation algorithms applied along the data collection tree. Such algorithms and policies, however, do not exist currently and need to be developed keeping in mind the unique characteristics of metering data like variable packet sizes, stochastic arrivals, and the presence of messages with and without deadlines. Current DCUs on the market lack the ability to reduce the volume of data flowing through them and real-time aggregation capabilities. They only provide simple integration of sensing and WAN communications options with the intention to follow the PRIME standard [10] which gives the utilities the freedom to choose meters from various vendors and avoid being reliant on proprietary solutions from a single source.

In this part of the report we design and comparatively evaluate a suite of online message concatenation algorithms at DCUs in the AMI scenario that minimize usage of network capacity in transporting data through the meter data collection network while meeting quality-of-service (QoS) constraints imposed by applications on individual messages. The specific contributions of this work include:

1. A formulation of the message concatenation problem at DCUs in smart metering networks to minimize network capacity utilization

2. Hardness results for the formulated message concatenation problem that proves it as NP-complete

3. Six different heuristic-based algorihms that can be employed at DCUs for the message concatenation problem

4. A comparative performance evaluation of proposed heuristic-based algorithms for message concatenation

5. Exploration of feasibility of message concatenation under practical settings considering network and processing delays, tighter application deadlines, and lossy backhaul links

Our results indicate that the proposed heuristic-based concatenation algorithms can reduce data volume in the range of 10-25% for typical backhaul technologies used, with greater benefits seen for scenarios with higher data traffic rates. These benefits are obtained operating only on packet headers without compressing or aggregating the underlying information in messages. Our results are also shown to hold up well under various practical issues such as network and processing delays, tighter application deadlines, and lossy backhaul links.

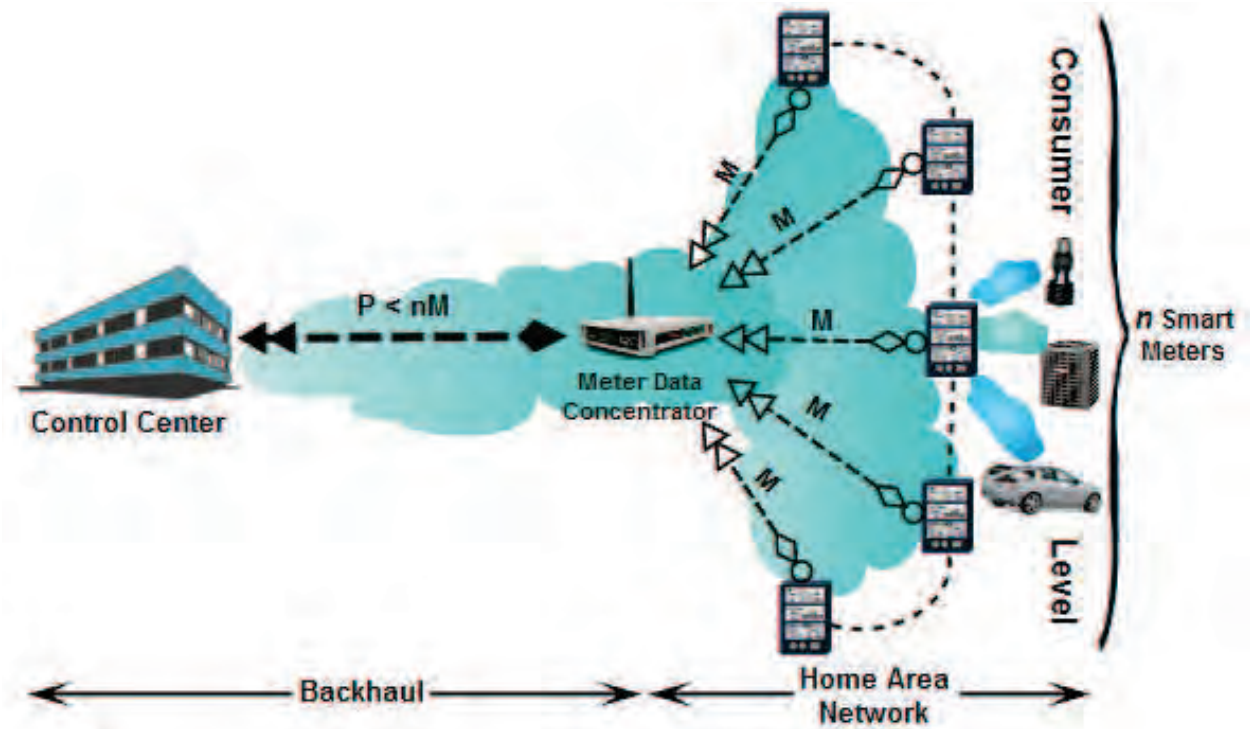


Figure 1.1: Data Concentrator Unit's envisioned role of message concatenation at the power distribution level.

2 Problem Formulation

2.1 Motivation

In most communication protocol suites (e.g. TCP/IP) used for sending smart metering messages, the small size of packets will result in a high amount of protocol overhead due to packet headers. For example, for messages of size 100 bytes from the source smart meter, there may be 40-60 bytes of additional header overheads due to TCP/IP protocols and specific versions used. If a data concentrator collects multiple packets and strips off all individual headers and includes only one header for the larger aggregated message, there could be significant reductions in network capacity utilization. Studying the messaging format for the ANSI C12 smart meter communications standard in [11] provides an idea of message sizes involved and the amount of protocol overhead to expect. As shown in Figure 2.1, each smart meter generated message includes parameters like meter identification number, equipment status, type of message, among others. This information is enough to uniquely identify a message source with no additional protocol header information required for source identification. Thus, source protocol headers can be stripped away to rely only on a common aggregated packet header to route the packet to the destination.

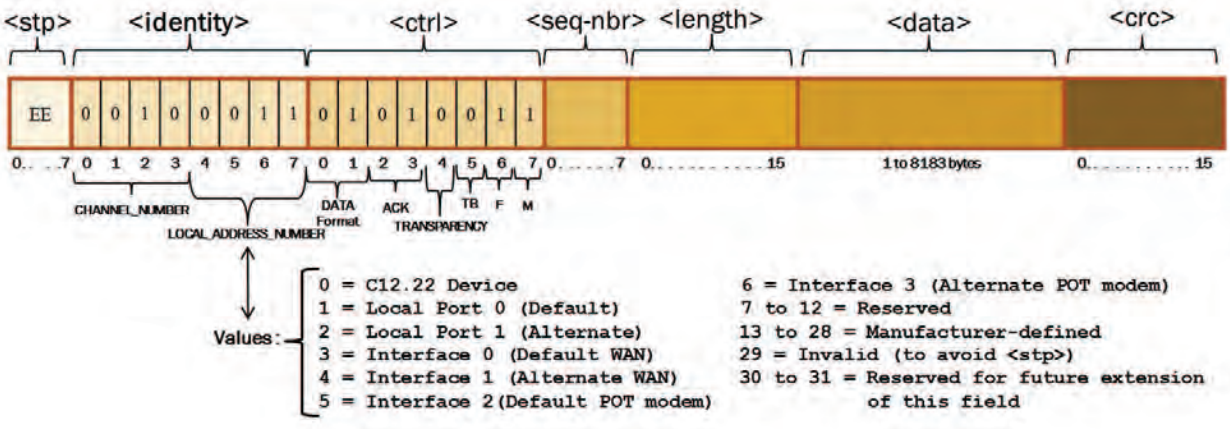


Figure 2.1: Smart meter datagram structure.

In Table 2.1 (abstracted from [5]) basic message types along with their properties are listed. It can be seen that messages can be of various sizes (from 20-500 bytes), and can have loose or strict deadlines (2-5 seconds), or no deadlines at all. Some messages may be generated randomly at any time to indicate critical events that need to be responded to immediately. Data concentrators will have the challenge of handling these varying message sizes that may or may not have deadlines, with possibly stochastic arrivals, at the same time

guaranteeing that each message meet any specified deadline. Stochastic message generation and critical events with short deadlines exclude the use of polling based algorithms to collect data at DCUs.

Table 2.1: Smart meter data message types

Message/ Traffic Description	**Size (Bytes)**	**Inter-arrival interval**	**Inter-arrival unit**	**Delay Objective**
Meter clock sync.	64	Day	1	2 secs
Interval data read	480	Day	1	Best effort
Firmware patch/ upgrade confirmation/ acknowledge	20	Year	1	Best effort
Meter ping (on demand read)	64	Week	4	2 secs
Meter remote diagnostic	500	Day	4	2 secs
Tamper notification	64	Week	26	5 secs
Meter remote disconnect/ reconnect response	500	Day	1	2 secs

2.2 Related Work

There have been much prior work on data aggregation in the field of WSNs [12]. Typical approaches to WSNs have focused on efficient data gathering and energy-latency tradeoffs under deadline constraints (e.g. [13, 14, 15, 16, 17, 18]). These schemes propose algorithms for grouping smaller packets into larger ones by delaying data transmissions at the relaying nodes whenever slack times are positive with significant reductions in packet transmissions, congestion, and battery energy use. In this project, our goal is similar in proposing data concentration at DCUs as relay nodes. However, power or energy consumption of the nodes employed are not considered because the AMI infrastructure is expected to have access to electric power at all times with backup batteries. This shifts the focus of the problem from battery life of nodes involved to the reduction of network capacity utilization. The work in [6] does look at data volume reduction in smart metering networks, but does not include aspects such as message concatenation and the application of aggregation functions.

Another direction of related work has been in terms of designing a reliable, flexible, and cost-effective data concentrator [19, 20]. Many studies have considered the problem of data concentration for synchrophasors (e.g. [21]). The latest activities in standardization of Wide Area Monitoring, Protection, And Control (WAMPAC) systems, and design and implementation issues, such as maintaining time-sync at PMUs, missing phasor data frames, handling multiple input data rates and latency from PMUs, etc. with data concentrators are discussed in [22]. This work on the other hand designs data concentration algorithms specifically for smart metering and reduce information volume through the network.

Scheduling under deadlines poses well-known challenging problems with many new applications. It was shown by Karp [23] that optimal offline scheduling for problem of deadline scheduling is NP-complete. On the other hand, simple online scheduling algorithms that achieve the best competitive ratio do exist. For example, the earliest deadline first (EDF) algorithm works on the job with the earliest deadline, and it switches to a newly arrived job if the new arrival has an earlier deadline. It is known that such a simple scheduling algorithm is optimal when the traffic load is light. See in particular the seminal work of Liu and Layland [24], the work of Mok [25] and Locke [26], recent applications in scheduling jobs for cloud systems [27] and large-scale EV charging [28]. In this work, EDF-based online algorithms are developed for use at the DCU.

Finally, from an information needs perspective, there has been recent work on a futuristic approach to information-sharing mechanisms in smart grids, including at the distribution level [29, 30, 31, 32, 33]. The GridStat effort [34, 35, 36], primarily from Washington State University, has set about defining communication requirements for power grids for the last 5-10 years. GridStat has further inspired the NASPINet effort to develop an "industrial grade," secure, standardized, distributed, and expandable data communications infrastructure to support synchrophasor applications in North America [37]. None of the prior work in the area has looked at information needs of grid operators from a purely information volume perspective and its impact on the design of a communications network for the distribution system.

The topic of information needs also raises the question of information security and consumer privacy, which has been an area of research [38, 39, 40, 41, 42, 43, 44, 45]. Some of these approaches, although providing strong guarantees of confidentiality, are very taxing from a communication and computational stand-point and may not be feasible on low-end smart meters. Given the frequency of the data being sent and possible bandwidth limitations, this can lead to unacceptable delay and network overhead, and needs mitigating mechanisms. This complementary task of reducing overhead introduced by security and privacy mechanisms for smart metering is presented as a separate part of this report.

2.3 The Smart Metering Message-Concatenation Problem

The smart metering message concatenation (SMMC) problem considered in this part of the report is as follows. A DCU receives different types of messages from smart meters with a stochastic arrival process (we will discuss this arrival process later in Section 4). Each message can be of a different size and comes with an application specific end-to-end deadline by which it must reach the common destination that is the utility control center. Each message has protocol overhead as it is packaged into a packet before being sent to the DCU. The DCU can either send each packet to the destination as it arrives as a single message or wait and concatenate multiple messages before sending them out over the backhaul to the destination. The objective considered is to minimize the number of individual packets (and hence protocol overhead) sent upstream by the DCU so as to reduce network capacity requirements of the backhaul. The constraints are that all packets meet their deadline (if any) and that each concatenated packet generated (including a common packet header) has a upper size limit, W, governed by the maximum transmission unit (MTU) of the upstream

link from the DCU. The objective function chosen helps reduce total overhead required to send all messages within a given time period T by maximizing the size of each concatenated packet for a fixed header size H. In this work we assume that messages are not compressed from their original sizes (zero-compression) and the solution to the SMMC problem at DCUs would serve as a lower bound for the possible reduction in network utilization by additional schemes (possibly that compress message sizes themselves) developed in the future for the smart metering scenario. We focus on only a single DCU and its concatenation operation in this work; in future work we envision considering a more wider view of the backhaul network and the use of multi-level DCUs along the communications network.

A formal statement of the SMMC problem is provided in the following definition.

Definition 1. *Assume that over some period of time T, all smart meters together generate n messages $M = \{m_1, \cdots, m_n\}$. Each message $m_i \in M$ has size s_i and an associated protocol header h_i accompanying it till the DCU with (s_i, h_i, $s_i + h_i \in [0, W]$), an arrival time at the DCU of a_i ($a_i \in [0, T]$), and a deadline d_i ($d_i \in [a_i, \infty]$) by which it must leave the DCU, where $i = 1 \cdots n$. Then, the SMMC problem is to determine an integer number of packets $k (k \leq n)$ and a k-partition $P_1 \cup P_2 \cup \cdots \cup P_k$ of the set M such that (i) $\sum_{i \in P_j} s_i + H \leq W$, $\forall j = 1 \cdots k$, and (ii) each message $m_i \in M$ meets its deadline with $\max_{i \in P_j} a_i \leq \min_{i \in P_j} d_i$. A solution is optimal if it has minimal k.*

The SMMC problem can also be stated as a $0-1$ Integer Linear Program (ILP) as follows:

$$\text{minimize } \ k = \sum_{i=1}^{n} y_i \tag{2.1}$$

subject to constraints

$$\sum_{j=1}^{n} s_j x_{ij} + H \leq W y_i, \quad \forall i \in \{1 \cdots n\}$$

$$\max \ a_j x_{ij} \leq \min \ d_j x_{ij}, \quad \forall i \in \{1 \cdots n\}, j \in \{1 \cdots n\}$$

$$\sum_{i=1}^{n} x_{ij} = 1, \forall j \in \{1 \cdots n\}$$

$$y_i \in \{0, 1\}, \forall i \in \{1 \cdots n\}$$

$$x_{ij} \in \{0, 1\}, \forall i \in \{1 \cdots n\}, \forall j \in \{1 \cdots n\}$$

where $y_i = 1$ if packet i is used and $x_{ij} = 1$ if message j is put into packet i.

In the formulations above, the term deadline refers to the local deadline for a message at the DCU by which a particular message must be picked up for the packet creation and transmission over the network. This local deadline can be set by subtracting away an estimate of processing delay at the DCU and the network delay over the backhaul from the end-to-end deadline specification of an application for messages. We will discuss and incorporate the impact of processing and network delays later in Section 5. In the problem definition above, for any set of messages assigned to a packet, none of the messages in the packet will miss their local deadlines at the DCU if the arrival times of all messages are at least some value

ϵ before the first expiring deadline value among all messages of that set. This value ϵ could be set to the maximum processing delay to be encountered at the DCU in forming a packet and could be an input to the problem; more discussion about estimation of processing delays will be presented in Section 5.

3 Algorithms for the SMMC Problem

3.1 SMMC Hardness Result

To prove that the SMMC problem is NP-Complete we first show that SMMC is in NP, or in other words, has a polynomial time verifier. An instance of a solution to the SMMC problem is an integer number of packets k and a feasible k-partition $P_1 \cup P_2 \cup \cdots \cup P_k$ of the set of messages M. Such an instance can be verified in polynomial time in terms of the input consisting of the following fields `<message identifier, arrival time, deadline, message size, header size,` W`>` for n messages. Further, in polynomial time (in terms of input length) we can check that each message falls in exactly one of the k partitions/packets, and that each packet meets the condition of having its total size less than or equal to W. We can further check in polynomial time if any message in the packet will miss its local deadline. Thus, we can verify whether a given instance is a solution to SMMC in polynomial time, and hence, SMMC $\in$ NP.

To prove that the SMMC problem is NP-hard we reduce the known NP-complete Bin Packing problem [46] to the SMMC problem. These problems have many similarities but differ in terms of the notion of arrival times and deadlines for the SMMC problem. The Bin Packing problem takes as input a set of $n^{'}$ items $I = \{it_1, \cdots, it_{n'}\}$ of sizes $S^{'} = \{s_1^{'}, s_2^{'}, \cdots, s_{n'}^{'}\}$ and a set of bins $B = \{b_1, \cdots, b_{k'}\}$ each of size $W^{'}$. An assignment of items to bins is sought that minimizes the number of bins $k^{'}$ into which all items are packed. That is we seek a $k^{'}$-partition $B_1 \cup B_2 \cup \cdots \cup B_{k'}$ of the set of items I.

We will transform an instance of the Bin Packing problem to that of the SMMC problem as follows. For each item i in I, add dummy variables $A^{'} : a_i^{'} = 0$, and $D^{'} : d_i^{'} = \infty$. This transformation can be trivially done in polynomial time (in terms of input length) and the modified instance used as an input to the SMMC problem with $M = I$, $S = S^{'}$, $D = D^{'}$, $A = A^{'}$, $W = W^{'}$, and $P = B$.

Any resulting solution from the SMMC problem can be transformed back to a solution for the Bin Packing problem as follows. A solution to the SMMC problem gives an integer k and a k-partition of M that maps individual messages to specific concatenated packets. We can take this solution and apply the following transformation: $k^{'} = k$ and $B_i = P_i$, $i = 1 \cdots k$. This transformation gives the required solution assignment for the Bin Packing problem and can be easily done in polynomial time again.

Theorem 1. *SMMC is NP-complete.*

Proof. By transforming (in polynomial time) any input instance of the Bin Packing problem to that of an SMMC problem, and the resulting solution of the SMMC problem back to Bin Packing problem, we have thus reduced Bin Packing to SMMC. Thus, SMMC is an NP-hard

problem. And since we had proved SMMC $\in$ NP earlier, we can conclude that SMMC is NP-complete. $\square$

The problem as stated so far is an offline version where all packet arrival times and deadlines are known beforehand and the DCU needs to solve the problem looking forward at the entire window of messages that could arrive over duration T. This problem can occur in practice when all message types and their arrival times are known deterministically, for example, when all messages are scheduled deterministically. However, in most cases the problem will be an online one with stochastic message types and arrivals where the DCU will only have access to those messages (with their arrival time and deadlines) that have reached the DCU and are waiting to be concatenated before being sent out over the backhaul. Thus, any proposed heuristics will need to perform in an *online* fashion.

3.2 Heuristics

Due to the proven hardness of the SMMC problem, in this work we develop online heuristic-based algorithms for solving the SMMC problem. Our heuristic solution approach is to rely on Earliest Deadline First (EDF) scheduling where a concatenated packet is created at the DCU starting with a message within a specific threshold of its deadline and then filled with other messages so as to maximize the packet size that can be sent out. Proposed heuristics differ in terms of what other messages they decide to fill in the concatenated packet in addition to the message whose deadline is about to expire.

Six different heuristic-based algorithms are proposed for scheduling of messages at a DCU for the SMMC problem as listed in Table 3.1. All six algorithms initiate creating a packet when one of the local message deadlines at the DCU is about to expire; they differ in terms of what other messages (in addition to the message whose deadline is about to expire) are put in the packet being sent out. In all six schemes, a *Classifier* module checks the arrived messages to see whether they are best-effort or have a specific deadline (if the selected heuristic needs to differentiate between them). Two different queues are formed based on the classification done. All deadline messages are kept in a priority queue sorted by earliest deadline. It is assumed there are two queues in the system, one for the messages with specific delay objective and another for those without a delay objective (the best effort messages). If no classification is required then all arrived messages will be sorted and placed in a single buffer. All of the proposed heuristics (except EDF-FCFS) employ the *0-1 knapsack* algorithm [46] to decide which messages to fit into the packet among the various options available. More details of the implementation of our proposed heuristics and associated pseudocode can be found in our prior work in [47].

3.3 Reference Algorithms

EDF-based Integer Linear Programming (ILP) Formulation
To get a solution for the SMMC problem one can use mathematical optimization algorithms. We have formulated the SMMC problem as a mixed-integer linear program which optimally schedules the remaining messages in addition to the EDF message to begin a packet with

Table 3.1: The proposed concatenation heuristics

Algorithm	Description
EDF-DKB	Inserts deadline messages as much as possible inside the packet and the remaining space will be filled through knapsack selection over best-effort messages that have been queued.
EDF-SDKB	Only a single deadline message sits inside the packet with any available space filled with non-deadline messages in the non-deadline queue through knapsack selection.
EDF-FCFS	Messages will be placed in the packet according to their arrival sequence from a common queue of deadline and non-deadline messages on a first-come first-served basis.
EDF-KN	Messages are chosen from a common pool of deadline and best-effort messages selected through the knapsack algorithm.
EDF-KDKB	A sequence of knapsack selections first on all queued deadline messages and then over the queued best-effort messages if needed to fill the packet.
EDF-KBKD	Reverse order of knapsack process in EDF-KDKB working first on the queued best-effort messages and then on the deadline messages if needed.

index. The problem is formulated as follows for a packet with index i:

$$\text{maximize} \quad P_i = \sum_{j=1}^{n_t} s_j x_{ij} \tag{3.1}$$

subject to constraints

$$\sum_{j=1}^{n_t} s_j x_{ij} + H \leq W,$$

$$x_{ij} \in \{0, 1\},$$

where $x_{ij} = 1$ if message j is put into packet i. In the formulation above, n_t ($n_t \leq n$) is the set of messages queued at the DCU and available for concatenation at time t ($t \leq T$). Any messages that are found to not meet deadline constraints are forwarded immediately with no concatenation process applied. This formulation is different from Equation 2.1 in that it is EDF-based and message deadlines are not a constraint as messages closest to their deadlines are selected and sent out before their deadlines occur. This formulation tries to fit in as many messages as possible (among those available) in a packet to be sent out. The given constraint specifies the maximum packet size that can be sent over the backhaul technology with a specific MTU size. The drawback of this approach in practice (as opposed to our heuristics) is the brute force nature of this integer linear programming solution procedure which makes it practically infeasible for real-time applications and those that involve large-scale data.

Theoretical Optimum

This method is theoretically the minimal number of packets that needs to go out of a DCU

for a given number of messages generated from the smart meters over a period of time. This value is not constrained by arrival times or deadlines of messages; it is computed through the equation $\left\lceil \frac{\sum_{i=1}^{n} s_i}{MTU-H} \right\rceil$ where n is the total number of arrived messages during a time interval, and s_i is the size of a message i. MTU and header size H are the parameters defined according to the backhaul technology. Although this solution is not feasible in practice, it gives a theoretical reference for the performance evaluation of any SMMC algorithms, not limited to EDF based heuristics.

4 Evaluation

4.1 Methodology

We outline below more details about the simulation environment, message arrival process, and distribution of various message types.

Simulation environment

A discrete-event simulator was developed using MATLAB to evaluate the proposed heuristic-based algorithms and compare to the reference algorithms. The network topology consisted of a group of smart meters generating messages as a poisson process and sending messages to the DCU to be routed to the control center.[1] Due to the assumption of individual meter message generation as a poisson process, we can sum the individual average message generation rates to get an cumulative average arrival rate at the DCU of λ which is used as a parameter in our simulations. We have considered three different λ values of 0.1, 0.5, and 1 at the DCU which would correspond to 90, 450, and 900 smart meters sending 1 message on average every 15 minutes. The service capacity of the DCU is considered to be infinite; however, we do study the impact of processing delays in the following section.

Message types distribution

During a day, different types of the messages may be exchanged between smart meters and the utility control center through the AMI. In our evaluations we have considered all seven basic types of messages first reported in [5]. Based on geographic location, power distribution infrastructure, and utility preferences, the transmission of messages could come from different distributions of these basic message types which will have an impact on the performance of our proposed heuristics. In our evaluations we used different Beta distributions across these message types by varying shape parameters $\alpha > 0$ and $\beta > 0$.

For our experiments we generated five different message type distribution using the shape parameters mentioned in Table 4.1 to test the performance of our proposed algorithms.

4.2 Simulation Results

Simulations were conducted with 100 runs and the mean value plotted in results shown along with 95% confidence intervals. Each scheme was evaluated in terms of the overall reduction in bytes of data transmitted out into the backhaul network by the DCU as compared to the overall incoming data in bytes from smart meters, including all headers. Each packet header was assumed to be of a fixed size of 50 bytes corresponding to the 40-60 byte range for TCP and IP headers. Figure 4.1 displays the output of our proposed algorithms and reference

[1] Prior work in [48] supports this assumption that smart meters message generation can be modeled as a poisson process.

Table 4.1: Pre-defined message arrival distributions

Distribution	Description
Uniform ($\alpha = 1$, $\beta = 1$)	The traffic would have almost equal percentage of all message types.
More smaller ($\alpha = 2.8$, $\beta = 1.9$)	Most of the arrived messages are of the smaller size of message types.
More larger ($\alpha = 0.18$, $\beta = 0.25$)	There is higher percentage of large message size and very few numbers of small size messages.
More deadline ($\alpha = 1$, $\beta = 1.8$)	Most of the times there are incoming messages with deadline restriction.
More best-effort ($\alpha = 2.5$, $\beta = 0.5$)	There are very few numbers of messages with a deadline and so many best-effort messages.

algorithms over five message types distributions with 95% confidence intervals. Results are shown for packet arrival rates at the DCU of $\lambda = 0.1, 0.5$, and 1. It can be seen that overall data volume reduction varies from 5-25% depending on message type distribution, message arrival rate at DCU, and specific algorithm used. Three questions answered are:

How do the proposed heuristics stack up against each other and reference algorithms?

Taking a look at the bar charts in Figure 4.1 one can observe that the algorithm EDF-KN has the best performance among all other heuristics and comes very close to the performance of the EDF-based ILP across all λ and message type distributions. This is due to the fact that EDF-KN is using a common pool of messages whether they be deadline or best effort, giving more options to maximize packet size before it is sent out. Since typically there are enough queued messages before a deadline reaches, the algorithm has a good collection of options to maximize the packet before sending it out. It can be noticed that in general EDF-based approaches do well compared to theoretical volume reduction, where the latter increases with MTU size and decreased with the size of H.

What is the impact of message type distribution?

The uniform distribution of all message types serves as the reference case to compare other distributions. For the more deadline case with a majority of all messages having deadlines, overall data volume reduction is smaller for all algorithms. Presence of more messages with deadlines than best-effort necessitates packets to be sent out of the DCU without having the luxury of waiting for the right combination to maximize packet size. However, when there are more best-effort messages present, algorithms can wait longer before being forced to send out packets; this allows each packet to be larger, and hence reduces packet overheads. The case for more smaller size messages is similar to the more deadline message case in that it helps reduce packet overheads significantly through concatenation as header sizes are comparable to data sizes. Smaller messages are also easier to pack into a packet. Conversely, the more larger messages case results in greater difficulty to fill messages into a packet; also larger underlying message sizes already have a reduced overhead making much improvements

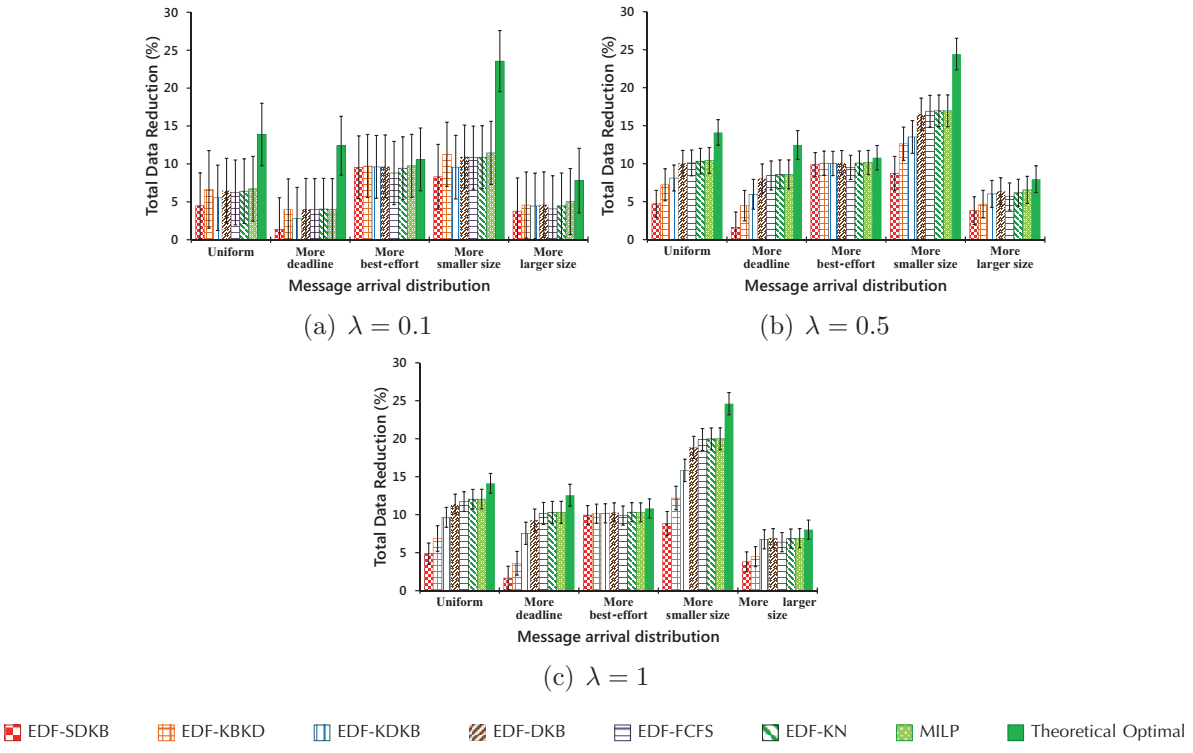


Figure 4.1: Overall data reduction percentage using proposed heuristics over different message arrival rate and message type distributions.

through concatenation difficult.

What is the impact of λ?

The value of λ signifies the packet arrival rate at the DCU; hence, larger values indicate that more messages are arriving at the DCU increasing opportunities for a concatenation algorithm to find a best fit of messages in an outgoing packet from the DCU to reduce overall protocol overhead. The EDF-KN data volume reduction approaches very close to that of even the theoretically optimal solution with increasing λ. Thus, greater the rate of packet arrivals, the proposed EDF-based concatenation algorithm over a common queue of messages maximizes the reduction in data volume.

5 Impact of Network and Processing Delays

Network delays between the DCU and the utility control center, and processing delay at the DCU itself are two factors we had assumed to be negligible in the results presented so far. The magnitude of these delays may not be negligible in all practical cases, and can cut down the amount of time a DCU can wait to maximize the size of outgoing packets sent out. Thus, there will be a direct correlation between network and processing delays on the ability of a DCU to reduce protocol overhead. An interesting challenge here is that the DCU cannot accurately predict these delays beforehand; each concatenated packet will suffer variable network and processing delays due to many factors related to number of messages processed and characteristics of the communication backhaul. Thus, the DCU needs to rely on an estimate of network and processing delays it needs to budget into computing the local deadline of each message. An overestimate will reduce the amount of time a DCU will have to wait and concatenate a large packet; an underestimate on the other hand can mean some messages will miss their deadlines. This section describes how such delays can be estimated and what what impact it will have on data volume reduction through message concatenation.

5.1 Estimation of Network and Processing Delays

To estimate the processing delay, we need to break it into the major individual components that cause delay. These components are (i) concatenation delay: the time required to put all selected messages into a packet and add a common header, (ii) knapsack delay: the time required by some of the schemes that use a knapsack operation to select messages from a queue of messages, and (iii) sorting delay: the time required to maintain the queue sorted in terms of earlier deadlines. These components are present in each heuristic in possibly different ways based on the nature of the algorithm. Table 5.1 summarizes how each of these components (C_C, C_S and C_K time costs for concatenation, sorting and selection through Knapsack respectively) sum up to the total processing delay for each heuristic scheme. These schemes operate on either a single common queue of n items, or one of two queues (with sizes n_1 and n_2) having deadline and non-deadline messages, or both queues one after the other. The next step was to populate realistic values into the processing delay estimation model. For this, we measured actual processing delays when executing each of the three operations: concatenation, knapsack selection, and keeping a sorted queue. These values were computed on a Dell Optiplex 64-bit PC with a 2-core 2.8GHz CPU and 5GB RAM for a full range of values of n from 1 to 1000 to study all possible queue sizes we are likely to encounter for message arrival rates used in the evaluations in Section 4.[1] By populating these values for a

[1]We assume that when our algorithms are deployed, an estimate can be re-calculated for the specific system employed in the DCU as opposed to using the estimates discussed here. DCUs on the market

Table 5.1: The heuristics processing time calculations

Heuristic	Processing Delay
EDF-FCFS	$C_C + C_S$
EDF-KN	$C_C + C_S + C_K(n)$
EDF-DKB	$C_C + C_S + C_K(n_1)$
EDF-SKB	$C_C + C_S + C_K(n_2)$
EDF-KDKB	$C_C + C_S + C_K(n_1) + C_K(n_2)$
EDF-KBKD	$C_C + C_S + C_K(n_2) + C_K(n_1)$

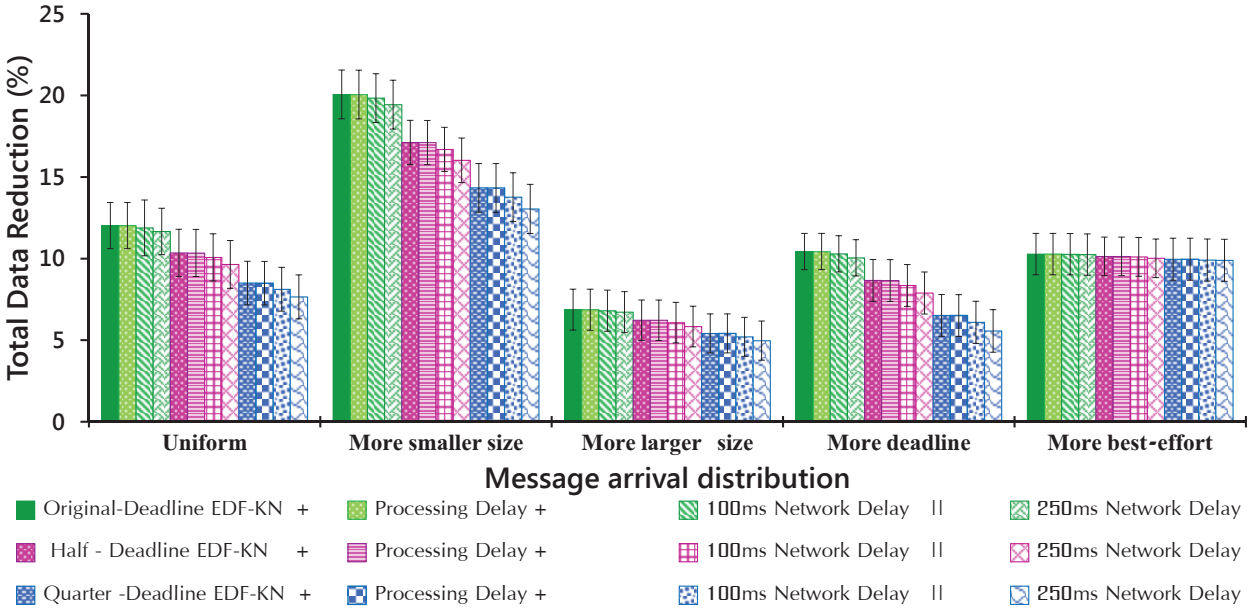


Figure 5.1: Data reduction trend vs. Delay addition.

given n in the processing delay model presented in Table 5.1, the DCU could easily construct an estimate.[2]

5.2 Evaluation Results

Here we re-evaluate our proposed heuristic-based algorithms with varying values of network and processing delays, and study the impact on achievable reductions in protocol overhead. For these evaluations we have chosen the EDF-KN heuristic, one of the better performing heuristics among those evaluated in section 4.2. Figure 5.1 presents the results for $\lambda = 1$ and shows the protocol overhead reduction achieved with varying values of network and processing delays, including the case where such delays are set to nil. In addition, to further

can have high processing capabilities as described in [49] and we expect the values used in this work to over-estimate actual processing delays.

[2]We refer the reader to [50] for a description of how network delays can be predicted with an exponentially weighted moving average over a sliding window of previously seen delays. We will study the impact of various possible network delays to assess the impact on benefits of message concatenation in evaluations that follow.

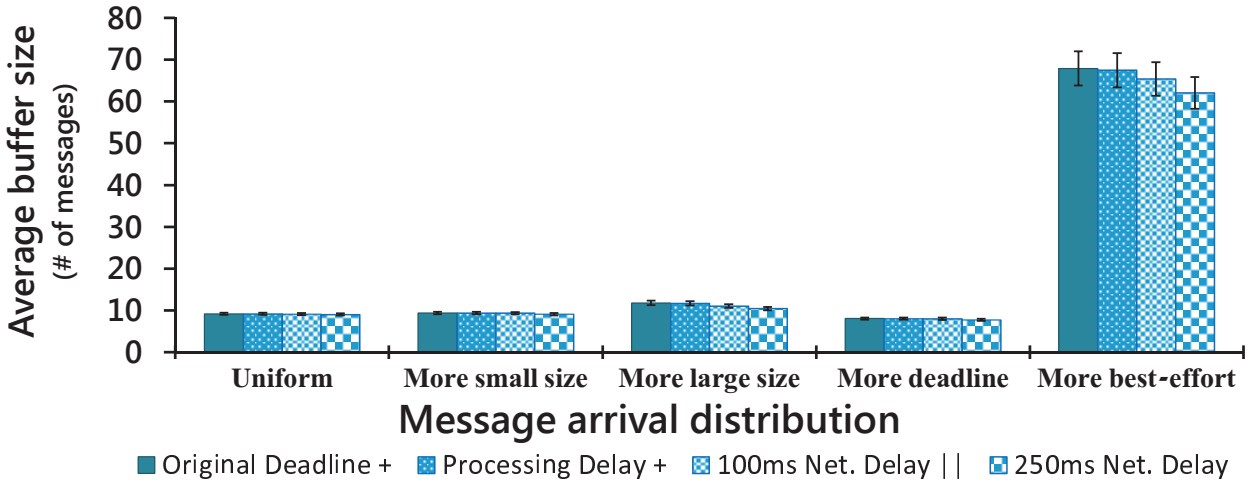


Figure 5.2: Average buffer size vs. Delay addition

explore the lower limits of possible benefits of message concatenation, we experiment with deadline values half and quarter the amount of that used in our evaluations in Section 4. The impact of tighter deadlines will be similar to that of additional network and processing delays, with both factors essentially reducing the time the DCU has to concatenate messages into larger packets.

The results in Figure 5.1 show that as processing and network delays increase, the percentage overhead reduction decreases. Similarly, as deadlines get tighter, the data volume reduction achievable reduces. Even for such extreme cases considered, there is at least a 5% reduction in data volume possible. The biggest impact of network and processing delays, or tighter deadlines is with the "more deadline" message distribution with a greater fraction of messages needing to be concatenated and sent out quickly. The smallest impact of delays or tighter deadlines is seen for the "more best-effort" case where most messages are not hard-pressed to meet deadlines.

A more accurate depiction of what happens inside the DCU can be seen by studying the average queue or buffer size for various message type distributions for estimated processing delays and varying network delay values of 100 ms and 250 ms. A similar trend can be expected for tighter deadline values. As Figure 5.2 confirms, the "more deadline" message distribution has the smallest average queue size, implying that messages do not stay in the buffer for long periods. The "more best-effort" message distribution at the other extreme results in the largest average queue size implying messages stay in the buffer for a much longer duration. A large average queue size does add additional processing delay; however, for the more best-effort case, there are few messages with deadlines that are impacted by the larger processing delays. For all the other schemes, evident from the results, the average queue size stays small enough to not adversely impact data volume reduction.

6 Data Volume with Lossy Links

Another practical aspect that needs to be considered is the impact of lossy backhaul links on the large concatenated packets expected to be sent out from the DCU by proposed heuristic-based algorithms. Larger packets will typically suffer more re-transmissions (and thus adding to data volume transported) when sent through networks with a fixed bit-error rate (BER) due to their larger size. Thus, it is imperative to explore the impact of various backhaul technologies, each with different BER characteristics, on benefits of message concatenation[1].

6.1 Theory

The most important factor in analyzing the impact of lossy networks is considering the BER of the technology being used. The transmission BER is the number of detected bits that are incorrect before error correction, divided by the total number of transferred bits (including redundant error codes). Different communication technologies have different BER. The goal here is to translate a given BER for a technology and estimate the corresponding data volume reduction ratio. Let e_b be the BER of a given technology. A packet is declared incorrect if at least one bit is erroneous. Thus, for a packet of size L bits, the resulting packet error rate (PER) of the technology, e_p, then is $e_p = 1 - (1 - e_b)^L$. Let D be the volume of data in bytes (including payload and control overhead) that would have been sent over the backhaul in a time period T when message concatenation is not employed. Let $D^{'}$ be the volume of data sent (again including payload and control overhead) over the backhaul after message concatenation. With a PER of e_p and $e_p^{'}$ respectively, the corresponding data volume sent through the backhaul will be $(1 + e_p)D$ and $(1 + e_p^{'})D^{'}$ respectively. Thus, the data volume reduction ratio ρ with a lossy backhaul can be computed as

$$\rho = \frac{(1 + e_p)D - (1 + e_p^{'})D^{'}}{(1 + e_p)D} \tag{6.1}$$

With larger packet sizes $e_p^{'} > e_p$, thus reducing the data volume reduction ratio as compared to the case when lossiness of the backhaul network is ignored.

6.2 Numerical Evaluation

The technologies for the backhaul considered are fiber optic, WiMAX, 3G Cellular; these three technologies are currently commonly used to connect the AMI at the customer to the

[1]Due to space restrictions, we do not explore the analogous issue of packet loss due to network congestion; the eventual impact on the benefits of message concatenation is expected to be similar regardless of the underlying reason for packet loss

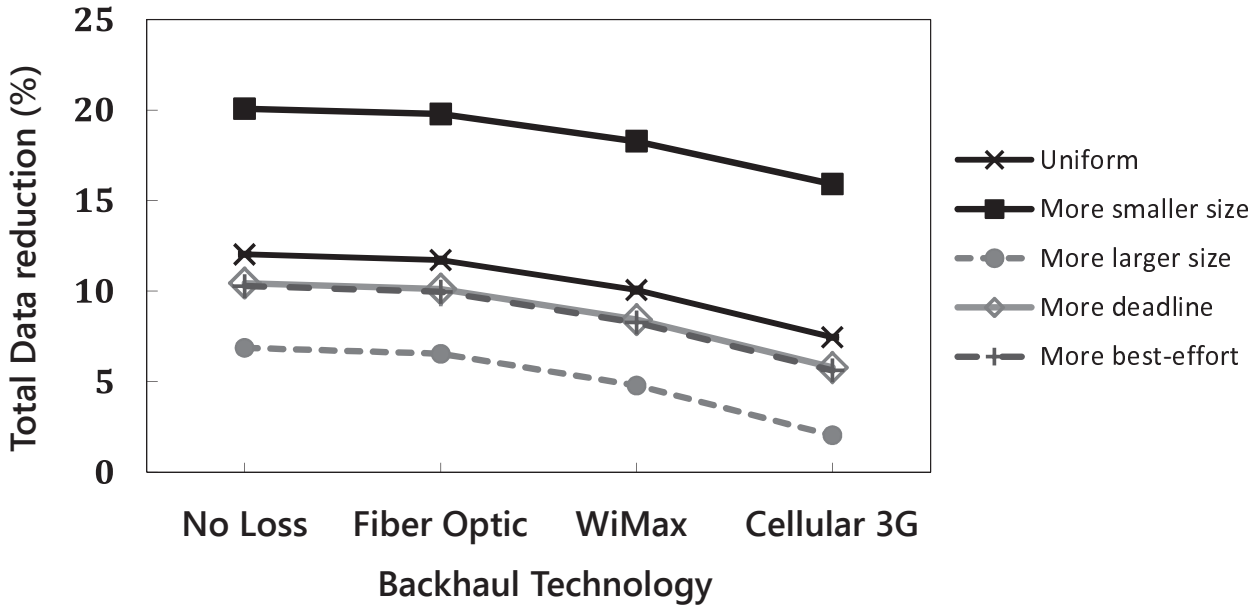


Figure 6.1: Data reduction savings vs. Different backhaul technologies

backbone network and tend to be lossier than the core network. We picked BER values for these technologies based on known ranges in [51, 52, 53] to study the impact of message concatenation algorithms. The BER values e_b used in the following evaluation were 5E-07, 3.16E-06, and 7.5E-06 for fiber optic, WiMAX, and 3G technologies respectively. For each technology, we computed PER's e_p using the equation above. For the case with no message concatenation, we considered an average packet size of 100 bytes ($L = 800$ bits) in computing a PER of e_p; for the case with concatenation, we used a packet size of 1000 bytes ($L = 8000$ bits) to compute $e_p^{'}$ which is roughly the average size of concatenated packet seen in our simulations from the earlier sections. Finally, using Equation 6.1, we computed ρ for each of the three technologies with D and $D^{'}$ computed based on our simulations earlier in Section 4 for the EDF-KN scheme with a message arrival rate of $\lambda = 1$.

It can be seen from Figure 6.1 that for even the most lossy technology considered (3G) with worst-case BER characteristics chosen, data volume reduction with message concatenation only falls by 3-4% compared to the reference ideal BER case. Thus, the benefits of message concatenation seems to hold up for the most commonly used technologies. These results are likely to be better with the use of forward error correction (FEC) techniques employed to minimize packet loss.

7 A Case Study of Practical Benefits of Proposed Algorithms

With many hundreds of thousands of customers in a geographic location, utilities will be thus sending data in the order of Mbps to Gbps through their backhaul networks connecting to control centers. This scale of data flow through AMI networks is also supported by the literature (e.g. [54, 55, 56]). This section presents a case study of actual data rates flowing through neighborhood networks of different sizes and how it may impact a given backhaul communication technology and the applicability of proposed data concentration algorithms.

Assume a power system topology with a feeder connecting to 1350 customers in an area with 450 distribution transformers, with one transformer connecting to 3 customer smart meters. This chosen topology is typical of for a suburban area in the U.S (see for e.g. [57]). A logical communications network overlaid on the physical topology of this distribution system topology could be as follows. Based on the manner in which the communications network is organized, its communication range, and the customer meter density, x smart meters could be connected to a DCU. For the topology assumed, x could take on any values from 1 to 1350. The total number of DCUs required would depend on the value of x. The DCUs are then further connected through a backhaul to the communications network. With x meters each sending a message every y seconds, the average data arrival rate at each DCU will be $\frac{x}{y}$ messages per second. For message sizes averaging 350 bytes and a 50 byte header (typical sizes from [5]), this amounts to a data rate of $\frac{3.2x}{y}$ Kbps at each DCU employed. For $x = 450, 900, 1500$ and for $y = 900$ seconds (15 minute intervals), this amounts to data rates per DCU of 1.6 Kbps, 3.2 Kbps, and 4.8 Kbps. For more fine-grained data collections in the future for analysis (e.g. as motivated in [58]) or just applications such as EV load control and appliance-level load monitoring, y could be of the order of few seconds. For 10 second intervals, this results in data rates of 144 Kbps, 288 Kbps, and 432 Kbps for $x = 450$, 900, and 1350 respectively.

A technology like power line communications can only support data rates in the order of Kbps [59]. Thus for neighborhood deployments of the order of 500-1500 smart meters, with a low bandwidth technology like PLC, it is imperative that data volume through such backhaul links be managed carefully. Other higher bandwidth backhaul links such as cellular, Wi-Fi, WiMax can support higher data rates (at higher costs) and will be less stressed by smart meter deployments. With electric utilities either leasing communications capacity from telecom companies, or building their own telecommunications networks and then leasing capacity to recuperate costs, they will benefit from reducing the amount of data sent through their networks regardless of the scale of a smart meter deployment and bandwidth of communication links. A 20% reduction in data volume (as can be achieved by the proposed heuristics in this work) should translate to a similar reduction in network infrastructure costs under a scenario of per byte capacity costs. Such reduction in costs is expected to also benefit

all customers, whether they are equipped with smart meters or not. As the penetration of smart meters increases, the applicability of this work will keep increasing with more benefits for greater traffic volumes as found in our results earlier in this report. The FERC survey in 2012 [60] indicated AMI penetration to be about 23% (a 14% increase over 2010 levels) and is expected to have increased at a similar rate since then.

A scenario where the applicability of the proposed data concentration approach would be reduced is if network capacity is not metered per byte of data transported, but instead is a fixed capacity cost, and if the smart meter deployments are small enough to not stress deployed networks. The data flow analysis in the previous paragraph shows that in such a case, for neighborhoods as small as 500-1500 smart meters connected to a single DCU, the proposed data concentration schemes may be useful only if a low bandwidth technology like power line communications is used for the backhaul. However, if multiple such neighborhoods are clustered together behind a single data concentrator (with appropriate network topology configurations), the data concentration schemes would still be useful for even high bandwidth technologies. For larger number of meters, such as 3000 and above, data generated at (10 second interval collection) would be of the order of Mbps and can stress higher bandwidth links and be very useful even for those links.

8 Conclusion and Future Work

This part of the report demonstrated that message concatenation algorithms can be an important element of data concentrators deployed in smart grids to solve the looming challenge of transporting massive data volumes through last mile bandwidth-constrained backhaul networks. Effective message concatenation algorithms at DCUs (such as the EDF-KN algorithm proposed in this work) were shown to be able to reduce overall data volume by 10-25% for each DCU. This reduction was achieved just by a reduction in protocol overhead with no compression of the original data sent by smart meters; this provides enough motivation to develop additional data concentration mechanisms at DCUs that also act on the payload of messages. Another direction of future work is to look at how concatenation can be done at multiple levels of the communications network, not limited to just the first hop from smart meters.

Some preliminary related work and ongoing work to that presented in this part of the report can be found in [61, 62, 63, 64].

Bibliography

[1] National Energy Technology Laboratory, "Advanced Metering Infrastructure," February 2008. [Online]. Available: http://www.netl.doe.gov/smartgrid/refshelf.html, lastaccessedOctober27,2011.

[2] US Department of Energy, "What the Smart Grid Means to Americans," 2008, available online at http://www.doe.gov/sites/prod/files/oeprod/ DocumentsandMedia/ConsumerAdvocates.pdf.

[3] D. Bernaudo et al., "SmartGrid/AEIC AMI Interoperability Standard Guidelines for ANSI C12.19 / IEEE 1377 / MC12.19 End Device Communications and Supporting Enterprise Devices, Networks and Related Accessories," The Association of Edison Illuminating Companies, Meter and Service Technical Committee report ver. 2, November 2010.

[4] E. E. Queen, "A Discussion of Smart Meters And RF Exposure Issues," Edison Electric Institute (EEI), Washington, D.C, A Joint Project of the EEI and AEIC Meter Committees, March 2011.

[5] W. Luan, D. Sharp, and S. Lancashire, "Smart grid communication network capacity planning for power utilities," in *Transmission and Distribution Conference and Exposition, 2010 IEEE PES*, april 2010, pp. 1 –4.

[6] M. Allalouf, G. Gershinsky, L. Lewin-Eytan, and J. Naor, "Data-quality-aware volume reduction in smart grid networks," in *Smart Grid Communications (SmartGridComm), 2011 IEEE International Conference on*, 2011, pp. 120–125.

[7] Engage Consulting Limited, "High-level smart meter data traffic analysis," for the Energy Networks Association(ENA), UK, Document Ref. No. ENA-CR008-001-1.4, may 2010. [Online]. Available: http://www.energynetworks.org/electricity/futures/ smart-meters.html,lastaccessedonAugust8,2013.

[8] "Arcadian's Smart Grid: Licensed Spectrum Network to Own or Rent," greentechgrid. [Online]. Available: http://www.greentechmedia.com/articles/read/ arcadians-utility-offering-licensed-spectrum-to-own-or-rent

[9] M. Kennedy, "Leveraging investment in fiber optic communications," IEEE Smart Grid. [Online]. Available: http://smartgrid.ieee.org/june-2011/ 105-leveraging-investment-in-fiber-optic-communications

[10] PRIME Alliance, "PoweRline Intelligent Metering Evolution." [Online]. Available: http://www.prime-alliance.org,lastaccessedonAugust8,2013.

[11] A. F. Snyder and M. T. G. Stuber, "The ANSI C12 protocol suite-updated and now with network capabilities," in *Power Systems Conference: Advanced Metering, Protection, Control, Communication, and Distributed Resources*, March 2007, pp. 117–122.

[12] R. Rajagopalan and P. Varshney, "Data-aggregation techniques in sensor networks: a survey," *Communications Surveys Tutorials, IEEE*, vol. 8, no. 4, pp. 48–63, quarter 2006.

[13] L. Becchetti, A. Marchetti-Spaccamela, A. Vitaletti, P. Korteweg, M. Skutella, and L. Stougie, "Latency-constrained aggregation in sensor networks," *ACM Trans. Algorithms*, vol. 6, no. 1, pp. 13:1–13:20, Dec. 2009.

[14] Y. Yu, B. Krishnamachari, and V. K. Prasanna, "Energy-latency tradeoffs for data gathering in wireless sensor networks," 2004.

[15] S. Habib and P. Marimuthu, "Data aggregation at the gateways through sensors' tasks scheduling in wireless sensor networks," *Wireless Sensor Systems, IET*, vol. 1, no. 3, pp. 171 –178, september 2011.

[16] S. Zhu, W. Wang, and C. V. Ravishankar, "Pert: A new power-fficient real-time packet delivery scheme for sensor networks," *Int. J. Sen. Netw.*, vol. 3, no. 4, pp. 237–251, Jun. 2008.

[17] S. Hariharan and N. Shroff, "Maximizing aggregated revenue in sensor networks under deadline constraints," in *IEEE Decision and Control (CDC)*, 2009, pp. 4846–4851.

[18] S. Hariharan and N. B. Shroff, "Deadline constrained scheduling for data aggregation in unreliable sensor networks," in *WiOpt*, 2011, pp. 140–147.

[19] Y. Chen, J. K. Hwang, and S. M. Wu, "A reliable power line carrier and wireless data concentrator for broadband energy information network," *Consumer Electronics, IEEE Transactions on*, vol. 49, no. 4, pp. 1054 – 1060, nov. 2003.

[20] J. Schroeder, E. Doherty, and M. Nager, "Overvoltage protection of data concentrators used in smart grid applications," in *Innovative Smart Grid Technologies (ISGT), 2011 IEEE PES*, jan. 2011, pp. 1 –4.

[21] K. Zhu, A. T. Al-Hammouri, and L. Nordstrom, "To concentrate or not to concentrate: Performance analysis of ict system with data concentrations for wide-area monitoring and control systems," in *Power and Energy Society General Meeting, 2012 IEEE*, 2012, pp. 1–7.

[22] M. Adamiak, M. Kanabar, J. Rodriquez, and M. Zadeh, "Design and implementation of a synchrophasor data concentrator," in *Innovative Smart Grid Technologies - Middle East (ISGT Middle East), 2011 IEEE PES Conference on*, dec. 2011, pp. 1 –5.

[23] R. M. Karp, "Reducibility among combinatorial problems," in *Complexity of Computer Computations*, R. E. Miller and J. W. Thatcher, Eds. Plenum Press, 1972, pp. 85–103.

[24] C. L. Liu and J. W. Layland, "Scheduling algorithms for multiprogramming in a hard-real-time environment," *J. ACM*, vol. 20, no. 1, pp. 46–61, Jan. 1973.

[25] A. K. Mok, "Fundamental design problems of distributed systems for the hard-real-time environment," Cambridge, MA, USA, Tech. Rep., 1983.

[26] C. D. Locke, "Best-effort decision making for real-time scheduling," Ph.D. dissertation, 1986. [Online]. Available: http://opac.inria.fr/record=b1052010

[27] S. Chen, T. He, H. Y. S. Wong, K.-W. Lee, and L. Tong, "Secondary job scheduling in the cloud with deadlines," in *IPDPS Workshops*, 2011, pp. 1009–1016.

[28] S. Chen and L. Tong, "iEMS for large scale charging of electric vehicles: Architecture and optimal online scheduling," in *Smart Grid Communications (SmartGridComm), 2012 IEEE Third International Conference on*, 2012, pp. 629–634.

[29] H. Sanders, "Enabling price responsive demand," Discussion Paper, January 2012. [Online]. Available: http://www.ngusummitna.com/media/whitepapers/2012/CaliforniaISO_EnablingPriceResponsiveDemand.pdf

[30] J. Medina, N. Muller, and I. Roytelman, "Demand response and distribution grid operations: Opportunities and challenges," *Smart Grid, IEEE Transactions on*, vol. 1, no. 2, pp. 193 –198, sept. 2010.

[31] M. Braun and P. Strauss, "A review on aggregation approaches of controllable distributed energy units in electrical power systems," *International Journal of Distributed Energy Resources*, vol. 4, no. 4, pp. 297–319, 2008.

[32] Z. Fan, P. Kulkarni, S. Gormus, C. Efthymiou, G. Kalogridis, M. Sooriyabandara, Z. Zhu, S. Lambotharan, and W. Chin, "Smart grid communications: Overview of research challenges, solutions, and standardization activities," *Communications Surveys Tutorials, IEEE*, vol. PP, no. 99, pp. 1 –18, 2012.

[33] C.-H. Lo and N. Ansari, "Decentralized controls and communications for autonomous distribution networks in smart grid," *Smart Grid, IEEE Transactions on*, pp. 1–12, 2012.

[34] R. A. Johnston, C. H. Hauser, K. H. Gjermundrød, and D. E. Bakken, "Distributing time-synchronous phasor measurement data using the gridstat communication infrastructure," in *HICSS*, 2006.

[35] D. E. Bakken, A. Bose, C. H. Hauser, D. E. Whitehead, and G. C. Zweigle, "Smart generation and transmission with coherent, real-time data," *Proceedings of the IEEE*, vol. 99, no. 6, pp. 928–951, 2011.

[36] C. H. Hauser, D. E. Bakken, I. Dionysiou, K. H. Gjermundrød, V. S. Irava, J. Helkey, and A. Bose, "Security, trust, and QoS in next-generation control and communication for large power systems," *IJCIS*, vol. 4, no. 1/2, pp. 3–16, 2008.

[37] P. Myrda and K. Koellner, "NASPInet - the internet for synchrophasors," in *System Sciences (HICSS), 2010 43rd Hawaii International Conference on*, 2010, pp. 1–6.

[38] I. Rouf, H. Mustafa, M. Xu, W. Xu, R. Miller, and M. Gruteser, "Neighborhood watch: security and privacy analysis of automatic meter reading systems," in *Proceedings of the 2012 ACM conference on Computer and communications security*, ser. CCS '12, 2012, pp. 462–473.

[39] A. Barenghi and G. Pelosi, "Security and privacy in smart grid infrastructures," *2012 23rd International Workshop on Database and Expert Systems Applications*, pp. 102–108, 2011.

[40] Z. Erkin, J. R. Troncoso-Pastoriza, R. L. Lagendijk, and F. Pérez-González, "Privacy-preserving data aggregation in smart metering systems: An overview," *IEEE Signal Process. Mag.*, vol. 30, no. 2, pp. 75–86, 2013.

[41] M. Line, I. Tondel, and M. Jaatun, "Cyber security challenges in smart grids," in *Innovative Smart Grid Technologies (ISGT Europe), 2011 2nd IEEE PES International Conference and Exhibition on*, 2011, pp. 1–8.

[42] H. S. Fhom and K. M. Bayarou, "Towards a holistic privacy engineering approach for smart grid systems," in *Proceedings of the 2011IEEE 10th International Conference on Trust, Security and Privacy in Computing and Communications*, ser. TRUSTCOM '11, 2011, pp. 234–241.

[43] X. He, M.-O. Pun, and C.-C. J. Kuo, "Secure and efficient cryptosystem for smart grid using homomorphic encryption," in *Proceedings of the 2012 IEEE PES Innovative Smart Grid Technologies*, ser. ISGT '12. Washington, DC, USA: IEEE Computer Society, 2012, pp. 1–8. [Online]. Available: http://dx.doi.org/10.1109/ISGT.2012.6175676

[44] F. Cohen, "The smarter grid," *IEEE Security & Privacy*, vol. 8, no. 1, pp. 60–63, 2010.

[45] R. L. Lagendijk, Z. Erkin, and M. Barni, "Encrypted signal processing for privacy protection: Conveying the utility of homomorphic encryption and multiparty computation," *IEEE Signal Process. Mag.*, vol. 30, no. 1, pp. 82–105, 2013.

[46] T. H. Cormen, C. Stein, R. L. Rivest, and C. E. Leiserson, *Introduction to Algorithms*, 2nd ed. McGraw-Hill Higher Education, 2001.

[47] B. Karimi, V. Namboodiri, and M. Jadliwala, "On the scalable collection of metering data in smart grids through message concatenation," in *Smart Grid Communications (SmartGridComm), 2013 Fourth IEEE International Conference on*, October 2013.

[48] H. Li, Z. Han, L. Lai, R. Qiu, and D. Yang, "Efficient and reliable multiple access for advanced metering in future smart grid," in *IEEE Smart Grid Communications (SmartGridComm)*, 2011, pp. 440–444.

[49] S. Evanczuk, "Data Concentrators Combine AFEs, MCUs, and Radios to Key Smart Grid Efficiency," March 2013. [Online]. Available: http://www.digikey.com/us/en/techzone/energy-harvesting/resources/articles/data-concentrators-combine-afes-mcus-radios.html

[50] V. Namboodiri and L. Gao, "Energy-efficient VoIP over Wireless LANs," *Mobile Computing, IEEE Transactions on*, vol. 9, no. 4, pp. 566 –581, april 2010.

[51] O. V. Sinkin, "Calculation of bit error rates in optical fiber communications systems in the presence of nonlinear distortion and noise," Dissertation, University of Maryland, 2006, submiited in partial fullfilment of the requirements for the degree of Doctor of Philosophy.

[52] H. Kaur and M. L. Singh, "Bit error rate evaluation of ieee 802.16 (wimax) in ofdm system," *International Journal of Computer Applications*, vol. 40, no. 12, pp. 10–13, February 2012, published by Foundation of Computer Science, New York, USA.

[53] P. G. P. B. Sebastien Tomas, Mattias Ahnoff, "Tms320c6416 coprocessors and bit error rates," Texas Instruments, Application Report SPRA974, November 2003.

[54] D. Alahakoon and X. Yu, "Advanced analytics for harnessing the power of smart meter big data," in *Intelligent Energy Systems (IWIES), 2013 IEEE International Workshop on*, Nov 2013, pp. 40–45.

[55] M. Ringwelski, C. Renner, A. Reinhardt, A. Weigel, and V. Turau, "The hitchhiker's guide to choosing the compression algorithm for your smart meter data," in *Energy Conference and Exhibition (ENERGYCON), 2012 IEEE International*, Sept 2012, pp. 935–940.

[56] W. Luan, D. Sharp, and S. LaRoy, "Data traffic analysis of utility smart metering network," in *Power and Energy Society General Meeting (PES), 2013 IEEE*, July 2013, pp. 1–4.

[57] M. Oens and C. Lange, "Improvements in feeder protectionproviding a primary and backup relay system utilizing one relay per feeder," Schweitzer Engineering Laboratories, Inc., 2013. [Online]. Available: https://www.selinc.com/WorkArea/DownloadAsset.aspx?id=3417

[58] B. McCracken, M. Crosby, C. Holcomb, S. Russo, and C. Smithson, "Data-driven insights from the nations deepest ever research on customer energy use," Pecan Research Institute, 2013. [Online]. Available: http://www.nhpci.org/publications/NHPC_White-paper-Making-Sense-of-Smart-Home-final_20140425.pdf

[59] M. Hoch, "Comparison of plc g3 and prime," in *Power Line Communications and Its Applications (ISPLC), 2011 IEEE International Symposium on*, April 2011, pp. 165–169.

[60] Federal Energy Regulatory Commission (FERC), "Assessment of demand response and advanced metering," Staff Report, December 2012. [Online]. Available: http://www.ferc.gov/legal/staff-reports/12-20-12-demand-response.pdf,lastaccessedonSeptember1,2014.

[61] B. Karimi, V. Namboodiri, and M. Jadliwala, "Scalable collection of metering data in smart grids through message concatenation," *IEEE Transactions on Smart Grids*, vol. 6, pp. 1697–1706, 2015.

[62] B. Karimi and V. Namboodiri, "On the capacity of a wireless backhaul for the distribution level of the smart grid," *IEEE Systems Journal Special Issue on Smart Grid Communications Systems*, vol. 8, no. 2, pp. 521–532, 2014.

[63] V. Namboodiri, V. Aravinthan, B. Karimi, and W. Jewell, "Towards a secure, wireless home-area network for metering in smart grids," *IEEE Systems Journal Special Issue on Smart Grid Communications Systems*, vol. 8, no. 2, pp. 509–520, 2014.

[64] V. C. Dev, U. Das, V. Namboodiri, S. Chakraborty, V. Aravinthan, Y. Guo, and A. Srivastava, "Towards application-aware data concentration schemes for advanced metering infrastructures," in *2015 IEEE Conference on Smart Grid Communications*, November 2015.

Part II

Impacts of Communication and Control on Distribution System

Visvakumar Aravinthan

Muhammad Usman Khan and Abbas Gholizadeh, MS Students
Suvagata Chakraborty, PhD Student

Wichita State University

For information about Part II, contact:

Visvakumar Aravinthan
Assistant Professor
Electrical Engineering and Computer Science
Wichita State University
Email: visvakumar.aravinthan@wichita.edu
Phone: 316-978-6324

Power Systems Engineering Research Center

The Power Systems Engineering Research Center (PSERC) is a multi-university Center conducting research on challenges facing the electric power industry and educating the next generation of power engineers. More information about PSERC can be found at the Center's website: http://www.pserc.org.

For additional information, contact:

Power Systems Engineering Research Center
Arizona State University
527 Engineering Research Center
Tempe, Arizona 85287-5706
Phone: 480-965-1643
Fax: 480-965-0745

Notice Concerning Copyright Material

Table of Contents

List of Tables

List of Figures

1 Introduction

1.1 Background

One of the critical components of distribution system advancement is the communication infrastructure. Real-time management of distribution system requires improved two-way communication. Due to the types of equipment at distribution level, such as voltage regulators, capacitor banks, automated switches etc. the requirements distribution level is different, when compared to the transmission level requirements. Any distribution level communication infrastructure should incorporate the following to ensure the effectiveness.

- *Consumer level benefits:* One of the prime modes of real time control of distribution level is residential consumer level demand response. The two-way communication infrastructure will determine control frequency. This will have an effect on the expected and actual benefits to the consumer.
- *System level benefits:* Communication infrastructure will determine the level of aggregation required. When consumer data is aggregated the demand interval will be modified. Impact of demand interval on accuracy predicting the system benefits, is a new concern. An approach to estimate the error needs to be developed.

This work focuses on developing an evaluation approach for consumer and system level impacts. The following two approaches were taken in this work.

1.2 Demand Response Impacts

Demand side management (DSM) among other smart grid initiatives, has gained increased attention in an attempt to defer the investment in upgrading the power grid in terms of more generating units and transmission lines [1]. The technologies utilized in smart grid pilot projects include advanced metering infrastructure (AMI), automated meter reading (AMR), distributed generation, energy storage, smart appliances and dynamic pricing. A well-developed Demand Side Management (DSM) program is expected to have impact on several applications such as peak reduction, power factor improvement, equipment life, and consumer cost reduction.

Even though significant number of work has been done to develop demand management schemes, this work differs from the literature, as it compares the impact of control interval (demand interval) on the actual benefits to the distribution grid. The control parameters are based on different control sampling rate selection and represent the impact on both the user and the utility. The motivation is to develop a framework to analyze the tradeoffs when choosing different sampling rates; based on the ASHRAE 55 [2] standard for thermal comfort.

Two key demand response resources are (*i*) thermostatically controlled loads (TCLs) and (*ii*) electrochemical batteries used in plug-in electric vehicles. This is due to the fact that both of them are capable of storing energy in some form and release when it is needed with minimum interaction with the consumers and their comfort level. Scheduling either of them is that they are not only good at reducing peak demand, but also are good candidates for load shifting or valley filling. Among other benefits, HVAC systems are able to store energy i.e. cooling in off peak hours and coasting during peak hours. Moreover, the user comfort level violations can be reduced by raising or dropping the thermostat to levels that does not cause much thermal discomfort as long as it's within the limits defined by ASHRAE 55 standard for thermal comfort.

On the other hand, other controllable appliances may impose some time delay in operation, thus create more discomfort. For example, a clothes dryer cannot be used before clothes washer, so the consumer will have to wait to operate the dryer once the washer cycle is complete. Also, the availability of programmable thermostats may help avoiding the investments in developing advanced equipment as is the case with other appliances that may require special circuits to be able to communicate with the network and control those complicated appliances like dish washer, washer-dryer etc.

1.3 Feeder Level Impacts

Current advancements in the power system have focused on developing solutions and technical frameworks for the next generation distribution system in order to make it more flexible, robust and cost effective [1] and to allow more participation of residential customers. Advanced metering infrastructure and distributed intelligent devices allows better monitoring and controlling but data storage is a concern using smart metering. Data storage requirements along with bandwidth limitation of communication network leads to use of different demand interval using information engineering concepts [1]. Aggregating data from high resolution to low resolution and time period over which aggregation is done is known as aggregation interval. Aggregating demand data from the consumer end at a certain sampling interval could be linked to demand interval [3].

Depending on the type of signal to be measured the data aggregation interval is of great importance. Important details are lost due to averaging if long aggregation interval is used and on the other hand short interval leads to copious amounts of data that is difficult to assess. This excessive information may not be meaningful and leads to storage problem if the data is to be retained [4]. Thus the accuracy of the information obtained at different intervals is analyzed in this paper. Nine different aggregation intervals are used for the analysis.

An algorithm to determine the importance of aggregation interval needs to be evaluated using an appropriate statistical tool. This must guarantee to improve the consumer side and system side benefits. For example: in case of voltage transformer with tap changer as the main tool [4]. If the tap of transformer and capacitor regulators were frequently changed, these devices would be easily damaged [2]. This is due to the repeated switching operation of the tap changer which results in wearing down the metal contacts. It has been observed that failure in the population of power transformers is mainly due to ageing and the tap changer is the component with highest contribution of failures [5]. Thus accurate predictions of tap changing operations are necessary to improve the failure prediction in advance, which if accurately predicted could reduce the downtime. Furthermore, one of the objectives of the smart grid is to reduce the cost of operations. Reducing distribution level losses could reduce the cost of operation. However, it is vital to determine actual reduction in losses in the presence of communication and load control.

1.4 Organization of the Report

The this work has two parts and the report is organized as follows, Chapter 2 provides the state of the literature, Literature Review, Chapter 3 presents the consumer impact modeling and analysis part, Chapter 4 presents the system side impact modeling in the presence of data aggregation and conclusion is provided in Chapter 6.

2 Literature Review

Dislike the traditional grid, which relies more on human interaction, the smart grid is however expected to have lesser human interaction. For instance, in case of a power outage in a traditional system, the consumer has to complain about the outage and bear the delays; meter reading and billing requires more man power; these tasks can be taken care of with some form of communication with the grid. One of the objectives of smart-grid is to increase the consumer response through active residential consumer participation (demand response).

The demand response requires certain residential appliance to respond to the control signals sent to them by the distribution system operator. The sensors and processors should work closely to respond to different types of disturbances. Traditionally the grid's flow was bottom up, i.e. the generation units used to respond to the demand. Demand response enables top to bottom approach in which the appliances/loads respond in accordance with the available resources.

2.1 Feeder Level Impacts

Appliance's usage profiling can be very helpful for the smart grids and can support DSM programs since it enables the system to understand the user activities and thus better forecast the load profiles. Appliance identification strategies were developed in [3], [6]-[8]. In [3], a V-I trajectory based taxonomy was presented. The appliances, based on their trajectories are then divided into groups representing similar appliances. In [6], a load monitoring system is presented based on S-Transform. However, the methodologies require load monitoring at appliance level and not the aggregation point, i.e. the main meter. This problem was targeted in [7] and [8] where the appliances are identified at the aggregation point and thus the need of individual monitoring was suppressed. Usage profiling will also help in developing load specific models and better forecast each appliance's demand and will ultimately supplement the development of controlling algorithms.

In terms of controllability, household load can be classified as controllable loads (CL) and non-controllable loads (NCL). CLs can be further categorized as thermostatically controllable loads (TCL) and non-thermostatically controllable loads (NTCL). Examples of TCLs are HVAC systems, refrigerator, freezer etc. Appliances that can be switched on/off directly or can be programmed for their operation fall in the other category.

Work done in [9] and [10] is primarily focused on the controlling of Non-thermostatically Controllable Loads (NTCLs) responding to DSM signal. However a direct control method as proposed in [9] and [10] is not feasible for Thermostatically Controllable Loads (TCLs) due to the constraints like acceptable thermal range and minimum compressor on/off time. Controlling appliances locally, i.e. for each individual household separately brings another concern that there remains no coordination at the community level. Moreover, expanding them to multiple households require information about the usage patterns of other appliances. In [11] an energy consumption scheduling model is presented in the presence of local micro generation which can be expanded from household to the community level. The test cases used for simulations are with arbitrary load profiles of some appliances, unable to control TCLs, and does not represent an actual system. The goal, similar to most of the work available, is to save energy usage cost to the consumer. Furthermore, the algorithm is unable to handle large number of appliances and demands high computational power. Similarly, in [12] a power consumption scheduling scheme, using arbitrary load profiles of NTCLs, handles multiple tasks to schedule. The analysis of algorithm is based upon the impact of number of tasks on the execution time for a constrained environment, which actually did far better than a non-constrained environment.

Significant work is done to manage the demand of multiple customers. For example, in [13]-[16] game theory is utilized to schedule loads at a community level with minimum information exchange. The algorithm proposed in [13] was able to reduce Peak to Average Ratio (PAR) and the total cost in the system. Basically, the game among the consumers is to schedule appliances so that the overall cost of supplying the energy demand is reduced, and ultimately reduce the cost to each consumer. Real time pricing is utilized in [14] to optimally schedule household loads in an attempt to reduce cost and waiting time to the consumer. The automated system is proposed so that the consumer does not have to respond manually to continuously changing prices as it requires training and understanding of the system as well as constant and careful input from the consumer. The work was extended in [15] to realize PAR reduction when compared to percentage of schedulable loads available. The work was further extended in [16] to utilize battery storage system for balancing the supply and demand by charging the batteries at low demand periods and discharging when the demand is high. Finally, the effect of battery capacity and number of users equipped with battery storage system was analyzed. However, throughout the work related to game theory was mostly based on reduction in cost and PAR, and not the user discomfort.

In [17]-[23], TCLs are the main focus. A real time scheduling of deferrable load such as electric vehicles and TCLs is presented in [17] and performance of three scheduling algorithms is compared. The analysis was done from the grid point of view i.e. the impact of scheduling appliances on the reserve capacity. In [18], an algorithm to schedule water heater based on different cost and comfort settings is presented; utilizing the forecasted temperature and price data. However, the model is local to the each house and the idea of immediately turning off an appliance when the cost is high is not implementable to HVAC units as HVACs pose constraints such as *minimum off time for compressor* (which is usually 5 minutes) hence cannot be switch ON/OFF frequently. The HVAC units have some operating constraints that restrict the way they need to be controlled. For instance, when the compressor of an HVAC system is turned "off", the air pressure in the chamber is high and a certain amount of time is needed for the pressure to even out. Restarting the compressor under pressure may cause physical damage [19]. The aggregated models and control strategy proposed in [20] explicitly takes into account the lockout effect of HVAC units which prohibits the unit from turning back "ON" before a certain time. Also this constraint is incorporated in the work done in [19], which demonstrates a comparison of varying the temperature upper and lower bounds. Temperature readings from an office building were used to model this system in order to maintain thermal comfort and power consumption, where multiple units are working in coordination to maintain temperature in a facility. This is very helpful when there is a system available to identify appliances, learn the usage pattern and tune the respective models to help the scheduling algorithms schedule. Then a comparison of different algorithms is compared for the performance metrics such as time to reach comfort band, number of switching i.e. ON/OFF and discomfort duration. Although comparing different comfort bands for the user, it is not giving the control to the user, thus forcing the consumer to stay at higher temperatures for a while.

The aggregated models and control strategy proposed in [20] also explicitly takes into account the lockout effect of HVAC units which prohibits the unit from turning back "ON" before a certain time. Notice that this is a concern when the control signal frequency is very high, hence an algorithm with low frequency for control signal can also help avoid this concern. The implementation of such a system is however, does not seem practical as first, it requires high computational power at the aggregation point to schedule 5000 HVAC units and secondly, the benefits cannot be seen at the distribution transformer as the main idea is to reduce peak on system wide level. Then in case a lockout of majority of the HVAC's population, the algorithm will not be able to perform well. Lastly, the communication requirements for the data and control signals will increase in order to serve a large number of HVAC units at once. A day-ahead scheduler is presented in [21] promising savings in consumer cost, but the user is not given flexibility of choosing temperature set points and deviation from the set point. In [22], a low computational cost scheme using look ahead control approach is proposed, however the controller requires more than one day data.

In majority of the work mentioned above, the main goal remained the reduction in cost. Those who did something about user comfort, has either a high computation resource consuming algorithms or have not given any control to the consumer. Moreover, none has actually discussed the interdisciplinary goals such as the tradeoffs among the power system and the communication system. Improving power system may need some compromising in communication system and vice versa. This work deals in analyzing the trade that can be seen while trying to improve either of the systems so that they can work together in the most efficient manner and the system designers can have improved view before making decisions. To this, a comparison of control signal frequencies is presented. A higher control frequency could be better for the power system as it gives more information of the system and thus better control but at the same time is a burden from the communication and computation point of view.

2.2 System Level Impacts

2.2.1 Importance of Aggregation Interval

End use load shape provides detailed time-of-use information and for evaluating the impact of various types of utility demand side programs. End-use forecasting requires description of the load shape in terms of economic data, customer demographics, dwelling characteristics (i.e. the characteristics of the equipment which causes the demand and weather).The introduction of smart grid helps electric utilities to enable greater monitoring and control of their distribution system. A consequence is more data and data flow over communication network and hence storage and management of data become a big issue. Data collected from each smart meter is approximately of five bytes. So for feeder system with 50,000 meters will have 250,000 bytes of information for every meter read every time [24]. The amount data for different periods of time is shown in Table 1.

Table 1: Data at different time intervals [24]

Time Interval	Number of Meters	Data	Amount of Data
1 Day	50K	5 Bytes	24 MB
1 Month	50K	5 Bytes	720 MB
1 Year	50K	5 Bytes	

Smart meters alone are not sufficient to measure all the parameters to maintain power quality standards. Hence distribution system operators are installing power quality monitoring system (PQMS) based on fixed power quality monitors [25], thus increasing the size of the data to be analyzed. Due to this instrumentation limitations and memory restrictions, non-standard data aggregation intervals are used [26].

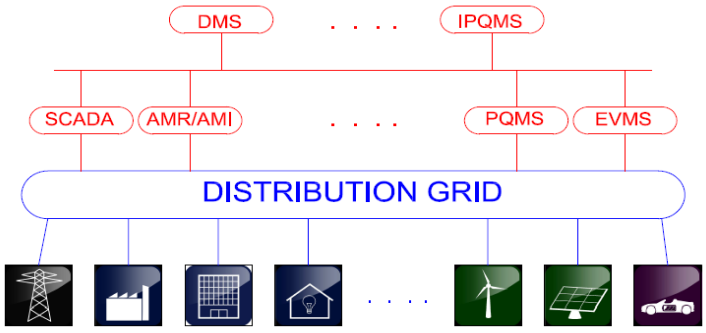


Figure 1: Principle architecture of the smart distribution grid [25]

Other than power quality monitoring, utilities use aggregation intervals to maintain customer privacy. From daily energy consumption data, information such as household occupancy and occupant activities can be derived, hence short intervals can compromise customer privacy, thus leads to use longer aggregation intervals [27].

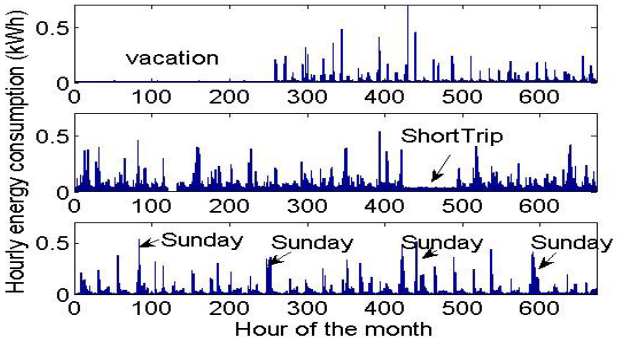


Figure 2: Hourly energy consumption plots showing occupancy status [27]

2.2.2 Voltage Regulation

In a feeder system the voltage drop along the line is calculated as a product of line impedance and total line current. Figure 3 shows the voltage ranges for primary feeder and secondary customer service points for a feeder system. Voltage regulation under ANSI C84.1 standard and the unidirectional nature of power flow can be performed using an on-load tap changing transformer or by using capacitor banks.

Tap changer varies the number of turns in one side of the transformer and thereby, change the transformer ratio. Normally, this can vary between 10-15% in steps of 0.6-2.1%. There are several options to design the control of the voltage. One of them is to set a nominal value of the voltage with a dead band in a point of the line, and to control it with an integral controller [28]. Figure 3 show the working principle of a tap changer incorporated with a line drop compensator circuit which is used to compensate for the voltage drop between the regulator and the load center. In order to prevent excessive operation of tap changer, a time delay is used in order to keep the voltage fluctuation within a desired or predetermined bandwidth [29].

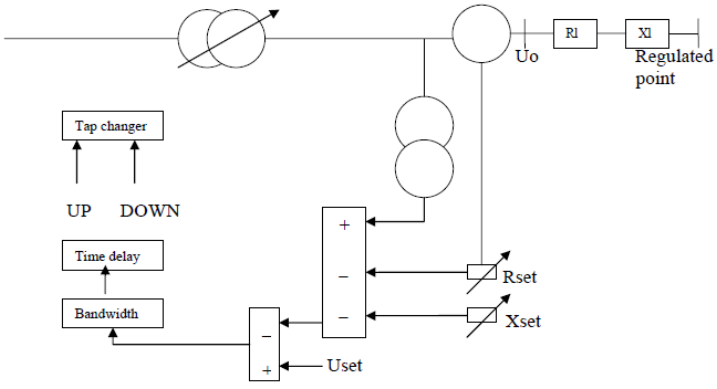


Figure 3: Line-Drop Compensator Circuit [30]

In order to introduce reactive power to system capacitor banks can be used and can be fitted anywhere in the feeder. The line current is reduced if the capacitor is closer to the load center and hence improves the feeder voltage profile. In order match the supplied reactive power to the load, capacitor banks can be fixed permanently or switched in order to prevent overcompensation of reactive power which might lead to

increase feeder level voltage [29].

2.2.3 Voltage Quality Requirement

In order to unify values for the different electrical parameters there are standards to preserve acceptable voltage quality for customers. EN 50160 is presented which gives the main voltage parameters and their permissible deviation in public low voltage (LV) and medium voltage (MV) electricity distribution systems. The technical and economical possibilities needed for the supplier to maintain public distribution systems are provided in EN 50160 [31]. Since we are not considering abnormal operating conditions, EN 50160 is best suited.

3 Residential Level Demand Response Impacts

3.1 Household Load Modeling

The first step of this work is to determine the load shapes of different appliances used at the residential level. At the beginning of this work, limited appliance level data was available for modeling purpose. Therefore this work recorded at developed load models for different residential level appliances. The available residential level load curves that are available in the literature were used to validate the aggregated individual house load shape. The following subsections details the modeling of individual residential loads.

3.1.1 Data Recording

Power profile of few appliances was recorded using Eagle 120 power monitor to analyze their operating behavior which helped in generating the base load profile for different houses. Figure 4(a) represents electric load profile for a household refrigerator. It was noticed that the changes in compressor on-time for a refrigerator is due to the following three reasons: (*a*) door opening (b) High room temperature and (*c*) high cooling load.

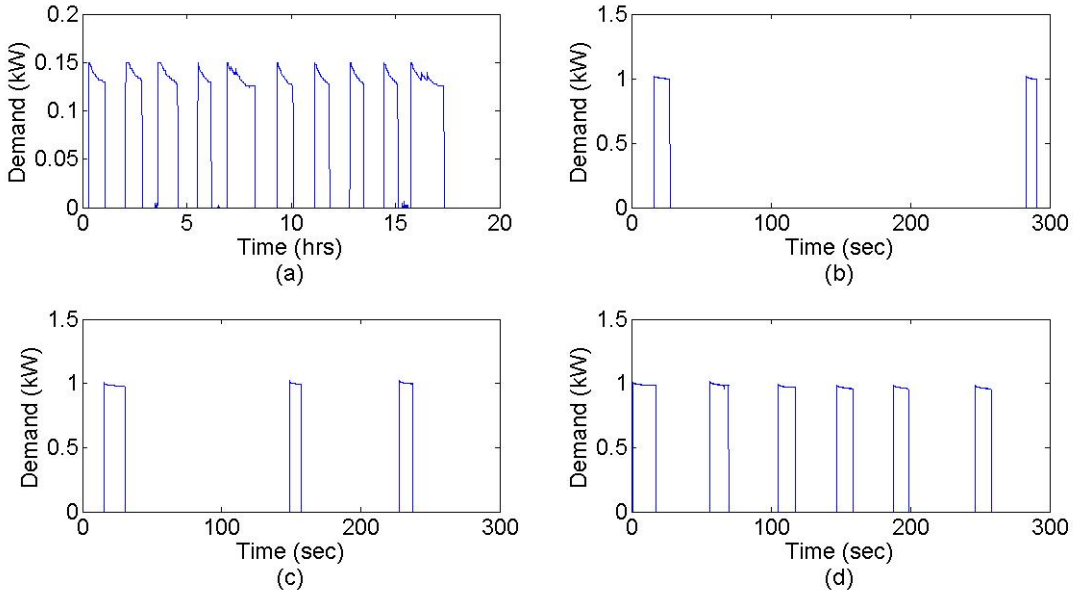


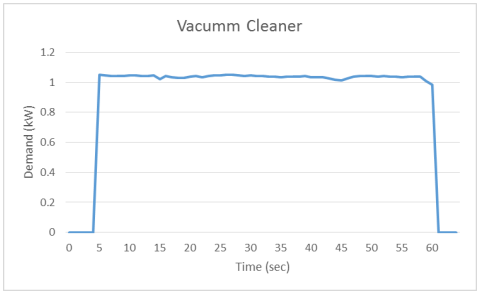
Figure 4: Recorded individual load data (a) Refrigerator, (b) Electric iron with minimum setting, (c) Electric iron with medium setting, (d) Electric iron with maximum setting

It should be noted in the Figure 4(a) the small spikes recorded between actual ON cycles are due to the door opening event which causes the internal bulb to turn on and changes the ON time of the immediately following cycle. The compressor usually remains on for 40 to 80 minutes depending upon the usage. Different selector settings were not analyzed for refrigerator as the same ETP model as that of air conditioning is used to model this load, as explained later.

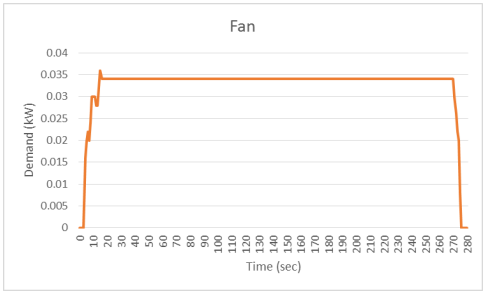
Figure 4(b) represents an electric iron's load profile with the selector knob set to minimum. Notice that the cycles are very less frequent and consequently the average load contribution for this case is insignificant.

The selector knob was then set to midpoint as shown in Figure 4(c). Again, the on cycles are very less frequent i.e. only 3 cycles of maximum 20 seconds duration, within the 4 minutes of recording. Finally, Figure 4(d) represents the load profile with selector set to maximum setting. Notice the difference in frequency of ON cycles which will have significant impact on the average demand posed by this appliance. It was noticed that the average ON duration in a minute remains 10 to 20 seconds. Lower selector settings also keep the iron on for approximately the same duration, as it has to just maintain the iron plate's temperature, however, the cycles become less frequent. The electric iron remains on for 25% to 40% of the time during the operation, and consequently consumes from 25% to 60% of the max rating per minute when the selector is set to max. The different ON-times during the operation are due to pick up from cold iron, pressing/ironing and on stand events.

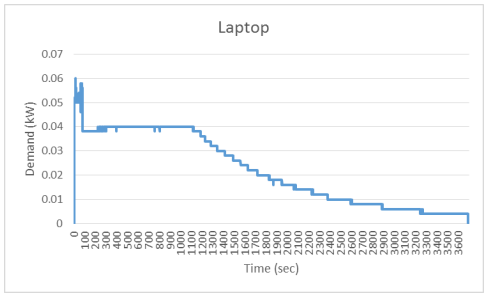
The rest of the electrical appliances selected for this work i.e. microwave oven, fans, lights and laptops are all non-cyclic loads and remain at their rated power level when ON and zero when OFF. Some sample recordings are given in Figure 5.



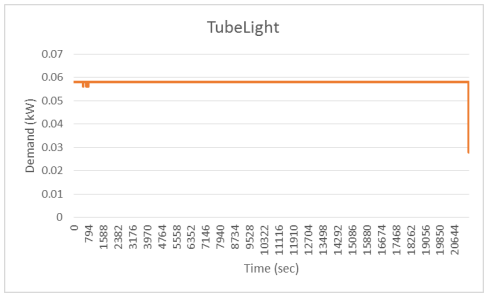
(a) Vacuum cleaner



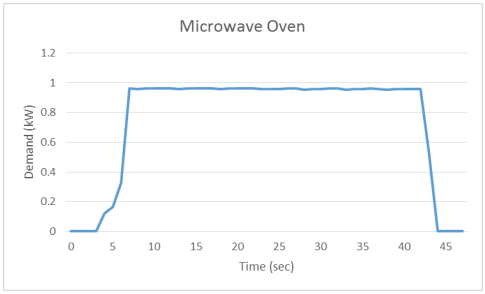
(b) A fan load profile



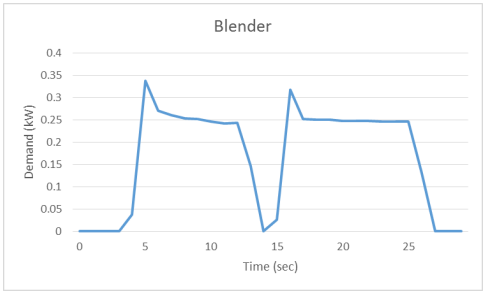
(c) A laptop charging and using



(d) Fluorescent light



(e) Microwave oven



(f) A blender, turned on/off during operation

Figure 5: Load profile

3.1.2 Base Load

A base load profile was needed for each house in order to imitate a real system. Base load is defined as all load types except air conditioning and, which remains there most part of the day. Some of the appliances such as electric iron, laptops and microwave oven are included due to their high power demand or usage frequency. Since household load modelling is not the main focus of this work, a simple, intuition based scheme was developed. More sophisticated techniques including probabilistic methods and Markov chain based models can be found in [32]-[34].

3.1.3 Load Classification

Household appliances can be classified as cyclic (CYC) and non-cyclic (NCYC) based on their demand profile pattern. Cyclic loads change states during their operation e.g. HVAC, refrigerator, and freezer, electric iron etc., whereas non-cyclic loads remain at a certain power level while operating e.g. space lighting, fans, laptops, microwave ovens, LCDs etc.

3.1.3.1 Cyclic Loads

Air Conditioner

The modeling approach that is used to estimate thermal loads is called an equivalent thermal parameter (ETP) modeling approach. This modeling approach has been chosen for the current work because it has been proven to reasonably model residential (and small commercial building) loads and energy consumption and also because it is based on first principles [34].

$$\dot{T}_{air} = \left\{ \frac{1}{R1 \cdot C_{air}} - \frac{1}{R2 \cdot C_{air}} \right\} \cdot T_{air} + \frac{T_{mass}}{R2 \cdot C_{air}} + \frac{T_{out}}{R1 \cdot C_{air}} + \frac{Q}{C_{air}} \tag{1}$$

$$\dot{T}_{mass} = \frac{T_{air}}{R2 \cdot C_{mass}} - \frac{T_{mass}}{R2 \cdot C_{mass}} \tag{2}$$

where, C_{air} is air heat capacity (Btu/^{0}F), C_{mass} is mass (of the building and its content) heat capacity (Btu/^{0}F), T_{out} is ambient temperature (^{0}F), T_{air} is air temperature inside the house (^{0}F), T_{mass} is mass temperature inside the house (^{0}F), Q_h is heat rate for HVAC (Btu/hr.), Q_i is heat rate from other appliance, lights, people etc. in the residence (Btu/hr.), Q_s = heat gain from solar (Btu/hr. or watts), $R1 = 1/UA_{insulation}$, $UA_{insulation}$ is heat gain/loss coefficient (Btu/^{0}F.hr) to the ambient, $R2 = 1/UA_{mass}$, UA_{mass} is heat gain/loss coefficient (Btu/^{0}F.hr) between air and mass, $Q = Q_i + Q_s + u*Q_h$ and u is on/off control variable.

Therefore the Euler's equivalent of the model is,

$$T_{air}(k+1) = T_{air}(k) + h \cdot \left\{ \frac{1}{R1 \cdot C_{air}} - \frac{1}{R2 \cdot C_{air}} \right\} \cdot T_{air}(k) + h. \left\{ \frac{T_{mass}(k)}{R2 \cdot C_{air}} + \frac{T_{out}(k)}{R1 \cdot C_{air}} + \frac{Q}{C_{air}} \right\} \tag{3}$$

$$T_{mass}(k+1) = T_{mass}(k) + h \cdot \left\{ \frac{T_{air}(k)}{R2 \cdot C_{mass}} - \frac{T_{mass}(k)}{R2 \cdot C_{mass}} \right\} \tag{4}$$

where, h is sample height in hours i.e. the step size, in our case is 1/60 hour or 1 minute.

The relay in TCLs needs to be molded when a controller is developed. The output frequently changes according to minute temperature changes (temperature does not remain constantly at a particular level due to various changes in the environment and can force the relay to respond to those changes), and shortens

the life of the output relay or unfavorably affects some devices connected to the temperature controller. To prevent this from happening, a temperature band called hysteresis is created between the ON and OFF operations [35] as shown in Figure 6.

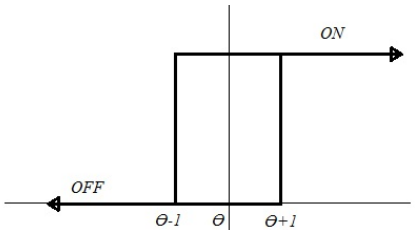


Figure 6: Hysteresis loop for the thermostat relay

The demand profile can be generated thus,

$$P_{ac}^{n}(k) = u_{ac}(k) \cdot P_{ac\_rated}^{n} \tag{5}$$

The thermal parameters for each house can be different and demand a survey or knowledge of typical ranges for them to be used. This is where appliance identification techniques could help in a real system by learning about appliance specific demand and behavior and then utilize the information to tune the respective model's parameters. Therefore, in an attempt to avoid the mentioned demands, i.e. doing surveys or having an appliance level identification system, we chose to tune the ETP model parameters to represent the most favorable conditions.

Based on the ASHRARE 55 [2], Table 2 shows limits on temperature drifts. Using this information, we tweaked the parameters to match the requirements for each house with 98 ^{0}F as the design day outdoor temperature for Wichita, KS and 74 ^{0}F as the desired set point [36]. Figure 7 shows sample internal temperature variation with respect to the outdoor temperature.

Table 2: Limits on temperature drifts

Time Period (hrs.)	0.25	0.5	1	2	4
Max Operating temperature change allowed (Degree F)	2	3	4	5	6

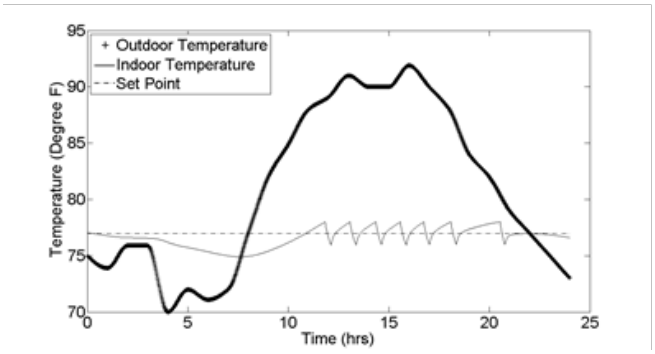


Figure 7: Indoor temperature profile against outdoor temperature for thermostat set to 77^{0}F

Refrigerator

The same model as that of air-conditioning is used for refrigerator, but with different thermostat settings and thermal parameters. Power profile of few refrigerators was recorded using Eagle 120 [37] power

monitor to analyze their operating behavior. The door opening event plays a critical role in defining the load profile for refrigerators. Based on this observation, in order to imitate random door opening events, the outdoor temperature (i.e. indoor house temperature) during those events was changed. To imitate different durations of door opening events, the range of outdoor temperature was randomly selected between 120 ^{0}F – 130 ^{0}F. For normal operations, the outdoor temperature is randomly selected between to 74 ^{0}F – 79 ^{0}F for all houses. The events are induced for morning between 5:00 a.m. to 7:00 a.m., afternoon between 12:00 p.m. to 1:30 p.m. and evening between 7:00 p.m. to 8:00 p.m. The demand profile is then generated using,

$$P^n_{refri}(k) = u_{refri}(k) \cdot P^n_{refri\_rated} \tag{6}$$

The range and power level for refrigerator's thermostat setting is selected randomly for each unit between 35 to 39 ^{0}F and 0.160 to 0.240 kW respectively. Figure 8 represents a sample demand profile for household refrigerator. Notice the change in cycle width around 6:00 a.m., 12:00 p.m. and 8:00 p.m.

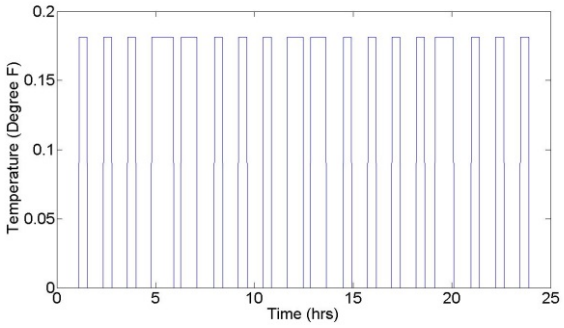


Figure 8: Demand profile for household refrigerator

Electric Iron

Electric iron is a short duration load, typically 15 to 30 minute, with high demand requirement. Although cyclic in nature, the load profile as mentioned in previously, cycles very frequently between ON and OFF state and thus does not require finer scale (per second) modelling. The cycles can start any time during each minute and thus on average can demand 25% to 60% of the rated value. It was assumed that electric iron is used more on weekends than it is used on weekdays.

First of all, total number of electric irons was chosen randomly between 1 and 2 for each house. Then the number of events for the whole day was chosen for each appliance with 1 being mean and 0.3 being the standard deviation. This means that the majority of the time, number of events for the appliance will remain between 0 and 2. Then the time of event is chosen from:

$$wd \in \{7, 21\} \text{ and, } we \in \{11, 17, 21\}$$

where, *wd* and *we* represent weekday and weekend respectively. The standard deviation of 10 minutes is chosen for each event. And finally, the mean and standard deviation for the duration of usage were chosen as 10 minute and 2 minute respectively. Using these daily event occurrence times, the per minute load profile for electric iron is generated from the range mentioned in equation (7). Notice that the electric iron's rated power is never achieved. This due to the fact that we are averaging the demand on per minute basis and our analysis on electric iron's load profiles showed that it remains below 60% of the rated power. We chose the mean value of 42% of P^{rated} with the standard deviation of 6% of the rated power to generate electric irons load profile.

$$P_{ei}(k) = \begin{cases} 0 & if\ sw = OFF \\ P_{rated} & if\ sw = ON\ \ p_k(on) \geq \tau \\ 0 & if\ sw = ON\ p_k(off) \leq \tau \end{cases} \tag{7.a}$$

$$0.25 * E_{ei}^{rate} \leq P_{ei} \leq 0.6 * E_{ei}^{rate} \quad (7.b)$$

where, P_{ei} is electric iron's demand in kW for the interval k, and P_{rated} is 1 kW. Normal distribution is used for each appliance with the mean and standard deviations mentioned in Table 3, representing the parameters for appliances including NCLs discussed later.

Table 3: Parameters used for household appliances for load generation

	Electric Iron		Laptop		Oven		Fan		Lighting	
	wd	we	wd	we	wd	we	wd	we	wd	we
Event Mean	{7, 21}	{11, 17, 21}	{19, 22}	{13, 19, 22}	{7, 19}	{7, 13, 19}	19	{11, 16, 19, 22}	{6, 19-22}	{6-7, 13, 15, 18-22}
Event Standard Deviation	10		10		10		20		20	
Mean Duration	10		90		5		200		5-40-120-300	
Standard Deviation of Duration	2		10		2		50		2-10-30-40	

3.1.3.2 Non-Cyclic Loads

Lighting Load

Using the parameters mentioned in Table II above, with an increment of 0.005 kW, the power ratings in kW, used for lighting load when ON are in the range,

$$0.010 \leq P_{lights}^{rated} \leq 0.060 \quad (8)$$

To imitate different usage patterns, lighting load is divided into two groups. More frequent, i.e. for rest rooms etc. and less frequent, i.e. for rooms etc. That is why the mean duration and standard deviations mentioned in the table for lighting load have varied range. Since the load profile was generated for multiple days, the status of lighting load was carried to the next day to keep it more realistic.

Fans

This load type is more consistent and can remain there for longer duration especially at night time and mid-day. Again, using the parameters mentioned earlier in Table II the load profile for fan load is generated. Similar to lighting load, the status was carried to the next day. The power rating of fan load in kW is chosen from the range,

$$P_{fan}(k) = \begin{cases} 0 & if\ p(sw) \leq \tau \\ P_k & if\ p(sw) \geq \tau \end{cases} \quad (9.a)$$

$$0.060 \leq P_{fan}^{rated} \leq 0.090 \quad (9.b)$$

Laptops

Laptops can take anywhere between 60 to 120 minutes to completely charge. Although, compared to the rest of the load, the demand is very low, 0.060 to 0.070 kW in most cases, we added this load as there can be multiple number of this load type in a house, charging at different times. Again, using the parameters mentioned in Table II, the load profile was generated with the power ratings in kW chosen from,

$$0.060 \leq P_{laptop}^{rated} \leq 0.070 \quad (10)$$

Microwave Oven

Similar to electric iron it can create short duration peaks mostly during early morning, afternoon and evening. It is modelled to have duration of anywhere between 1 minutes to 9 minutes in most cases, multiple times during mornings, afternoons and evenings and its power rating is represented mathematically by,

$$0.900 \leq P_{mw}^{rated} \leq 1.100 \quad (11)$$

3.1.4 Load Profile Generation

Each load I and its demand $P_i(k)$ for k^{th} interval is then used to first generate the per minute base load profile for each appliance. To simulate the fact that an event could occur any time within a minute, $P_i(k)$ is divided by a random number (*rnd*) for the first and last interval of operation, where $rnd \in \{1, 2, \ldots, 60\}$.

$$P_i = \begin{cases} P_i / rnd & for\ k = start \\ P_i & for\ start + 1 \le k \le end - 1 \\ P_i / rnd & for\ k = end \end{cases} \tag{12}$$

This is done for all loads other than air conditioner and refrigerator as these two are modelled differently. The aggregated load AL profile for N houses at any instant k is thus,

$$AL(k) = \sum_{n=1}^{N} P_{refri}^{n}(k) + \sum_{n=1}^{N}\sum_{i=1}^{I} P_i^n(k) \tag{13}$$

A sample aggregated load profile of 5 houses for the duration of 4 months, excluding air conditioning load is shown in Figure 9 Notice some part of the demand is increased in the midsection between 12:00 p.m. to 4:00 p.m. This represents weekend load and is meant to imitate that more people are at home during weekends. The low maximum demand for 5 houses is due to very small number of appliance used for each house. Inclusion of more appliances like electric stove, electric water heater, kettle, television, freezer etc. would help; however, the main interest was just to get a typical load profile.

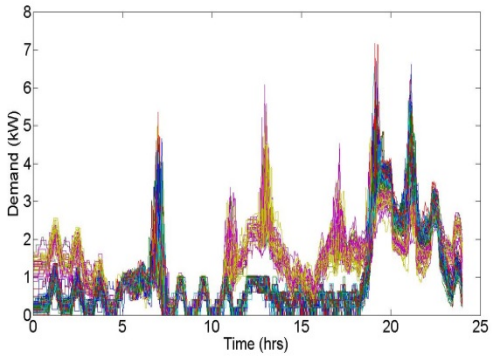


Figure 9: Aggregated load profile of 5 houses for 123 days

3.2 Algorithm and Simulation

The objective function is to minimize the variation in user's preferred thermostat setting for the air conditioning load respecting the power and thermal constraints, in other words, minimizing the user discomfort while attempting to reduce the peak demand at the aggregation point i.e. the transformer. The objective function is therefore,

$$\min \sum_{k=1}^{K}\sum_{n=1}^{N} \left\{\theta^n(k) - \tilde{\theta}^n(k)\right\}^2 \tag{14}$$

where, N is the number of air conditioning units, $\tilde{\theta}^n$ is the user's preferred thermostat setting and, θ^n is optimized temperature setting for interval k. Preferred temperature range provided by the user is considered constant for the whole day. Thus the optimization problem is subject to,

$$\theta^n_{\min} \le \theta^n \le \theta^n_{\max} \tag{15}$$

And the P_{max} constraint is,

$$\sum_{n=1}^{N} P_{ac}(k) \le P_{\max\_ac}(k) \tag{16}$$

where, P_{max_ac} is the maximum power constraint at k^{th} interval provided by the utility for the aggregation point i.e. the transformer serving N houses.

An algorithm using forward dynamic programming to find the optimum solution to the problem is written based on day-ahead data of temperature and price signal. The motivation behind using forward dynamic programming is that the appliance commitment problem is similar to unit commitment problem on generation side. In power plant unit commitment problem, there are a number of generating units available with their operating constraints and cost of operation known ahead of time. To serve the expected demand, combinations of units with minimum cost, respecting all the constraints are saved for the whole day initially with multiple routes. Then in the backward direction, the total minimum cost route is selected.

In appliance commitment problem, the same can be applied to TCLs units. With day-ahead information about the constraint signals and the expected demand from each appliance, to make decisions based on the deviation from thermostat settings as the cost and available combinations for the number of units to be served, the problem can be solved.

To determine number of maximum possible states per interval for each control frequency, the information from Table I is used. Figure 10 demonstrate the steps that can be taken per interval for 15, 30 and 60 minute control, respectively.

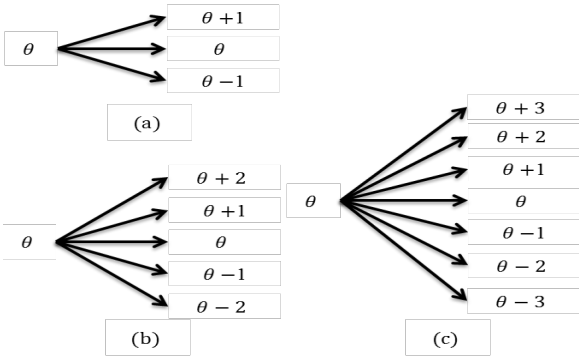


Figure 10: Possible steps for (a) 15 minute control, (b) 30 minute control, (c) 60 minute control

For instance Figure 10(c) shows that in one hour, there are 7 possible (steps) for the Θ to choose from i.e. it can choose to directly set the temperature 1, 2 or 3 degrees up or down. However, as can be seen in Figure 10(a) which is representing 15 minute control, the ASHRAE standard limits the range to only 1 degree up or down. The maximum possible states per stage based on the steps can therefore be calculated as,

$$states = steps^N \tag{17}$$

In order to avoid new peaks that usually show up when DSM programs are utilized, the power constraint signal was generated based on the price signal. The motivation behind using the price signal to determine the constraint signal at transformer level is that, firstly, the utility is able to determine the power constraint for each transformer at distribution level. Secondly, the price signal indicates the system operating

conditions incorporating both wholesale and retail markets. This can be used as an indicator to determine the load that can be connected. When the load is shifted in peak hours, the difference between the desired demand based on the price signal and the expected demand after the demand response is minimized. This will result in the actual price signal deviating less from the forecasted signal thus enabling the wholesale market to plan in advance and reduce the reserves. Moreover, managing load at the transformer level will also help in maintaining the desired load on each distribution transformer and thus will support equipment live improvement programs. Figure 11 represents the flow chart of the algorithm developed to solve the problem.

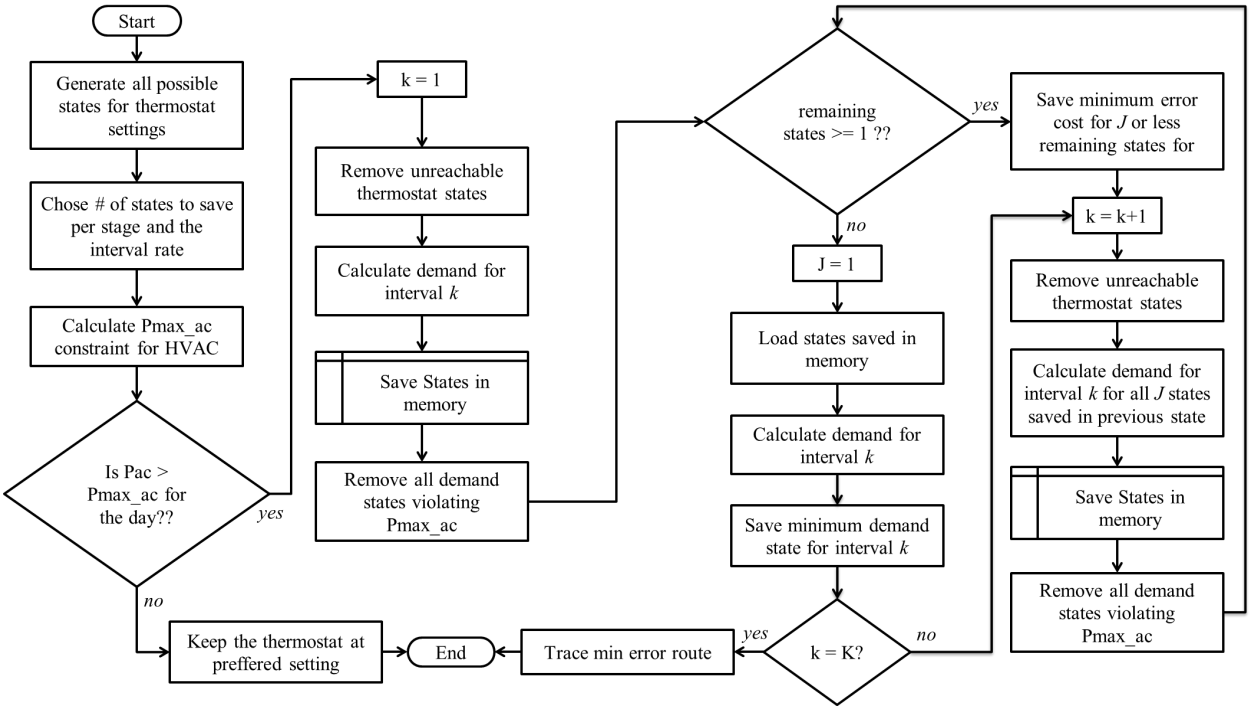


Figure 11: Algorithm using forward dynamic programming

3.2.1 Simulation Procedure

For simulation purpose, the residential load profiles were generated for five houses. Each house is assumed to have best insulation and same HVAC units as the all houses are considered of same size. The transformer is assumed to serve only these five houses. Since the HVAC parameters were tuned to perform as design day, choosing a house size is not significant.

The per hour price signal for May 1st to August 31st, 2012 from ComEd Illinois [38] was used in this work. Figure 12 represents daily price signal averaged for the entire 123 days. Notice that the maximum average value is approximately 5.5 cents.

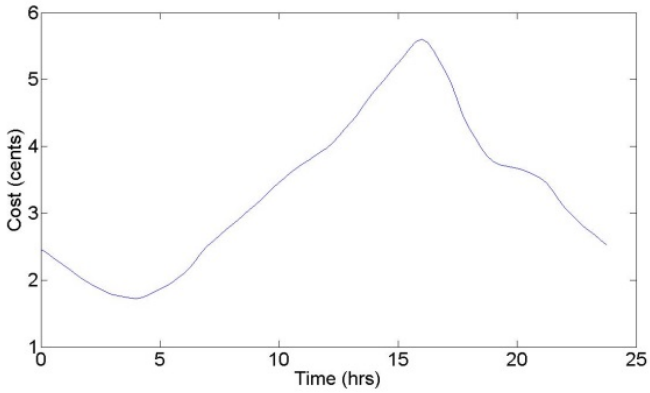


Figure 12: Average price signal for 123 days

Figure 13 represents histogram of the price signal and it is obvious from the histogram that very rarely the cost goes beyond 4 cents. This information was utilized to set inequality *p>4* for the power constraint signal function. The different colors represent days.

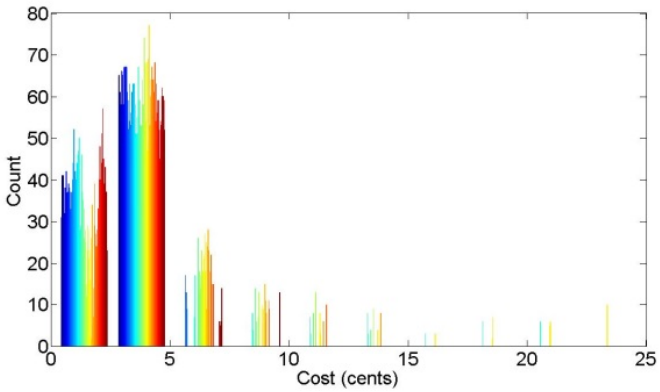


Figure 13: Histogram of the price signal

A simple function for constraint signal is then written as follows,

$$P_i = \begin{cases} PT, & 0 \le p \le 2 \\ 0.9 \cdot PT, & 2 < p \le 2.5 \\ 0.8 \cdot PT, & 2.5 < p \le 3 \\ 0.7 \cdot PT, & 3 < p \le 3.5 \\ 0.6 \cdot PT, & 3.5 < p \le 4 \\ 0.5 \cdot PT, & p > 4 \end{cases} \tag{18}$$

where, PT = 15 kVA is Transformer's rating, p is cost in cents.

As only air conditioning load is being controlled, therefore, the power constraint $P_{\max\_ac}$ for air conditioning load can be calculated by,

$$P_{\max\_ac}(k) = P_{\max}(k) - AL(k) \tag{19}$$

Since we have tuned our ETP model parameters for Wichita-KS, weather data for the same location was downloaded for the same summer duration as that for the price signal, from [39]. Average temperature with standard deviations is shown in Figure 14.

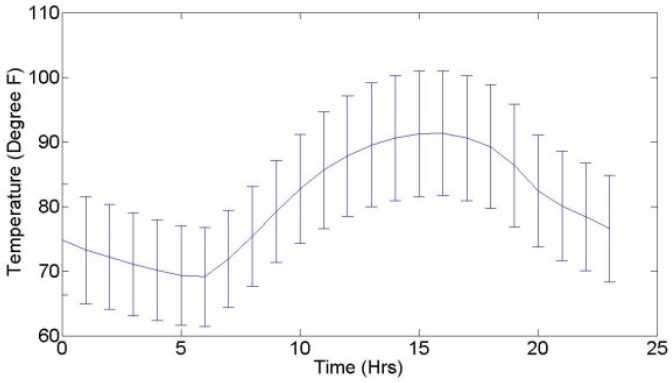


Figure 14: Average outdoor temperature of 123 days with standard deviation bars

Another challenge was to choose number of states to be saved per stage (succeeding interval) for the program. Table 4 represents the effect of number of users and steps on the possible combinations of states for 60 minute control rate.

Table 4: Effect of number of users and steps on computation requirements

		N	1	2	3	4	5	6	7	8	9	10
Steps	**3**		3	9	27	81	243	729	2187	6561	19683	59049
	5		5	25	125	625	3125	15625	78125	390625	1953125	9765625
	7		7	49	343	2401	16807	117649	823543	5764801	40353607	282475249

Notice the significance of increasing the number of houses (equal to number of HVAC units) *N* or number of *steps*; increasing either of them will increase the amount of computation required to solve the problem and will ultimately require more sophisticated system. Fortunately, in dynamic programming a constrained system can help avoid some portion of states. However, that portion can or cannot be significant help, thus requires some basic analysis to find out enough number of states to be saved for each control rate and its respective possible steps.

An initial test with the design outdoor temperature of 98 ^{0}F and 74 ^{0}F thermostat setting was run to target this problem. All the HVAC units were allowed full deviation range i.e. 69-79 ^{0}F, and were assumed to be having same power rating i.e. 3kW as well as the starting point temperature. We chose 3kW power rating for each HVAC unit so that the maximum coincident demand matches 15kW which is the power rating of the transformer supplying these loads. With only HVAC load in system, the simulation was run for each control frequency to acquire maximum achievable peak reduction and PAR reduction with all states saved per stage of the forward dynamic algorithm. Table 5 presents the results of the initial test.

Table 5: Initial analysis results to see the effect of reducing the number of saved states

Control Frequency	Possible Steps	Percentage Reduction	0.1	1	5	10	100
60	7	PAR	9.99	**9.76**	-	-	9.99
		Peak Demand	31.03	**31.03**	-	-	31.03
60	5	PAR	7.45	**9.8**	-	-	9.99
		Peak Demand	29.66	**31.03**	-	-	31.03
60	3	PAR		8.96	**9.93**	9.93	10.12
		Peak Demand		29.66	**29.66**	-	29.66
30	5	PAR	31.6	**42.77**	-	-	42.89
		Peak Demand	48.7	**56.52**	-	-	56.52
30	3	PAR		31.15	**42.02**	42.46	42.79
		Peak Demand		47.83	**55.65**	-	55.65
15	3	PAR		37.54	**53.05**	53.83	54.2
		Peak Demand		53.33	**64**	-	64

From the table, using the percentage of steps that give closest results to the column for 100% states saved, the number of states to be saved to get reasonable results can be calculated using,

$$sts = \frac{ps * steps^{N}}{100} \tag{20}$$

where, *sts* is states to save and *ps* is percentage of states chosen from the Table IV. Since number of states cannot be in fractions, *sts* is rounded to the nearest integer. With the respective *sts,* the algorithm is then run for 500 iterations, for N=5 users. The results are discussed in next chapter.

3.2.2 Results

All the results are within 95% confidence interval. In Figure 15, a comparison of different steps for the 30 minute control signal is shown representing benefits to the utility. Figure 15(a) and (b) represent percentage reduction in total number of P_{max} violations and the violation energy, before and after scheduling. It should be noted that the higher steps did not imprive the performance.

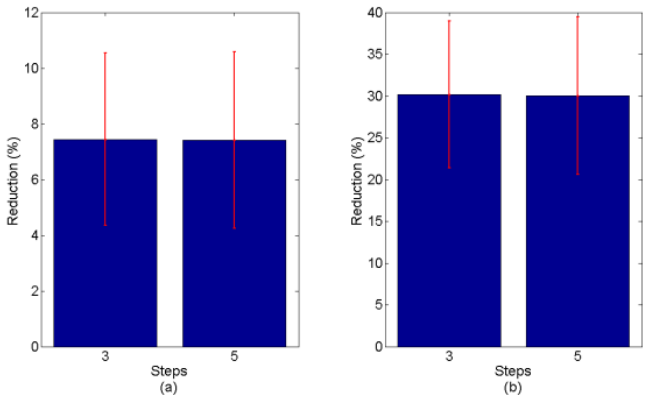


Figure 15: Percentage reduction in (a) per minute violation count, (b) violation energy

Similar is the case with Figure 16(a) which shows peak reduction, infact notice that the maximum duration of sustained violation was degraded as can be seen in Figure 16(b).

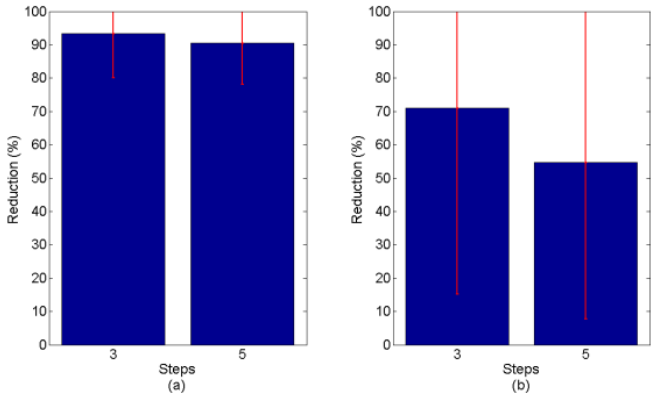


Figure 16: Percentage reduction in (a) peak demand, (b) maximum duration of sustained violation

From consumer point of view, the average total deviation from the setpoint (Figure 17(a)) throughout the day was not much improved, however the maximum duration of deviation from preferred thermostat settings is reduced as is reflected by the mean and standard deviation values in Figure 17(b).

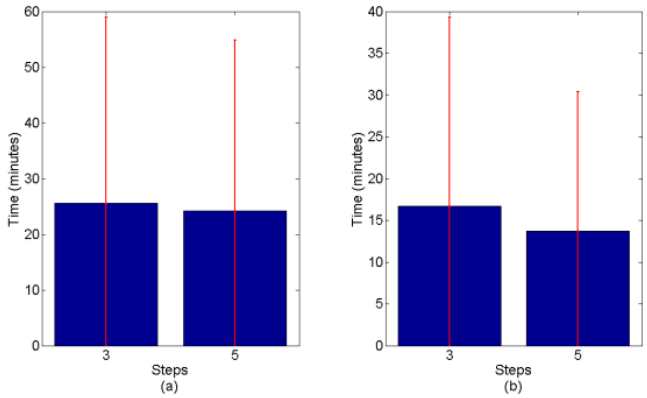


Figure 17: (a) Total deviation from preferred thermostat settings in minutes, (b) Maximum continuous deviation from preferred thermostat settings in minutes

Figures 18, 19 and 20 represent similar plots for the 60 minute control with 3, 5 and 7 steps possible, as discussed earlier. Although there can be seen some benefit in chosing more steps per stage, the overall reduction achieved in terms of the violation energy is very small, as shown in Figure 18(b). Also, the violations in any form that can be seen by the constraint signal naturally reduced due to the averaging for a much wider time slot. The standard deviation bars below zero represent that in some cases the violations were actually increased after scheduling. This is due to the lost of finer control when 60 minute control signal is chosen. Also, there wasn't much achieved in peak reduction, neither in suststained violation reduction when more steps were chosen, in fact the performance was actually degraded. Finally, nothing significant was achieved from the consumer point of view as well as can be seen from Figure 20(a) and (b).

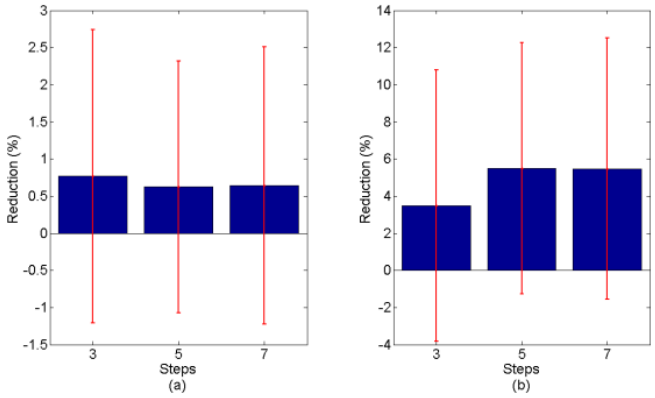


Figure 18: (a) Percentage reduction in per minute violation count, (b) Percentage reduction in violation energy

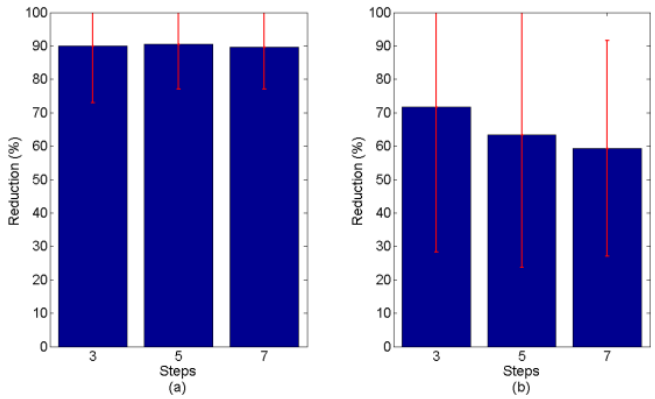


Figure 19: (a) Percentage reduction in peak demand, (b) Percentage reduction in maximum duration of sustained violation

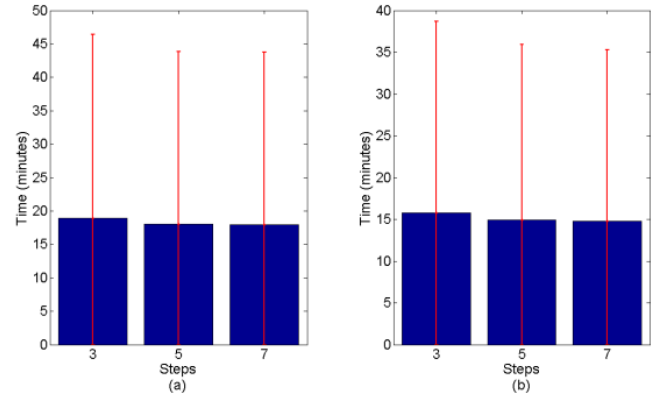


Figure 20: (a) Total deviation from preferred thermostat settings, (b) Maximum continuous deviation from preferred thermostat settings

Since the 15 minute control can not take more than 3 steps, Figures 21, 22 and 23 compares all three control sampling rates with 3 steps. It is very obvious from Figure 21(a) and 21(b) that 15 minute control did best, however, things changed when reduction in peak demand and maximum sustained violations were compared. Notice in Figure 22(a) and (b) that the percentage peak reduction achieved for each control signal is very similar and may not help much as a decision making factor but the maximum duration of sustained violation is improved a little bit. This is due to the control available for wider time slot in case of 30 and 60 minutes sampling i.e. once a decission is made about the entire time slot, it is for the entire duration of that wider slot, hence on average it performs better. The benefits to the consumer are sorted in Figure 23. As can be seen in Figure 23(a), 60 minute control did best in terms of deviation from preferred thermostat settings with mean value of 10 minutes lower than 15 minute control, however, the maximum continuous deviation from the preferred setting did not change significantly.

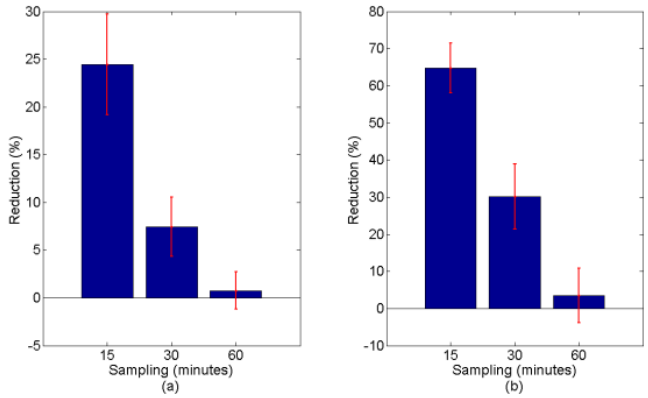


Figure 21: Percentage reduction in (a) per minute violation count, (b) violation energy

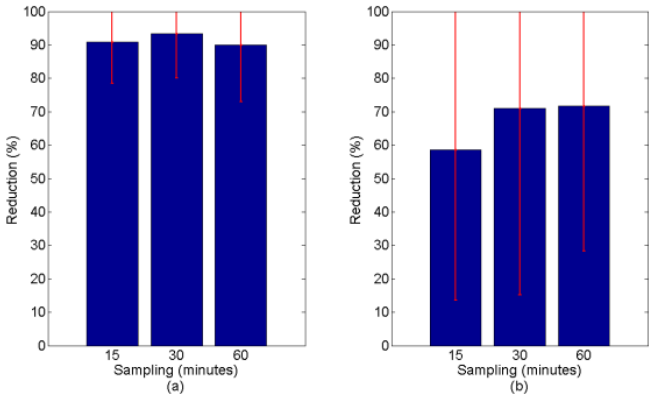


Figure 22: Percentage reduction in (a) peak demand, (b) maximum duration of sustained violation

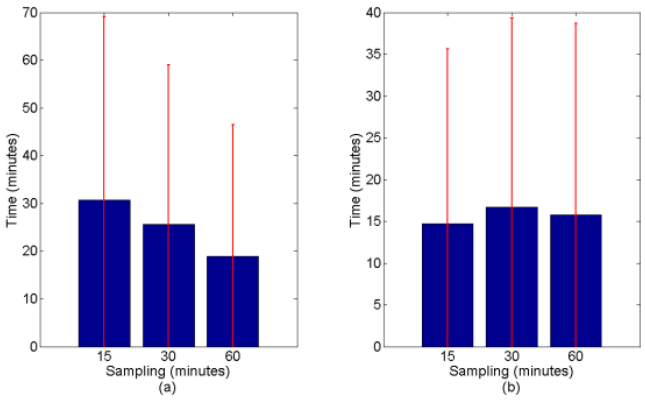


Figure 23: (a) Total deviation from preferred thermostat settings, (b) Maximum continuous deviation from preferred thermostat settings

4 Communication Impacts on the Grid

4.1 Data Aggregation

The focus of this work is to determine the link between the data and power network layers. Once the relationship between the power network and data layers are determined then similar analysis could be done to determine the relationship between the data and cyber network layers. Once the data is aggregated and the demand is computed for a particular aggregation interval, this demand data would be used for forecasting the performance of the system. Since the average demand is used, the spikes and dips within that period are masked as shown in Figure 24. This masking would create an error in the performance estimation.

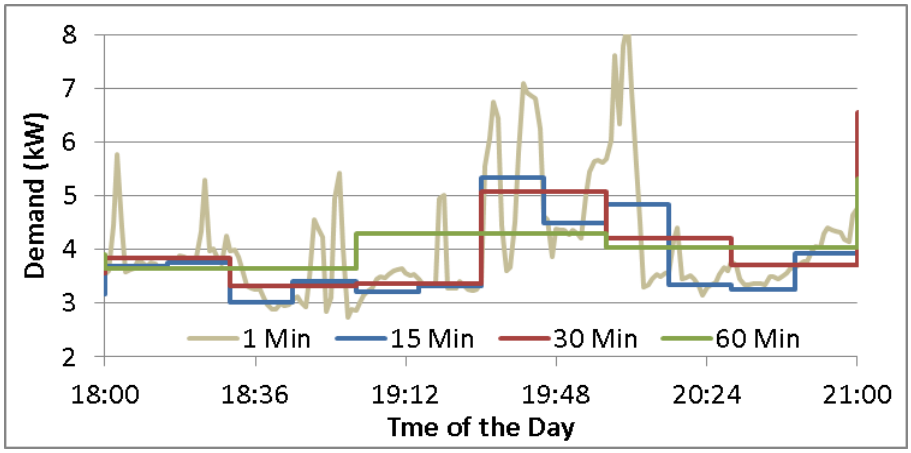


Figure 24: Loss of information due to aggregation

The following steps were taken to determine the relationship between data aggregation interval and forecasting error for tap-change estimation and power loss estimation.

Step 1: One minute demand for each consumer needs to be determined or generated.

Step 2: Determine the points of data aggregation. The data aggregation is geographical location dependent. Typically based on available bandwidth and packet size, optimal consumers within one aggregation node need to be determined. Furthermore, based on the total number of consumers, required levels of aggregation needs to be determined. Since focus of this work is limited to aggregation data interval and power system performance estimation, one level aggregation with different intervals is sufficient.

Step 3: Run time sequential voltage drop analysis and power flow analysis for the given time interval.

Step 4: Repeat step 3 for multiple time intervals.

Step 5: Determine the difference between the estimated values for both number of tap changes and power loss for each time interval and the reference time interval. The total prediction error is computed as

$$\delta_{\tau} = \frac{\kappa_{\tau} - \kappa_{ref}}{\kappa_{ref}} \tag{21}$$

where κ_{τ} is the measurement of the performance parameter (in this work it is either the total number of tap changes in a given month or total line-loss in a month) for the given aggregation interval τ, and κ_{ref} is the same for the reference aggregation interval. For accuracy the reference time interval will be the smallest time interval.

Step 6: Statistically determine the significance of different contributing factors for the model development. In this work the following are considered as contributing factors.

- Size and type of the distribution system.
- Season or month of the year. This is included to ensure any changes due to month are incorporated.
- Aggregation interval.
- Combination of these factors.

Step 7: Once the significant contributing factors are determined a relationship between the contributing factors and the prediction error will be determined.

4.2 Model Development

In this work IEEE 13 and 34 node test feeders were used [40]. The short and relatively highly loaded IEEE 13 node test feeder consist of unbalanced spot and distributed loads while one substation voltage regulator and two shunt capacitor banks regulates the feeder voltage. The very long IEEE 34-node test feeder consists of two-step-type voltage regulators and capacitor banks to satisfy the ANSI voltage standards.

Since both these feeders have limited information for time sequential analysis, appropriate one minute load shape needs to be modeled. Load profiles for this study were developed using available data [41]. The data were separated into three types of houses: i) Detached; ii) Semi-detached; and iii) Terraced. The proprietary data were statistically analyzed and a model was developed to extend the number of houses to the required level for IEEE 13 and 34 bus systems. It was determined that lognormal distribution (shown in Figures 25, 26, and 27) could be used to extend the load curve by changing the number of houses. Thus individual mean and standard deviation of each day of the three type of houses available were generated for a period of one year and were used to randomly generate more number of houses to create the load shape for different node of the feeder system.

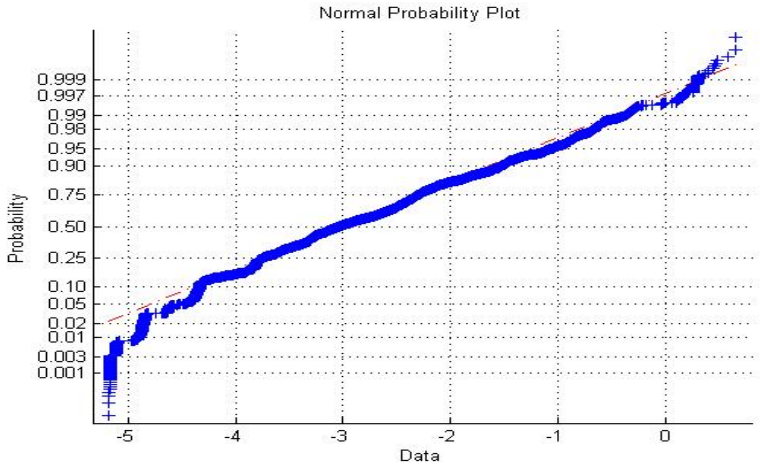


Figure 25: Lognormal plot for Detached houses

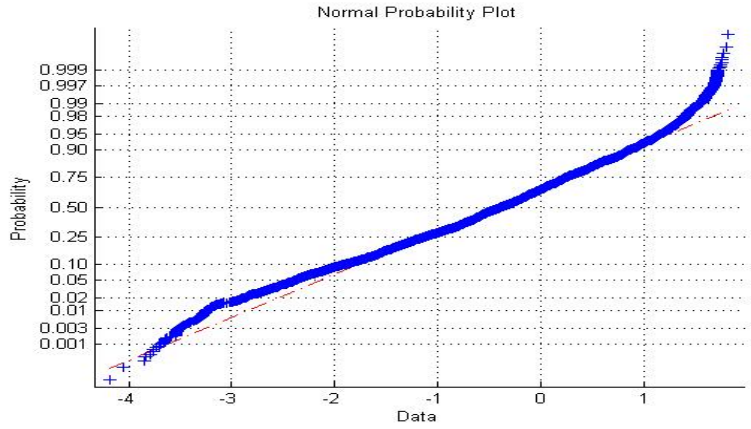


Figure 26: Lognormal plot for Semi-Detached houses

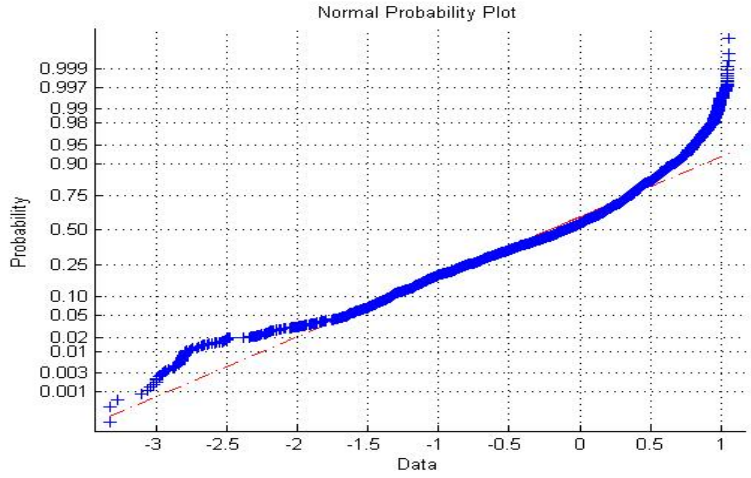


Figure 27: Lognormal plot for Terraced houses

It is assumed the system consists of residential customers and commercial load shown in Figure 28. Nodes with balanced three phase loads were considered as commercial loads. Total number of homes for each node was calculated using 0.55 coincidence factor [42].

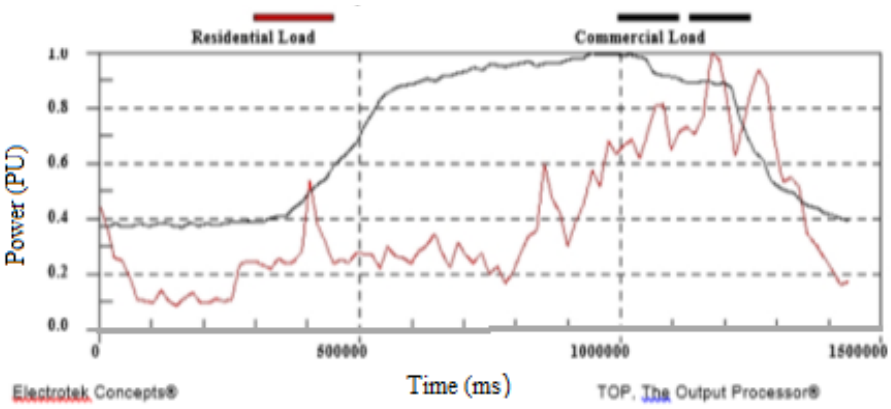


Figure 28: Time-series load profiles (one-minute interval)

4.3 Estimation Error Analysis

4.3.1 Tap Changer Modeling

The estimated number of tap changes and total power loss at different demand interval was monitored and compared with one minute values, which were considered as reference values. Due to the regulator delay settings one minute demand was considered as the real time demand. If needed, the same procedure could be repeated with smaller reference demand intervals. Initially the estimation error in the total number of tap changes for a year is determined for all the three contributing factors defined in the previous section, assuming all of them contribute to the prediction error. When more than one input / contributing factor is suspected to influence a relationship, Design of Experiments (DoE) can be used to determine the significance of each factor and to develop a predictive equation [43]. This work uses DoE to determine the significance of each contributing factor towards the prediction error. As the initial step, prediction errors δ_K for 13 node and 34 node systems were used to determine the influence of each contributing factor. The following aggregation intervals were used: 1, 5, 15, 20, 30, 40, 45 minutes and 1 hour. Each month of the year is considered as the contribution factor. Half normal plot for this experiment is given in Figure 29. Factor A is the aggregation interval, factor B is the month of the year, and factor C is the type of the system (13 bus or 34 bus).

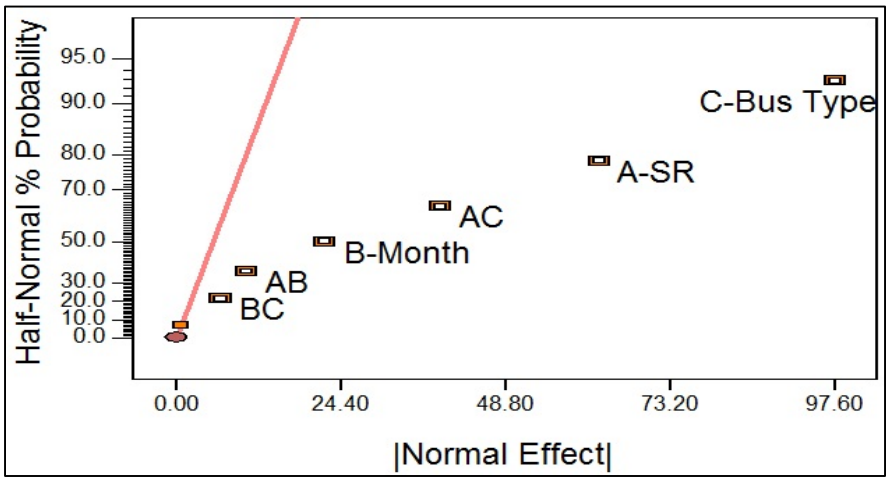


Figure 29: Half-normal plot for the combined 13 node & 34 node feeder analysis

Unimportant factors are those close to the indicated fitted line. If a factor is close to the fitted line, then it has an effect modeled by a normal distribution with near zero mean. The unimportant factors can be eliminated in the modeling [44]. From Figure 29 it can be seen that the type of feeder has the most influence on the prediction error and the aggregation interval is the next most influential factor. Therefore it is determined that each feeder needs its own model for error prediction. The following two sections describe the tap change and power loss modeling for 13 node and 34 node feeders separately.

4.3.1.1 IEEE 13 Node Feeder

The number of tap changes in the on load tap changer for each day was simulated for a whole year. Table 6 shows the percentage prediction error with respect to one minute aggregation interval.

Table 6: Percentage difference from one minute data

Sample Rate(min)	Average No. of Tap Changes per Month	% diff from 1min
1	1438.3	
5	1126.9	-21.65
10	1006.7	-30.01
15	959.1	-33.32
20	950.6	-33.91
30	896.7	-37.66
40	858.7	-40.30
45	826.4	-42.54
60	739.2	-48.61

This results were analyzed using Design of Experiment (DoE) with two factors at multilevel. Factor A represents the sample rate (SR) or demand interval and factor B represents individual month of a year. Figure 30 shows the Half-Normal probability plot.

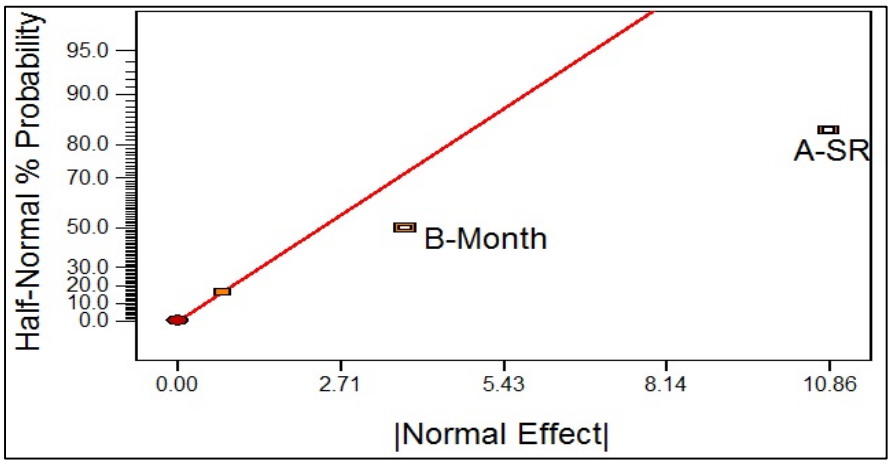


Figure 30: Half-normal plot for tap change analysis for 13 node feeder

From Figure 30 it can be inferred that months of a year (factor B) have negligible effect on number of tap changes whereas the demand interval (factor A) has a significant effect on number of tap changing. The results can be further analyzed using ANOVA table (Table 7).

Table 7: ANOVA table for whole year (13 bus voltage drop model)

Source	Sum of Squares	df	Mean Square	F Value	p-value Prob > F	
Model	1.189E+006	9	1.321E+005	17.37	< 0.0001	significant
A-SR	*9.998E+005*	*4*	*2.499E+005*	*32.88*	*< 0.0001*	
B-Month	*1.889E+005*	*5*	*37778.43*	*4.97*	*0.0041*	
Residual	1.520E+005	20	7601.58			
Cor Total	1.341E+006	29				

From the Table 7, since the *p*-value is less than 0.05, the model is significant. F value associated with the model is the ratio of the Model MS / Residual MS and shows the relative contribution of the model variance to the residual variance. A large number indicates more of the variance being explained by the model; a small number says the variance may be more due to noise. Hence from Table 7 it can be inferred that both aggregation interval and months of a year contributed to the model but the demand interval has much larger F value than individual months of a year. Hence months of a year have negligible effect on the experimented response (i.e. predicted error in tap change). The following relationship is developed using demand interval only. The estimation error in tap changes in terms of the demand interval is modeled as

$$\varepsilon_{VD}^{13N} = -10.9\ln(\tau) - 2.094 \tag{22}$$

The error in estimation is plotted against the data aggregation interval in Figure 31.

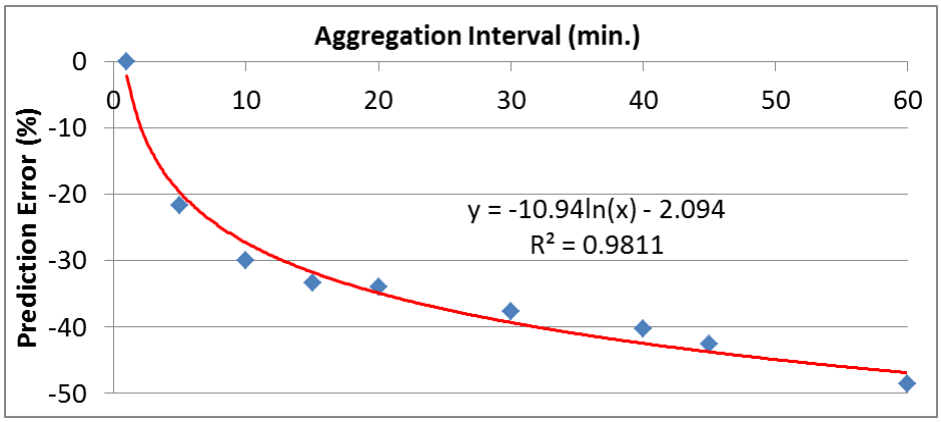


Figure 31: Voltage drop model for the 13 node system

4.3.1.2 IEEE 34 Node Feeder

Similar analysis was performed in IEEE 34 test feeder system. The number of tap changes in the on load tap changer for each day was monitored for a whole year. Half-normal plot for the 34 node system is given in Figure. 32.

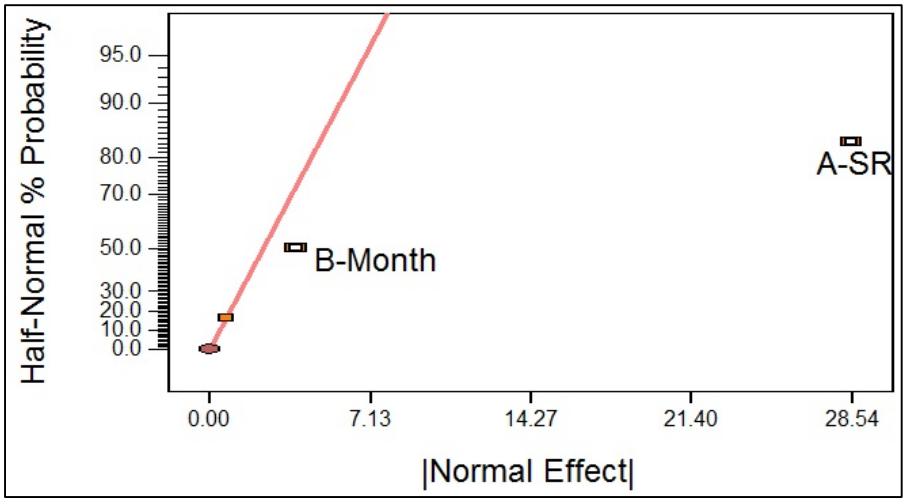


Figure 32: Half-normal plot for tap change analysis for 34 node feeder

ANOVA table for the 34 bus voltage drop analysis is shown in Table 8.

Table 8: ANOVA Table for whole year (34 bus voltage drop model)

Source	Sum of Squares	df	Mean Square	F Value	p-value Prob > F	
Model	2.454E+007	9	2.727E+006	187.21	< 0.0001	significant
A-SR	*2.417E+007*	*4*	*6.042E+006*	*414.89*	*< 0.0001*	
B-Month	*3.688E+005*	*5*	*73756.21*	*5.06*	*0.0037*	
Residual	2.913E+005	20	14564.01			
Cor Total	2.483E+007	29				

Both Half-Normal plot and ANOVA table shows months of a year (factor B) has negligible effect on the response signal. The estimation error in tap changes in terms of the demand interval is modeled as

$$\varepsilon_{VD}^{34N} = -14.3\ln(\tau) - 6.253 \tag{23}$$

The error in estimation is plotted against the data aggregation interval in Figure 33.

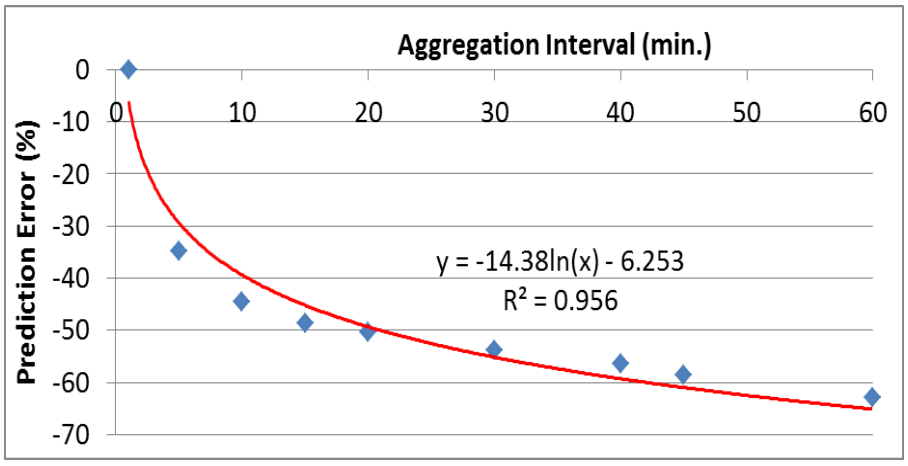


Figure 33: Voltage drop model for the 34 node system

4.3.2 Line-loss Analysis

Similar to voltage drop analysis DoE was used to determine the line-loss predicted error modeling. Half normal plot for 13 and 34 node systems are given in Figure 34 and 35. From Figure 34 and 35 and Table 9, 10 it can be inferred that the month of the year has higher contribution to the model than the aggregation interval when a model is developed for the year.

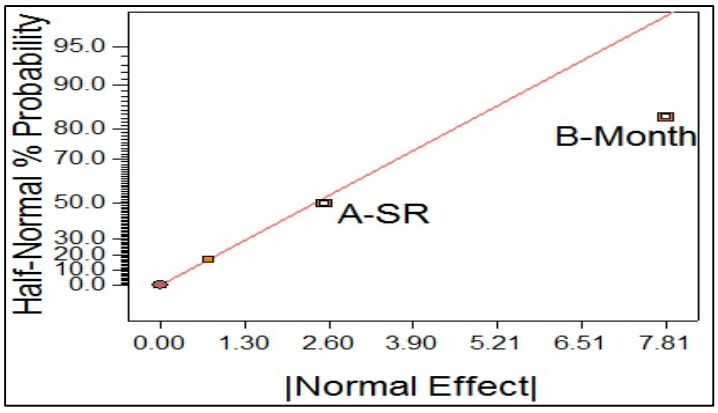


Figure 34: Half-normal plot for line loss analysis for 13 node feeder

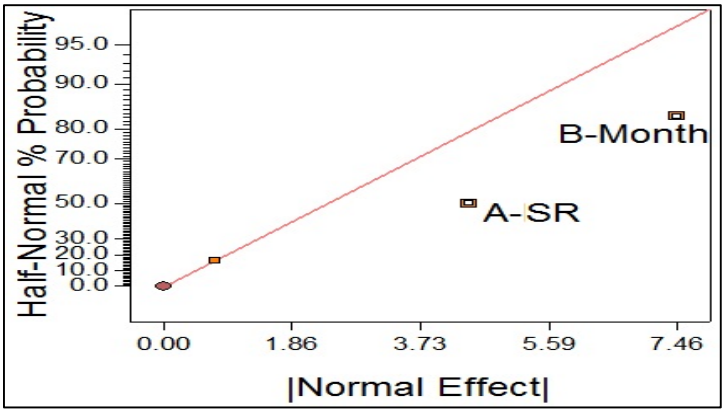


Figure 35: Half-normal plot for line loss analysis for 34 node feeder

Table 9: ANOVA table for whole year (13 node line-loss model)

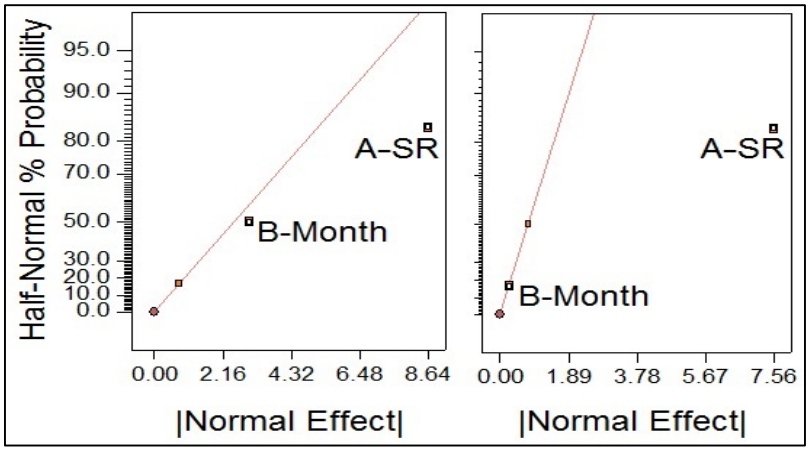
Source	Sum of Squares	df	Mean Square	F Value	p-value Prob > F	
Model	954.23	9	106.03	9.89	< 0.0001	significant
A-DI	*139.29*	*4*	*34.82*	*3.25*	*0.0331*	
B-Month	*814.94*	*5*	*162.99*	*15.20*	*< 0.0001*	
Residual	214.47	20	10.72			
Cor Total	1168.69	29				

Table 10: ANOVA table for whole year (34 node line-loss model)

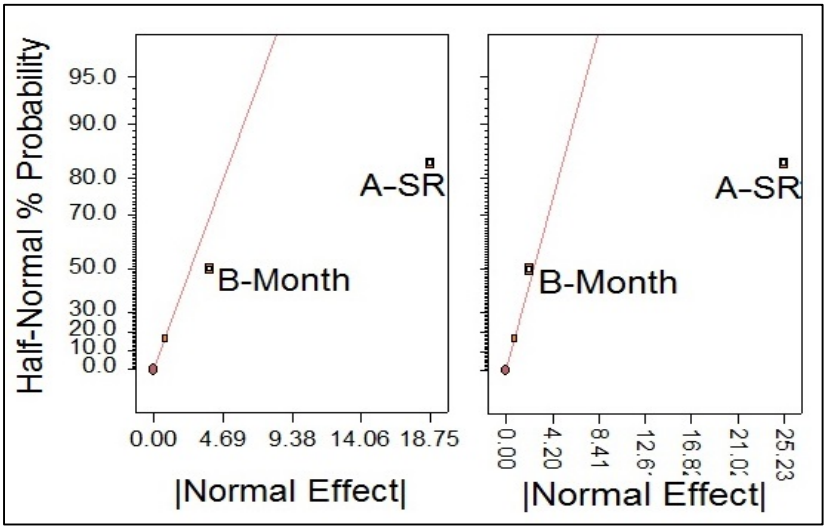
Source	Sum of Squares	df	Mean Square	F Value	p-value Prob > F	
Model	395.28	9	43.92	11.00	< 0.0001	significant
A-DI	*114.52*	*4*	*28.63*	*7.17*	*0.0009*	
B-Month	*280.76*	*5*	*56.15*	*14.07*	*< 0.0001*	
Residual	79.84	20	3.99			
Cor Total	475.12	29				

4.3.2.1 IEEE 13 Node Feeder

Half-normal plot of power-loss prediction error for each season for 13 node system is shown in Figure 36.



(a) Spring (b) Summer



(c) Fall (d) Winter

Figure 36: Half-normal plot for line loss analysis for 13 node feeder for four season

Based on the half normal plot for the four seasons in Figure 36, the individual month in a season has negligible effect on the response signal, therefore estimates of power loss error for demand intervals were developed and shown in the following equations.

$$\text{Spring}: \varepsilon_{PL,Sp}^{13N} = 0.722\tau^3 - 10.56\tau^2 + 53.77\tau - 42.18 \quad (24)$$

$$\text{Summer}: \varepsilon_{PL,F}^{13N} = 0.063\tau^3 - 1.610\tau^2 + 20.64\tau - 20.50 \quad (25)$$

$$\text{Fall}: \varepsilon_{PL,Su}^{13N} = -0.302\tau^3 + 4.471\tau^2 - 8.417\tau + 5.287 \quad (26)$$

$$\text{Winter}: \varepsilon_{PL,W}^{13N} = 0.453\tau^3 - 5.138\tau^2 + 24.9\tau - 18.58 \quad (27)$$

The error estimates for the different seasons are plotted in Figure. 37.

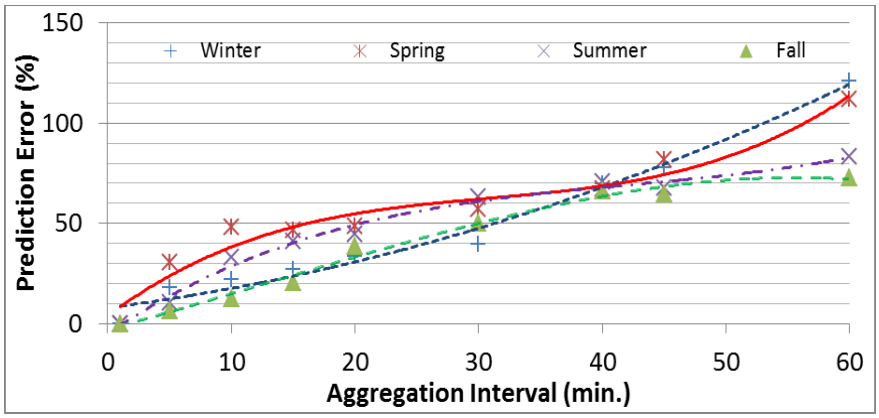
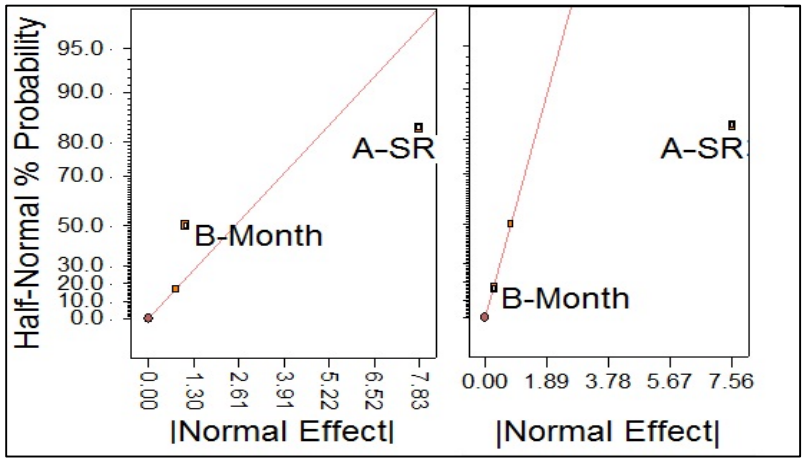


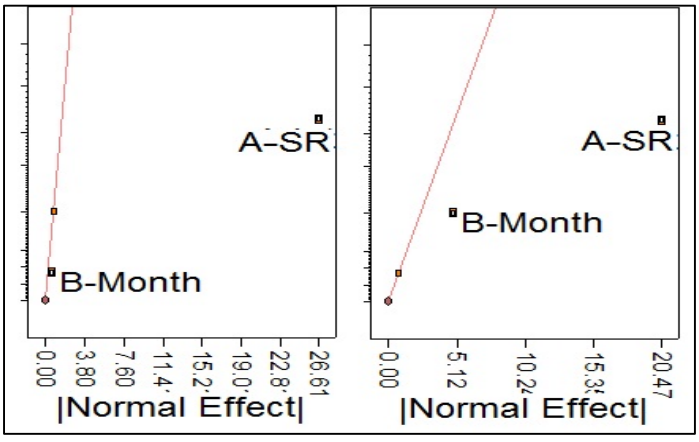
Figure 37: Line-loss error estimation for 13 node system for individual seasons

4.3.2.2 IEEE 34 Node Feeder

Half-normal plot of power-loss prediction error for each season for 34 node system is shown in Figure 38.



(a) Spring (b) Summer



(c) Fall (d) Winter

Figure 38: Half-normal plot for line loss analysis for 34 node feeder for four season

Based on the half normal plot for the four seasons in Figure 38, the individual month in a season has negligible effect on the response signal, therefore estimations of power loss error for demand intervals were developed and shown in the following equations.

$$\text{Spring}: \varepsilon_{PL,Sp}^{34N} = 0.004\tau^2 + 0.143\tau + 0.909 \tag{28}$$

$$\text{Summer}: \varepsilon_{PL,F}^{34N} = 0.004\tau^2 + 1.120\tau + 0.434 \tag{29}$$

$$\text{Fall}: \varepsilon_{PL,Su}^{34N} = 0.001\tau^2 + 0.444\tau - 2.083 \tag{30}$$

$$\text{Winter}: \varepsilon_{PL,W}^{34N} = 0.004\tau^2 + 0.198\tau + 0.195 \tag{31}$$

The error is estimation for the different seasons are plotted in Figure 39.

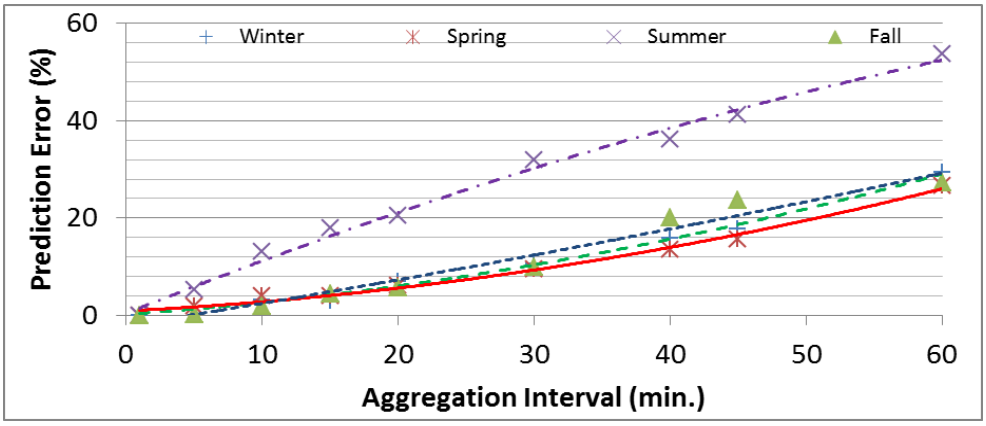


Figure 39: Line-loss error estimation for 34 node system for individual seasons

5 Conclusion and Future Work

A forward dynamic programming based algorithm is developed in order to analyze the impact of different control frequencies i.e. 15, 30 and 60 minute on both the consumer and the utility. The three control frequencies were compared to summarize tradeoffs when moving from one control frequency to the other. Each control frequency was also analyzed for its limits in terms of number of steps that can be taken while changing the thermostat settings.

Although the benefits of choosing more steps for the respective control frequency were there, the specific cases where 30 and 60 minute control did better than 15 minute control frequency (like percentage reduction in sustained violation and total duration of violation), there wasn't any benefit in choosing higher number of steps. Thus, it can be concluded that it is always feasible to choose lower number of steps and lower control frequency as long as one is dealing with the analyzed parameters.

A major factor that comes in the way of day-ahead scheduling is the price variations. However, if the constraint signal from the utility is sent in order to match a certain profile and thus promising the price signal, the variations in price signal can be reduced as is obvious from the peak reduction and violation energy reduction results. This will also help in forecasting the price signal more accurately. Moreover, while the reduction in peak demand would help in increasing life expectancy of transformers, the reduction in energy usage at the time of violation will help in reducing reserve capacity requirements.

One major concern while using dynamic programing is the number of possible combinations (states) per stage which increases with the number of users as was shown in Table III. For instance, for 10 units to be scheduled in a system there are 59,000 states. In worst case, where all the consumers are willing to let the thermostat deviate at maximum from the preferred thermostat settings, the system will have to solve the scheduling problem with most number of states. Thus a more constrained system poses computationally less burden on the system. Also, the number of states to be saved per stage gives reasonable results even when a small portion of maximum possible states is saved.

Active consumer participation through the smart grid initiative requires more demand information from the consumers. Requirement of large data creates additional burden in terms of larger bandwidth requirement for communication and data storage. One of the solutions to this problem is aggregating the data at the consumer level and forwarding the aggregated data. This is further expected to reduce the privacy concerns from residential consumers. However the aggregated data loses the granule information which is vital for accurate load forecasting and managing. This work analyzed the impact of data aggregation interval on the error of parameter prediction. Load tap changer operation estimation and line-loss estimation were used as two example applications in this work. Designs of Experiments were used in this work to determine the significance of contributing factors for a particular application or output. This analysis was performed with different houses generated with different load shapes (using statistical modeling) for each day for a year in order to simulate the actual feeder behavior. Based on the results, it is evident that prediction error can't be generalized for any distribution network, but could be useful for given distribution network. For each network such a model needs to be developed. The results for different bus systems shows that change in voltage in the distribution system is less prone to the individual month, while total power loss of the circuit is prone to individual months of a year but less prone within months of a season. This work presented the initial bench mark for quantifying the loss of important detail when a longer aggregation interval is used. The outcomes could be used for evaluating the impact of a certain aggregation interval on the distribution system parameters.

5.1 Future Work

As a future work, the demand level impact algorithm can be improved to serve as an on demand scheduler i.e. responding to instantaneous needs in order to achieve peak reduction. Currently, the algorithm is only able solve in one go; by going back and choosing different paths when the solution is not found can help in finding better solution. Furthermore, analyses of a complete system, i.e. with heating system as well as electric vehicles can be done as well, to realize the impact throughout the year for an entire feeder serving many transformers. The impact on transformer and other equipment's life expectancy can also be analyzed.

One of the limitations of this work is that the P_{max} constraint signal chosen for each control signal is same and naturally, unable to see majority of the violations for higher control frequencies. This can be taken as a future work to enforce different constraint signals for each control frequency and then analyze the system again. Also, the analysis can be further improved by choosing respective range parameters for the ETP model which would definitely require a faster algorithm as the range for each parameter will be varied and will need more iteration to conclude the results statistically.

As a future direction a more general relationship of demand interval for combined applications should be modeled. An optimal aggregation interval needs to be evaluated to support active consumer participation.

References

[1] P. Palensky, D. Dietrich, "Demand Side Management: Demand Response, Intelligent Energy Systems, and Smart Loads," *IEEE Trans. Industrial Informatics*, vol. 7, no.3, pp.381-388, August 2011

[2] ANSI/ASHRAE Std. 55-2013, "Thermal Environmental Conditions for Human Occupancy" https://www.ashrae.org/resources--publications/bookstore/standard-55.

[3] S. Rahimi, A. Chan, R. Goubran, "Usage Monitoring of Electrical Devices in a Smart Home," in Proc. *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, EMBC, pp.5307-5310, August 30 - September 3, 2011.

[4] http://tinycomb.com/wp-content/uploads/2009/05/smart-grid.jpg.

[5] C. Gellings, "The Concept of Demand-Side Management for Electric Utilities," *The Proceedings of the IEEE*, vol. 73, no. 10, October 1985.

[6] H. Lam, G. Fung, W. Lee, "A Novel Method to Construct Taxonomy Electrical Appliances Based on Load Signatures," *IEEE Trans. Consumer Electronics*, vol.53, no.2, pp. 653-660, May 2007.

[7] J. Martins, R. Lopes, C. Lima, E. Romero-Cadaval, and D. Vinnikov, "A novel nonintrusive load monitoring system based on the S-Transform," in Proc. *13th Int. Conf. on Optimization of Electrical and Electronic Equipment (OPTIM)*, 2012.

[8] A. Ruzzelli, C. Nicolas, A. Schoofs, G. O'Hare, "Real-Time Recognition and Profiling of Appliances through a Single Electricity Sensor," in Proc. *7th Annual IEEE Communications Society Conf. on Sensor Mesh and Ad Hoc Communications and Networks* (SECON), 2010.

[9] D. Alessandro, P. Laura, "An event driven Smart Home Controller Enabling Consumer Economic Saving and Automated Demand Side Management" *Applied Energy Journal*, vol. 96, pp. 92-103, August 2012.

[10] X. Gang, C. Chen, S. Kishore, A. Yener, "Smart (in-home) Power Scheduling for Demand Response on the Smart Grid," in Proc. *IEEE PES Innovative Smart Grid Technologies* (ISGT 2011), January 2011.

[11] D. Onur, F. Alberto, "Scheduling Energy Consumption with Local Renewable Micro-Generation and Dynamic Electricity Prices," in Proc. the *First Workshop on Green and Smart Embedded System Technology: Infrastructures, Methods and Tools*, 2010.

[12] L. Junghoon, P. Gyung-Leen, K. Sang-Wook, K. Hye-Jin, S. Chang-Oan, "Power Consumption Scheduling for Peak Load Reduction in Smart Grid Homes," in Proc. 2011 *ACM Symposium on Applied Computing*, pp. 584-588, 2011.

[13] A. Mohsenian-Rad, V. Wong, J. Jatskevich, R. Schober, "Optimal and Autonomous Incentive-Based Energy Consumption Scheduling Algorithm for Smart Grid," in Proc. *Innovative Smart Grid Technologies* (ISGT), January 2010.

[14] A. Mohsenian-Rad, A. Leon-Garcia, "Optimal Residential Load Control With Price Prediction in Real-Time Electricity Pricing Environments," *IEEE Trans. Smart Grid*, vol. 1, no.2, pp.120-133, September 2010.

[15] A. Mohsenian-Rad, V. Wong, J. Jatskevich, R. Schober, A. Leon-Garcia, "Autonomous Demand-Side Management Based on Game-Theoretic Energy Consumption Scheduling for the Future Smart Grid," *IEEE Tran. Smart Grid*, vol.1, no.3, 2010.

[16] K. Hung, J. Song, H. Zhu "Demand side management to reduce Peak-to-Average Ratio using Game Theory in Smart Grid," in Proc. *IEEE Conference Computer Communications Workshops (INFOCOM WKSHPS)*, pp. 91-96, March 2012.

[17] A. Subramanian, M. Garcia, A. Dominguez-Garcia, D. Callaway, K. Poolla, P. Varaiya, "Real-time Scheduling of Deferrable Electric Loads," in Proc. *American Control Conference* (*ACC*), pp. 3643-3650, June 2012.

[18] D. Pengwei, L. Ning, "Appliance Commitment for Household Load Scheduling," *IEEE Trans. Smart Grid*, vol.2, no.2, pp. 411-419, June 2011.

[19] G. Karmakar, A. Kabra, K. Ramamritham, "Coordinated scheduling of Thermostatically Controlled Real-Time Systems under Peak Power Constraint," in Proc. 19th *IEEE Real-Time and Embedded Technology and Applications Symposium* (*RTAS*), pp. 33-42, 2013.

[20] Z. Wei, L. Jianming, C. Chin-Yao, K. Kalsi, "Aggregated Modeling and Control of Air Conditioning Loads for Demand Response," *IEEE Trans. Power Systems*, vol. 28, no. 4, pp.4655-4664, November 2013.

[21] J. Dang, R. Harley, "Air Conditioner Optimal Scheduling Using Best Response Techniques," in Proc. *IEEE Power Engineering Society Conference and Exposition in Africa* (*PowerAfrica*), July 2012.

[22] F. Shariatzadeh, A. Srivastava "Look-Ahead Control Approach for Thermostatic Electric Load in Distribution System," in Proc. *North American Power Symposium*, September 2013.

[23] C. Chi-Min, J. Tai-Lang, "A Novel Direct Air-Conditioning Load Control Method," *IEEE Trans. Power Systems*, vol. 23, no. 3, pp. 1356-1363, August 2008.

[24] G. Koutitas and L. Tassiulas, "Smart Grid Technologies for Future Radio and Data Center Networks," *IEEE Communication Magazine*, vol. 52, issue 4, April 2014.

[25] Y. Ozturk, D. Senthilkumar, S. Kumar, amd , and G. Lee "Senior Member, IEEE An Intelligent Home Energy Management System to Improve Demand Response," *IEEE Trans. Smart Grid*, vol. 4, no. 2, 2013.

[26] P. Mancarella, and G. Chicco, "Real-Time Demand Response From Energy Shifting in Distributed Multi-Generation," *IEEE Trans. Smart Grid*, vol. 4, no. 4, December 2013.

[27] J. Wang, S. Kennedy and J. Kirtley Jr. "Optimization of Forward Electricity Markets Considering Wind Generation and Demand Response," *IEEE Trans. Smart Grids*, vol. 5, no. 3, May 2014.

[28] L.L. Grigsby. The Electric Power Engineering Handbook. CRC Press and IEEE Press.

[29] S. Maharjan, Q. Zhu, Y. Zhang, S. Gjessing, and T. Başar, "Dependable Demand Response Management in the Smart Grid: A Stackelberg Game Approach," *IEEE Trans. Smart Grid*, vol. 4, no. 1, March 2013.

[30] E. Romero, Voltage Control in a Medium Voltage System with Distributed Wind Power Generation, Department of Industrial Electrical Engineering and Automation, Lund University.

[31] H. Markiewicz, A. Klajn, Voltage Disturbances Standard EN 50160 - Voltage Characteristics in Public Distribution Systems, Wroclaw University of Technology.

[32] J. Gruber, M. Prodanovic, "Residential Energy Load Profile Generation Using a Probabilistic Approach," in Proc. *Sixth UKSim/AMSS European Symposium on Computer Modeling and Simulation*, pp. 317-322, November 2012.

[33] V. Jukka, D. Peter, "A Model for Generating Household Electricity Load Profiles," *International Journal of Energy Research*, vol. 30, issue 5, pp. 273-290, April 2006.

[34] Z. Taylor, K. Gowri, S. Katipamula, "GridLAB-D Technical Support Document: Residential End-Use Module Version 1.0", http://www.pnl.gov/main/publications/external/technical_reports/PNNL-17694.pdf".

[35] www.ia.omron.comsupportglossarymeaning3205.html.

[36] Residential Load Calculations Manual" Eighth Edition, Air Conditioning Contractors of America, http://www.energystar.gov/ia/partners/bldrs_lenders_raters/downloads/Outdoor_Design_Conditions_508.pdf.

[37] https://www.powermonitors.com/product/detail/eagle-120.

[38] Price signal. Available: https://rrtp.comed.com/live-prices/.

[39] Weather data. Available: http://www.isws.illinois.edu/.

[40] W. H. Kersting, "Distribution System Modeling and Analysis," CRC Press, Boca Raton, Florida, 2002.

[41] Richardson, I. and Thomson, M., One-Minute Resolution Domestic Electricity Use Data, 2008-2009 [computer file]. Colchester, Essex: UK Data Archive [distributor], October 2010. SN: 6583, http://dx.doi.org/10.5255/UKDA-SN-6583-1.

[42] A. Malekpour, and A. Pahwa, "Reactive Power and Voltage Control in Distribution Systems with Photovoltaic Generation," in Proc. *North American Power Symposium* (NAPS), September 2012.

[43] Design of Experiments Tutorial, American Society of Quality, available online: http://asq.org/learn-about-quality/data-collection-analysis-tools/overview/design-of-experiments-tutorial.html.

[44] Engineering Statistics Handbook, NIST, available online: http://www.itl.nist.gov/div898/handbook/index.htm.

Part III

Preserving Privacy of Advanced Metering Data using Efficient Aggregation and Prediction Techniques

Murtuza Jadliwala and Vinod Namboodiri

Arash Boustani and Anindya Maiti, PhD Students
Navid Alamatsaz and Zoya Khan, MS Students

Wichita State University

For information about Part III, contact:

Dr. Murtuza Jadliwala
Electrical Engineering and Computer Science Department
Wichita State University, Campus box – 83
Wichita, KS – 67260
Phone: 316-978-3729
Email: murtuza.jadliwala@wichita.edu

Power Systems Engineering Research Center

The Power Systems Engineering Research Center (PSERC) is a multi-university Center conducting research on challenges facing the electric power industry and educating the next generation of power engineers. More information about PSERC can be found at the Center's website: http://www.pserc.org.

For additional information, contact:

Power Systems Engineering Research Center
Arizona State University
527 Engineering Research Center
Tempe, Arizona 85287-5706
Phone: 480-965-1643
Fax: 480-965-0745

Notice Concerning Copyright Material

Table of Contents

List of Figures

List of Tables

1. Introduction

1.1 AgSec: Secure and Efficient CDMA-based Aggregation for Smart Metering Systems

Millions of people suffered from the biggest blackout in North American history in 2003. Investigations showed that the outage was because of lack of real-time monitoring and diagnosis and failure in proper load balancing [1]. Recently smart grid has been proposed as the next generation power grid. A smart grid is an electrical grid that utilizes communication technologies and information processing to collect process and act on gathered information in order to improve reliability, efficiency, economics and sustainability of the power grid [2]. This will help the utility companies to act on consumer information gathered from smart meters (SM) at the user's premises. The two-way communication capability will enable functions such as demand-response, demand-dispatch, self-monitoring, and self-diagnosis for the existing power grid [3]. It also promises reduced prices through dynamic pricing schemes, wide penetration of renewable resources such as wind and solar, and fewer power outages [4].

Smart grid researchers have been studying miscellaneous problems such as communication technologies and infrastructure [5]-[9], legal and policy concerns [10], [11], reliability, failure diagnosis and recovery [12]-[14], demand-response, load-shaping and peak-shaving [15]-[17], data aggregation [5], [18]-[20] and, last but not the least, security and privacy [3]-[5], [21]-[23]. Having access to fine-grained usage data reveals serious potential security and privacy threats to the users. For instance, it can be easily determined if a residential house is vacant or not by observing the fine-grained energy consumption patterns [24]. It is also possible to track the location of the residents of a house based on the appliance they are using [25]. Insurance companies can monitor and track eating, sleeping and possibly exercise habits of a household [26], [27]. In 2009, the Dutch Parliament prohibited the utilization of smart meters because of privacy issues. There are also many cyber security related challenges for the deployment of the smart grid [5]. The concept of smart grid is about "moving from a relatively small number of carefully controlled devices to an Internet-like distributed environment". This "Internet-like distributed environment" is vulnerable to many known and unknown cyber security attacks [28]. The security threats to the smart grid can target the confidentiality and the integrity of the gathered fine-grained user data. They can also threaten the availability of the power grid. Computerworld [29] reports more than 170 outages caused by cyber-security attacks. It goes without saying that without proper security and privacy-preserving mechanisms, large scale deployment and proliferation of the smart grid is difficult. Earlier security approaches have primarily used cryptographic techniques such as homomorphic encryption and secure multiparty computation in order to preserve user privacy while aggregating usage data [30]. These approaches, although providing strong

guarantees of confidentiality, are very heavy from a computational and communicational stand-point and may not be feasible on low-end smart meters with limited computation capabilities. Homomorphic cryptosystems usually generate an output of a huge fixed-length compared with the data generated by smart meters. This ciphertext can be up to one hundred times larger than the actual smart metering data [5]. Given the frequency of the data being sent and possible bandwidth limitations, this can lead to unacceptable delay and network overhead.

In this part of the project, we investigate the feasibility and efficiency of existing privacy-preserving data aggregation approaches. We devise a new efficient and computationally feasible secure data aggregation technique for smart meters using properties of spread spectrum communication technology. Details for this part of the project are organized as follows: Background and Related work on existing secure aggregation schemes for smart grid is outlined in Chapter 2. The network and adversary model assumed in this work is presented in Chapter 3. The proposed AgSec secure aggregation protocol is outlined in Chapter 4, and mathematical evaluation and results are discussed in Chapter 5.

1.2 Seer Grid: Privacy and Utility Implications of Two-Level Energy Load Prediction in Smart Grids

As part of the future smart electricity grid initiative, a smart grid communication network (SGN) is a large-scale integration of information and communication technologies within the electricity generation, transmission, and distribution systems of the traditional electricity grid. A combination of various smart technologies at different levels of the SGN promotes efficiency, reliability and stability in operations of the smart grid. One indispensable piece of technology in a SGN is a smart meter (SM) which collects and periodically reports the energy usage or consumption information of the customers to the electric (a.k.a. utility) company (EC), which in turn facilitates highly efficient energy generation and distribution and helps the EC to cope with changes in energy demand and supply. The monetary and natural resource savings due to the improved efficiency is a major factor in the fast growing adoption of SMs, with predictions that 800 million SMs will be in use globally by 2020 [48]. Despite their tremendous importance in a SGN, SMs can also be easily exploited by malicious adversaries (including the EC) who may attempt to infer private customer information from reported energy consumption patterns, such as occupancy of the house [49], specific appliances being used [50], and even daily routine of the residents [51] [52].

Various techniques for overcoming privacy issues due to the energy usage information generated and shared by SMs have been proposed in the research literature, and these solutions have primarily followed one of the following two approaches: (i) completely obscure the individual SM data from the perceived adversary, or (ii) hide privacy-sensitive signatures or patterns from the individual SM data by perturbation or down-sampling. In the first direction, protocols that take advantage of the homomorphic properties of public-

key cryptographic algorithms to perform neighborhood-level aggregation of SM data have been proposed in the literature [53]–[54]. These protocols enable the EC to learn the actual aggregated energy consumption information (at a neighborhood level) without leaking individual customer-specific information to the aggregator. In the second direction, many approaches have been proposed to efficiently perturb energy consumption data in order to meet certain privacy requirements. In-residence storage batteries have been employed to flatten or mask variances in the load or electricity usage information [56], [57]. Similarly, controlled perturbation [58]–[60] and down-sampling [61] of the energy consumption data to mask specific signal or appliance signatures have also been attempted. But as pointed out by [60], [61], the degree of correlation between the actual energy consumption and the data output by a privacy-preserving technique typically characterizes a tradeoff between privacy and utility (or usefulness). Higher correlation with the actual ground-truth makes the perturbed data more useful but reveals private information, whereas lower correlation (or increased perturbation) is good for privacy but reduces data usefulness or utility. As protocols following the first approach do not really perturb the electricity consumption data, the utility of the data (or any function computed from the data) is high. Also, as this data is cryptographically obscured from the aggregator, there is no leakage of private customer information. However, protocols using public-key cryptography are non-trivial to implement in practice and have very high computation and communication overhead [62]. Perturbation mechanisms, such as the ones using storage batteries [56], [57], are effective in masking private usage patterns but at the cost of drastically reducing the utility of the data. Moreover, installing and maintaining large capacity batteries in every household have also shown to be economically non-viable [63]. Similarly, [61] show that performance of smart grid operations can degrade due to reduction in sampling frequency.

Other perturbation mechanisms, such as, [60], that attempt to strike a good balance between privacy and data-utility by masking or suppressing specific appliance signatures assume that individual appliance electricity consumption information is readily available (or can be easily separated from the overall data) which may not always be feasible. Given the above state of- the-art, we feel that both data hiding and data perturbation approaches have inherent limitations, which motivates us to explore alternate paradigms (beyond hiding and perturbation).

Our goal in this part of the project is to explore alternate practical designs for privacy-sensitive generation and sharing of energy consumption information from the SMs to the EC which enables effective operation of the EC in terms of accurately predicting future demand and electricity generation and distribution. In order to achieve this goal, we move away from the classical perturbation/data-hiding techniques and use learning-based prediction mechanisms to generate (or predict) energy consumption patterns shared by SMs. Our prediction mechanism will replace variances in the individual household-level actual energy consumption patterns (which is typically indicative of loads) with relatively

smoother patterns that are free of load signatures but accurate enough to be useful in predicting energy consumption at the neighborhood level (which is the one that is actually used by the EC).

Due to this privacy-sensitive inference attacks will be much harder on the household-level data shared by the SM without significantly impacting the demand-response and electricity generation/distribution calculations at the EC. With Seer Grid, as our future work, we are going to propose a household-level prediction scheme comprising of a statistical learning algorithm (trained using past consumption pattern of the household) which predicts an entire day's electricity consumption pattern a day in advance. This prediction can be performed locally on the SM, on a local energy management unit or on a computing device that connects to such a unit. The household electricity consumption pattern predicted locally at the SM, with the load or appliance signatures masked or flattened, is then reported to an aggregator or data concentrator (referred here as a cluster head or CH) at the beginning of each day. All SMs within a neighborhood or cluster report their energy consumption predictions to their respective CH who in turn forwards an aggregated prediction (as described below) to the EC. As our localized prediction flattens or eliminates sharp variations (which may indicate load signatures) in the predicted consumption at the SM or household level, this difference can add up significantly while aggregating predictions for multiple households in a neighborhood or a cluster. This can reduce the accuracy of the aggregated prediction, thereby adversely impacting its utility or usefulness to the EC. In order to restore this utility lost due to prediction at the SM level, we introduce a second level of energy load prediction at the CH for compensating the difference in the aggregate of predicted and actual energy usage of individual SMs in the cluster. CH performs the load prediction based on past energy consumption pattern of the entire neighborhood or cluster, and reports the result of the second level prediction to EC just before beginning of each day. EC can then use this cluster or neighborhood wide load prediction to efficiently control electricity generation and distribution. To further improve efficiency and ensure fail proof operation of the SGN, we also incorporate real-time and privacy-preserving reporting of the aggregated variance between actual and predicted energy consumption of all SMs in the cluster.

We would like readers to note that, the Seer Grid's two level prediction mechanism offers several advantages over traditional privacy-preserving energy data reporting schemes in the literature. Unlike data hiding schemes that require multiple (one per each generated data value) encryption operations at the SM or household level, our prediction and reporting operation is performed just once (per day). Moreover, Seer Grid is communication-efficient (as no additional data or overhead needs to be communicated), does not require any specialized hardware (e.g., storage batteries) and does not need access to appliance-level consumption patterns. The contributions for this part of the project are organized as follows: In Chapter 6, we discuss the selection of statistical learning algorithms suitable for prediction in our SGN, followed by details of our SGN

architecture and its operation in Chapter 7. In Chapter 8, we evaluate the proposed Seer Grid architecture using real smart meter data by performing extensive experimental simulations. We empirically measure the correlation between predicted and actual consumption patterns at each level, using standardized metrics. Evaluation results strongly support our proposition of a practical SGN architecture which maximizes both privacy and utility of smart meters. Concluding remarks for both part of the projects are outlined in Chapter 9.

2. Background and Related Work

Below we outline existing cryptographic approaches to private data aggregation in Smart Grid Networks (SGN) and also study some data aggregation methods in other networking infrastructures with similar constraints such as WSNs.

2.1 Related Work on Secure Data Aggregation Mechanisms

Cryptographic schemes, especially encryption algorithms with homomorphic properties, have been adopted as a popular tool to achieve secure data aggregation in a variety of networking and communication systems. A public-key cryptosystem is known to have homomorphic properties if $E(m_1 \Diamond m_2) = E(m_1) \Delta E(m_2)$, where E is the encryption function, and $\Diamond$ and Δ are two different mathematical operations. Based on the supported operations, homomorphic cryptosystems fall into two broad categories: partially homomorphic and fully homomorphic. Partially homomorphic cryptosystems only support either addition or multiplication or in some cases polynomials up to certain degrees, whereas fully homomorphic cryptosystems support both addition and multiplication [5], [23]. We refer the readers to [31]-[34] for more details on homomorphic cryptosystems.

In SGNs, the utility companies are interested in statistics such as total consumption for billing in a specific time period [5]. Given that sum of consumed electricity of all smart meters in a residential neighborhood is of interest to the UC, homomorphic properties of the Paillier [34] encryption can be useful. Rather than adding the consumption data in plaintext, one can multiply the encrypted values and then decrypt the result to get the addition of plaintext data.

He et al. [23] present a secure data exchange scheme for the smart grid based on homomorphic properties of Goh cryptosystem [35]. Goh supports an arbitrary number of additions and a single multiplication on the ciphertext. It is worth noting that the aforementioned protocol is only a secure data communication scheme without any aggregation capabilities. Li et al. [18] utilize the homomorphic properties of Paillier to propose an incremental data aggregation scheme. In [18] every node passes its encrypted consumption data to its parent node on the aggregation tree. The parent node multiplies the received value into its own encrypted consumption data and passes the total result to the next parent node. Therefore, all the meters participate in the aggregation, without seeing any intermediate or final result. Garcia and Jacobs [36] present a privacy-preserving protocol using Paillier based on secret sharing. Their proposal hides consumption data from the electricity or utility company (UC) as it receives random shares of data which it cannot decrypt. The other nodes cannot retrieve meaningful information either since they only receive random shares. Kursawe et al. [37] propose two approaches to calculate total consumption in SGN. In their first approach, called *aggregation protocols*, smart metering data are masked in such a way that after summing the data from all smart meters masking values cancel each other out and the UC gets the

total consumption information. In their second approach, named *comparison protocols*, they consider that the UC roughly knows the total consumption. Erkin and Tsudik [38] propose a cryptographic protocol based on a modified version of the Paillier cryptosystem to calculate the total consumption of all the SMs in a given neighborhood as well as a single SM in an Advanced Metering Infrastructure (AMI). Acs and Castelluccia [39] suggest a solution using masking and differential privacy and utilizing the homomorphic properties of a computationally-cheap cryptosystem for private data aggregation. Lu et al. [40] propose an *Efficient and Privacy-Preserving Aggregation (EPPA)* for smart grid communications by structuring multidimensional data and encrypting them with the Paillier cryptosystem. Erkin et al. [5] study different existing secure signal processing mechanisms in SGNs and compare different existing cryptographic methods in terms of computational complexity, efficiency and imposed overhead.

He et al. [41] and Li et al. [42] propose similar integrity preserving data aggregation schemes, *iPDA* and *EEHA* respectively, for wireless sensor networks using the concept of data slicing and assembling. The authors propose three steps: i) constructing an aggregation tree using the well-known *LEACH* algorithm [43]. ii) Segmenting or slicing the data, and iii) merging the pieces of data at the aggregator and sending the merged data to the sink node. *iPDA* uses multiple aggregation trees, hence providing better integrity, by sending more than one copy of the data to the destination. However, transmitting more than one copy of the same data can cause extra communication overhead. Zanjani et al. [44], [45] propose a new energy-efficient aggregation mechanism for WSNs using the concepts of coding theory. The sensor nodes are assigned unique Orthogonal Chip Sequences (OCS) that are used to code and send their data on the CDMA channel. The authors claim that, utilizing *ESTOC*, data integrity can be protected while aggregating. Also, *ESTOC* reduces Bit Error Rate (BER) and interference caused by simultaneous transmission of nodes. Yan et al. [19] propose a secure in-network data aggregation scheme to aggregate the data from smart appliances inside a Home Area Network (HAN). Similar to *ESTOC* [44], the authors in this scheme utilize the properties of spread spectrum communications for efficient aggregation.

2.2 Main Motivation for AgSec

In the cryptographic approaches discussed in [5], [18], [23], [36]-[38], we observe that the power-usage information is generally of small size (e.g. 20 bits) [40], [3]. However, the plaintext input size of most existing homomorphic cryptosystems is huge [5], [40], for example 2048 bits for the widely-used Paillier cryptosystem [34], [36], [38], [40]. As a result, the input data has to be padded before encryption. Given the high frequency of data collection and the number of deployed smart meters, this will result in unacceptable communication overhead on the network, and also high processing burden on the smart meters with limited computational capabilities [40]. Aggregation schemes that construct and utilize the spanning-tree, for instance by Li et al. [18], also do not consider performance issues. The processing and communication overhead makes the protocol

less suitable in practical implementations. Moreover, depending on the depth of the spanning tree of the network, there can be large delays between the time power consumption data is reported by the meters and the time the aggregated data is received at the UC.

The aggregation schemes proposed in [41]-[45] do not consider any security issues. The main focus of the authors is increasing data integrity and energy efficiency in WSNs. Phulpin et al. [47] study the efficiency and benefits of network coding in both PLC and wireless SGNs. The authors also show that using coding theory in SGN reduces the delay by decreasing the number of time slots and saves energy through reducing the number of transmission.

We are proposing a secure aggregation scheme by using properties of spread spectrum communications and utilizing the slicing and assembling technique [41], [42] to efficiently aggregate energy usage while improving network performance and decreasing unnecessary computation load on smart meters. Our contention-free scheme will also decrease the delay, BER, and interference.

2.3 Related Work on Prediction Mechanisms

There have been multiple proposed schemes for load prediction at cluster level for short term [65]-[67] and long term [68]. Sevlian and Rajagopal [69] proposed short term electricity load forecasting on varying levels of aggregation, and concluded that aggregating more customers improves the relative forecasting performance only up to specific point. Recently, smart meter based short-term load forecasting was proposed [70] [71], as a household's historic energy consumption pattern is a better predictor of peak load than any other observable variables. In contrast, Seer Grid uses two level of prediction to retain the privacy benefits of aggregation, and utility benefits of individual household prediction.

There also have been extensive research efforts attempting to address SM privacy issues. Li et al. [53] proposed using Paillier's homomorphic encryption for distributed energy consumption data aggregation from SMs, where EC is able to know only the aggregated data upon decryption of the aggregated cipher. Garcia and Jacob [54] combined a secret sharing algorithm with Paillier's homomorphic cryptosystem to compute the aggregated energy consumption of a given set of users (for example, in a cluster), in a privacy preserving fashion. However, homomorphic cryptosystems induce a large computational overhead on the SMs, and real-time reporting in short time interval is impractical [62]. Alternatively, McLaughlin et al. [57] proposed a non-intrusive load leveling model using large capacity batteries. Large batteries smoothen the energy consumption pattern and effectively help in hiding signal signatures contained in actual consumption pattern. However, large batteries are economically inconvenient [63] due to their high capital cost and low energy efficiency.

Privacy through anonymization tries to unlink the energy usage data from individual SMs [72]. However, anonymization may turn out to be ineffective, as it is possible to

infer the data origin [73]. With the limited computational capabilities and practicality in mind, researchers suggested the use of perturbation techniques for hiding signal signatures. Consumer privacy can be preserved by deliberately introducing error into energy usage data [58] [59], and such perturbation techniques often try to achieve differential privacy in order to minimize the privacy-utility trade-off [60].

3. AgSec: Network Architecture

3.1 Network and Communication Model

We consider the widely-used wireless-wired architecture for the deployment of SGN. The wireless communication between smart meters, which are organized into clusters, and the aggregator or Cluster Head (CH) uses 802.15.4 or Zigbee due to characteristics such as low power, short delay, self-organization, scalability, and high security [8]. The aggregated data will be forwarded from the CH to the UC using a dedicated point-to-point wired link.

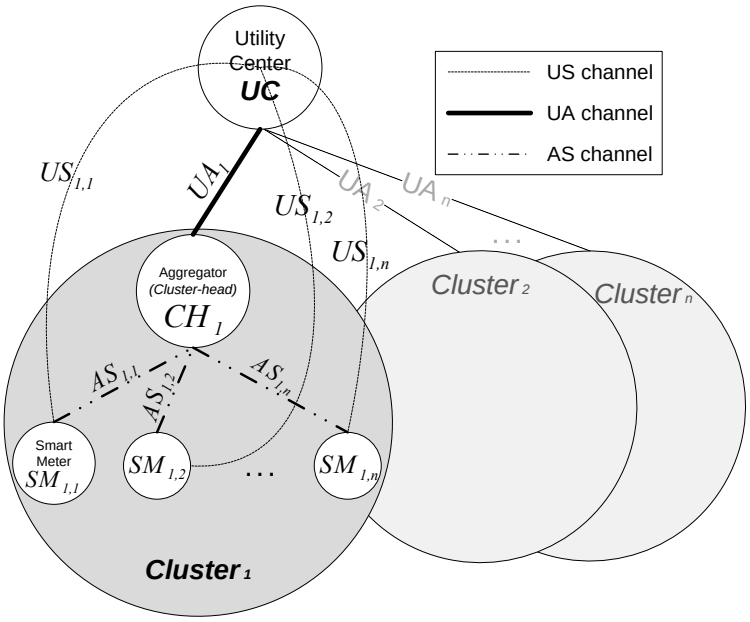


Figure 1. *Network Architecture for SGN*

Figure 1 depicts a three-level hierarchical network architecture. The communication between the UC and the i^{th} aggregator is denoted as UA_i. Similarly $AS_{i,j}$ represents the communication between the i^{th} aggregator and the j^{th} smart meter in the i^{th} cluster. The control and signaling messages between the UC and the j^{th} smart meter in the i^{th} cluster are exchanged on a channel referred to as $US_{i,j}$. The signaling messages, which are used in the initialization phase, are discussed in details in chapter 4. The Zigbee medium access protocol on all *AS* channels is CDMA. Also all *UA* communications are on a dedicated wired channel. Finally, our signaling channel is a high-range wireless WAN technology, such as GPRS, UMTS or LTE. Figure 2 illustrates the components implemented in different network entities.

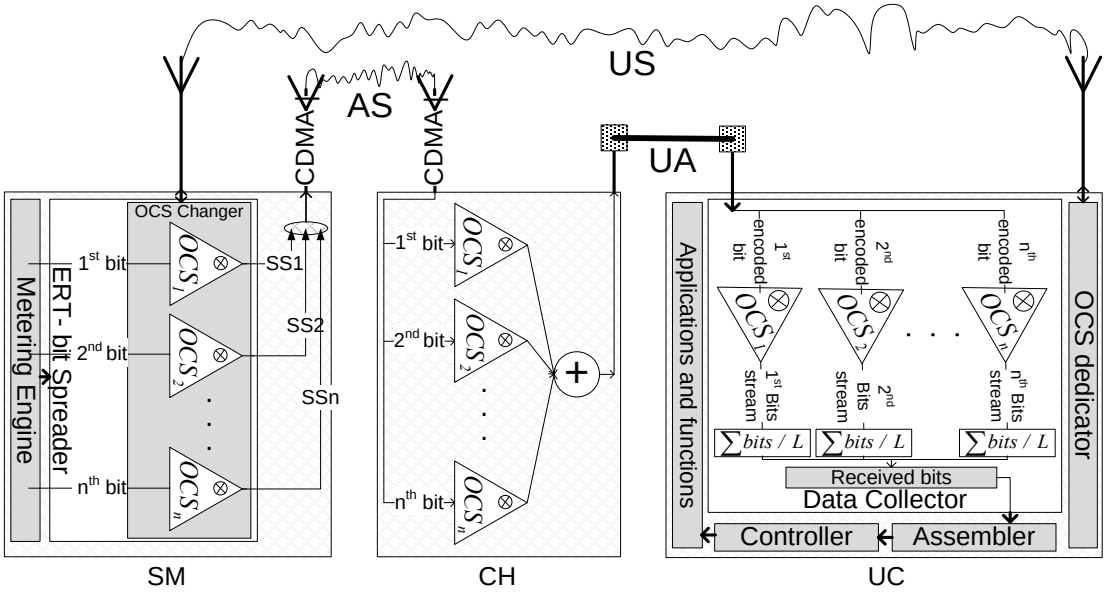


Figure 2. *Participating Entities in Secure Aggregation*

3.2 Communications on the CDMA Channel

All communication takes place over three separate channels as discussed in section 3.1. All smart meter data from the smart meter to the aggregator are sent over the CDMA-based data channel, represented as the AS channel (in Figure 1). The OCSs for encoding data transmission on the *AS* channel are generated using the Golay code generation algorithm [46]. The most important characteristics of OCSs that should be considered before choosing an algorithm are auto/cross correlation, length of the generated OCSs versus the number of possible OCSs, and fault tolerance capabilities. Golay OCSs can be generated recursively, as shown in Eqn. 1.

$$C_L = \begin{bmatrix} C_{\frac{L}{2}} & C_{\frac{L}{2}} \\ C_{\frac{L}{2}} & -\bar{C}_{\frac{L}{2}} \end{bmatrix} \qquad \forall L = 2^M, \quad M \geq 1,\ C_1 = \bar{C}_1 \qquad (1)$$

$$when \quad C_L = [A_L \quad B_L] \qquad and \quad \bar{C}_L = [A_L \quad -B_L]$$

In Eqn. 1, $L = 2^M$ is the total number of available OCSs, where $M \geq 1$ is the number of bits in each OCS. A_L and B_L are $L \times L/2$ sub-matrices.

Let us assume that time is divided into periods of random length denoted by a random variable ψ . During each period, each smart meter is assigned a subset of OCSs for use in that period by the UC. The assignment happens over the *US* signaling channel. The communications over the *US* channels are secured using symmetric key cryptography and shared keys between the smart meter and UC based on what has been proposed in [18], [31], [36]. The OCSs for each smart meter are randomly selected by the UC from a large

pool of available OCSs. Each smart meter will use the OCSs uniquely assigned to it in the time frame ψ. In order to spread data bits on the *AS* data channel, the smart meter calculates the inner-product of every data-bit with appropriate OCS. Every single bit of data is coded independently with an OCS different from the previous and next data bit. This will build the foundation of our secure scheme as described in section 4.2. It should be noted that it is possible for multiple smart meters to use the same OCS for data transmission in different parts of the network as long as they are not in the same cluster.

3.3 Adversary Model

In any networking scenario, all individuals in the network can fall into three broad categories based on their behavior. (i) *Honest* entities that fully follow the rules of the established protocol. (ii) *Malicious* or *cheating* nodes that not only do not follow the protocol but also try to manipulate, forge or deny access to possible resources. (iii) *Semi-honest* or *honest-but-curious* nodes follow the defined protocols but they will, or they can, infer privacy-sensitive data. In our proposed scheme we consider the UC as the only honest party. The aggregators are assumed to follow the semi-honest model. The neighboring SMs are, generally, semi-honest; however there can be some malicious nodes in the vicinity. Our objective in this part of the project is to secure all the SM communications against possible eavesdropping, spoofing (or integrity), and inference attacks by the malicious and/or semi-honest nodes.

4. AgSec: Secure Aggregation Technique

4.1 Initialization Phase

Upon initial deployment, the UC communicates control information to each smart meter through the WAN interface on the *US* channel. For each time duration ψ_i, the UC assigns each smart meter, SM_j, a set of attributes including, a temporary eight-bit identifier ($ID_{i,j}$) and a group of valid OCSs, denoted by $G_{O,i}^{j} = \{OCS_1^j, OCS_2^j, ..., OCS_g^j\}$. The integrity, authenticity and confidentiality of the communication between the UC and the SMs are ensured using appropriate symmetric or public-key cryptographic techniques (say, using pre-shared keys). In this phase every smart meter gets the information required for data transmission on the CDMA channel in the next *t* time-slots, as illustrated in Figure 3. It should be noted that, as this is a one-time process in every *t* time slots and $\psi_t >> \psi_i$, the imposed overhead due to it is fairly small. Also, we are not including any frame-level error checking mechanisms such as CRC because spread spectrum, by nature, can tolerate a certain amount of fault.

4.2 Secure Aggregation Protocol (AgSec)

After all smart meters are configured with appropriate OCS and ID information; they start to transmit their readings every τ seconds [3]. Different time intervals, ranging from 30 seconds to a few hours, could be found in the literature [3]. Each node *j* is assigned a group of OCSs ($G_{O,i}^{j}$) for each time interval ψ_i. The k^{th} bit of the data stream generated by SM_j will be coded with $O_{(k \bmod g)}^{j}$, where *g* is the total number of OCSs assigned to SM_j in a given timeslot ψ_i. The OCS $O_i(t)$ assigned to any SM_i at any instant in time *t* can be represented as shown in Eqn. 2.

$$O_i(t) = \sum_{j=0}^{L-1} O_{(j,i)}.p(t - jT_c) \qquad (2)$$

In Eqn. 2, $p(t)$ is a rectangular pulse which is equal to 1 for $0 \le t < T_c$ and zero otherwise. T_c is the chip duration of the OCS and $O_{(j,i)}$ is the j^{th} bit of the OCS assigned to SM_i (from the set of all OCSs C_L). The signal generated after encoding a data symbol of SM_i with the corresponding OCS is given by

$$x_i(t) = f_i \sum_{j=0}^{L-1} O_{(j,i)} p(t - jT_c), 0 \le t < T_f \qquad (3)$$

where, f_i is the data symbol of SM_i that needs to be encoded and $T_f = L.T_c$ is the duration of the encoded data symbol or data bit. The inner product of the sent bit with the OCS is

done bit-synchronously. Then, the overall transmitted signal *x(t)* of all *n* smart meters in a cluster can be given by Eqn. 4 [46].

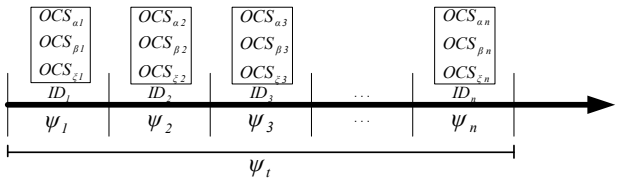


Figure 3. *Initialization Parameters for AgSec*

$$x(t) = \sum_{i=1}^{n} x_i(t) \tag{4}$$

CH will receive a signal including all the bits transmitted by all the smart meters. The received signal will be decoded by CH using all valid OCSs generated by the same algorithm with which they were initially produced by the UC. Given that SMs code their bits with different OCSs at every transmission it is difficult for the CH to decode and extract the actual data from the incoming signal. It should be noted that CH does not know the OCSs assigned to every single SM in the time period ψ_i, all it knows is a list of all possible OCSs in the network. Hence, after decoding the received signal it only has a bit-stream in which neither the IDs, nor the actual data, can be interpreted. After the decoding phase, CH has an *L* bit data stream for every available OCS. All corresponding bits of the decoded data with all possible OCSs will be added and placed in an *L*-element array. Each element of the array is between *–L* and *+L*. The produced array will be sent to the UC as a whole piece of data on the dedicated point-to-point *UA* link.

After the array is received at the UC it is easily decoded and interpreted into actual data transmitted by smart meters. Since UC maintains a table of assigned OCSs (in the same order that was agreed in the initialization phase) and IDs to every single SM in the network, it is able to retrieve the actual data by using appropriate OCS for every bit. We would like to note that the mentioned process is performed on the actual received data in upper layers rather than the physical layer.

Also, we would like to argue that the possible malicious nodes in the network are not able to eavesdrop any information. Given that every single bit of the data is coded with a different OCS, even if packets are captured, they cannot be decoded. The only entity in the network that knows about the set of assigned OCSs to the smart meters is the UC. Hence, all communications are secured against eavesdropping. Our proposed secure aggregation technique is outlined in protocols 1, 2 and 3.

```
1: Function (US operation)
2:   For each period ψ_k do
3:      Generate OCS table with Golay algorithm;
4:      Function (Initialization);
5:      Function (UA data channel);
6:   End For
7: End Function
8: Function (initialization)
9:    Establish a safe communication with each SM;
10:   Generate random SM ID;
11:   OCS dedicator unit grant some OCSs to each SM;
12:End function;
13:Function (UA data transmission)
14:    While data on UA channel do
15:       For (all valid OCS)
16:          Decode each received bit stream on a particular OCS by inverse inner product;
17:       End for
18:       Collect all data bits;
19:       Assemble bits based on ID & OCS;
20:    End while;
21:    Controller check decrypted data for being in thresholds;
22:    Utilize the aggregated data;
23:End function;
```

Protocol 1. *UC functions*

```
1: While data on CDMA data channel do
2:       Receive all signals from different carriers;
3:       Calculate the SUM of each corresponding bits' column of OCSs;
4:       Send calculated SUM values to UC on point-to-point channel;
5: End while
```

Protocol 2: *CH functions*

```
1: Function (SM operation)
2:   While network is ON do
3:      Function(US data)
4:      Function (Metering Engine);
5:   End While
6: End Function
7: Function (US data)
8:    While data on US control channel do
9:       If (receive signal come from UC) then
10:         Update OCSs' table and their orders;
11:      End if
12:   End while
13:End Function
14:Function (Metering Engine)
15:   While (Metering Engine produce value) do
16:      Get a OCS from OCS changer ;
17:      ADD, D random value to data frame;
18:      Encode k^th bit of data frame by (k mod g)^th OCS;
19:      Spread encoded bit stream on AS CDMA carrier;
20:   End while
21: End Function
```

Protocol 3: *SM functions*

5. AgSec: Evaluation and Numerical Results

As discussed in section 2.1, existing secure aggregation schemes impose a significant communication and computation overhead on SGNs with limited capabilities. Aggregation schemes that take advantage of the homomorphic properties of cryptosystems require fixed large size input blocks which is not ideal for small-sized data generated by SMs. The 20 to 30 bit [5] output data generated by SMs has to be padded, e.g., to 2048 bits for Paillier [34], before encryption. In our approach, by choosing OCSs with appropriate length, this overhead can be significantly reduced. Readers should note that in our scheme each bit will be spread to *L* bits after encoding.

We are evaluating our results with clusters of ten and also twenty smart meters and assuming that each smart meter is assigned three OCSs to use in every given time slot. Hence, using an OCS with L=32 and L=64 will be ideal for each scenario, respectively. The OCS length L limits the maximum number of users per cluster to $\frac{L}{|G_{O,i}^{j}|}$. The number of total users in the network is independent from the OCS structure used.

$$D_T = \frac{(F + H_{ID}) \times L}{R} \tag{5}$$

Where, F is the frame length, H_{ID} is the ID header, L is the OCS length and R is the link bit-rate. Given Eqn. 5, the transmission delay using L=32 and L=64, assuming a 200 kbps ZigBee link, is 4.8 *ms* and 9.6 *ms*, respectively. However, using traditional homomorphic cryptosystems as proposed by [18], we have:

$$D_T = \frac{(H_{ID} + D_C + T_{CRC})}{R} \tag{6}$$

Where, H_{ID} is the identifier header, D_C is the encrypted data (payload) and T_{CRC} is the error-checking trailer. Given the values used in [3], the transmission delay will be 10.44 *ms*. Hence, using an OCS with appropriate length we were able to decrease the overhead significantly, as seen in Table I. It should be noted that we are only considering the transmission delay. Moreover, given the high processing load and queuing delays due to the non-simultaneous transmission and high BER and retransmissions, the overall delay of the homomorphic approaches are too high compared with *AgSec*. Table 1 summarizes the transmission delay and total communication overhead $(= \frac{Transmitted\ Data}{Actual\ Data})$ for one smart meter.

Another shortcoming of the secure aggregation schemes based on homomorphic properties of well-known cryptosystems, such as [18], is that every node's data should be passed hierarchically to the upper level node in the aggregation tree. This process

continues until all the data is aggregated at the UC. However, this can increase the total delay which depends on the height of the aggregation tree. Our approach overcomes this issue as all nodes are able to transmit their data simultaneously and independently.

Table 1. *Transmission Delay and Communication Overhead for AgSec*

	AgSec L=32 bits	**AgSec L=64 bits**	**Homomorphic (Paillier)**
Transmission Delay (ms)	4.8	9.6	10.44
Communication Overhead	43.63	87.26	94.91

Moreover, cryptographic solutions usually require heavy processing and computational operations, which is not suitable for smart meters with resource constrained processors. However, our secure aggregation protocol does not put extra processing burden on the smart meters, as it only requires basic addition and multiplication which can also be efficiently accomplished at the circuit level.

6. Seer Grid: Assumed SGN Architecture

For this part of the project, we assume a similar hierarchical three-level SGN architecture (Figure 4).

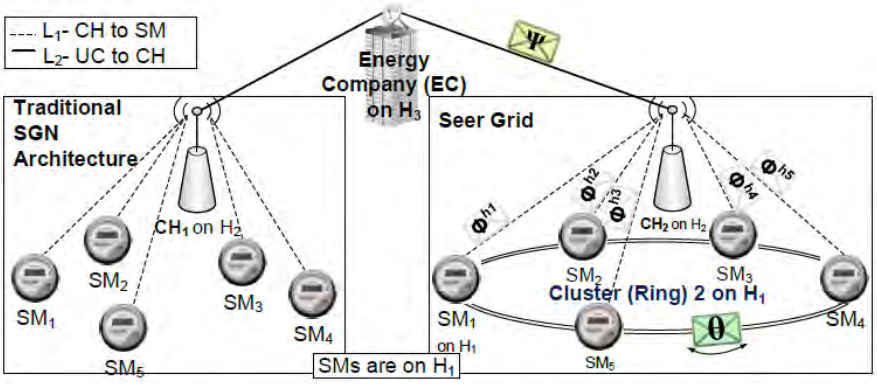


Figure 4. *Traditional SGN architecture on the left, and our proposed SGN architecture on the right*

At the lower level are the smart meters or SMs, physically located in households of end users or customers. At the middle level, each neighborhood has a cluster head or CH, and SMs report predicted energy consumption patterns to CH. At the higher level is the electric company or EC (also referred as UC in the previous part), to which all CHs report aggregated load of their respective neighborhood. The load reporting from all CHs aids EC in optimizing generation and distribution of electricity. Further, we assume that the CH is capable of measuring the actual electricity usage of the whole cluster for a given time interval. We also consider billing once as a month event, which can be done separately.

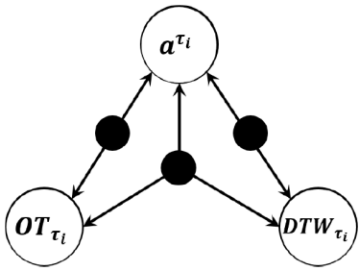


Figure 5. *Interaction between* $a^{(\tau_i)}$ *and* OT_{τ_i} *is 2-way. Interaction between* $a^{(\tau_i)}$ *and* DTW_{τ_i} *is also 2-way. And there exists a 3-way interaction between* $a^{(\tau_i)}$, OT_{τ_i} *and* DTW_{τ_i}. *The prediction model must learn these interactions in order to make effective predictions.*

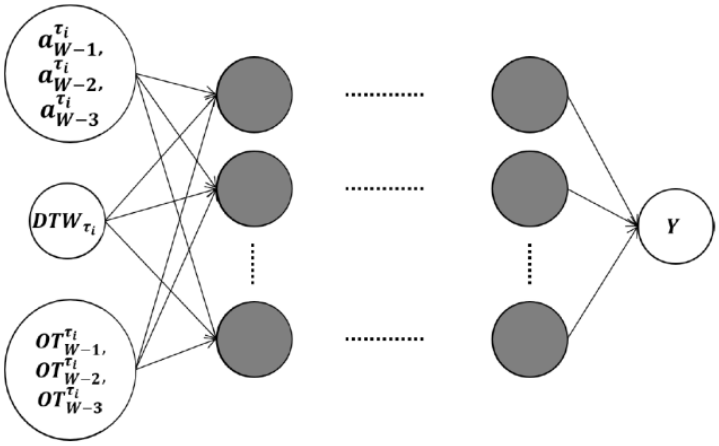


Figure 6. *The abstract structure of the MLP used of learning and prediction.* $a_{W-1}^{(\tau_i)}, a_{W-2}^{(\tau_i)}$ *and* $a_{W-3}^{(\tau_i)}$ *is the power usage in the* τ_i *th interval from last 3 weeks;* DTW_{τ_i} *represents day and time of the week,; and* $OT_{W-3}^{(\tau_i)}$ *and* $OT_{W-1}^{(\tau_i)} - OT_{W-2}^{(\tau_i)}$ *are the outdoor temperature (in Fahrenheit) in the* τ_i *th interval from last 3 weeks.*

We assume a passive adversary who tries to infer personal information of customers based on accessible energy consumption data. If given access to actual energy consumption data, the adversary is computationally capable of carrying out inference attacks by analyzing the data. We also assume that the adversary can access energy consumption data reported to CH and/or EC. However, CH and EC must be honest and cooperate with each other for the protocol to function properly. Thus, CH and EC can be considered as honest but curious. All SMs are assumed to be honest. As a result, we do not scrutinize collusion attacks between SMs and CH, or between SMs and EC.

7. Seer Grid: Technical Background and Prediction Mechanism

We carefully analyzed various statistical learning algorithms for predicting energy consumption patterns, in order to identify the algorithm apposite for preserving only the desired characteristics of the consumption pattern data. In this chapter, we first detail the constituents and properties of the consumption pattern data, followed by a discussion on how we select prediction algorithms for SM and CH.

7.1 Prediction at SM

Traditional SMs report energy usage data to EC in short time intervals, where each reporting conveys the energy used since last reporting. Let us denote the actual daily SM energy consumption pattern of a household h_k as $A^{h_k} = \{a^{(\tau_1)}, a^{(\tau_2)}, ...a^{(\tau_n)}\}$, where $a^{(\tau_i)}$ is the energy consumed since $a^{(\tau_i - 1)}$. The goal of using a prediction model at the SM is to predict a pattern $\phi^{h_k}_{day_j} = \{p^{(\tau_1)}, p^{(\tau_2)}, ...p^{(\tau_n)}\}$, such that there occurs high overlapping between $\phi^{h_k}_{day_j}$ and $A^{h_k}_{day_j}$, but $\phi^{h_k}_{day_j}$ is free of specific load signatures (such as spikes and plateaus). Predictive modeling leverages statistics to predict outcomes, i.e., the forecast of a day's consumption pattern is based on collection of past A^{h_k} (let's say for *m* days). After extensive analysis we identified the input variables critical to the outcome of the prediction model as (i) power usage history in each time interval $a^{(\tau_i)}$, (ii) outdoor temperature in each interval (OT_{τ_i}), and (iii) day and time of the week (DTW_{τ_i}). Each day of the week is considered differently so as to improve prediction based on weekly routines. All interactions present between these three variables are represented in Figure 5. Popular time series forecasting uses a statistical model for predicting future values based on previously observed values. However, basic time series forecasting does not capture the complex interaction between different input variables, resulting in inferior forecasting. Due to the highly complex interactions and some dependencies between input variables, multi-class classification and regression analysis will also result in non-optimal prediction. To achieve best prediction results, we use structured prediction using supervised machine learning techniques. Among candidate machine learning techniques for structured prediction, we decided to use multi-layered perceptron (MLP) because it is specifically designed to discover the complex interactions among input variables. MLP is a feed forward artificial neural network (ANN) model that uses a nonlinear activation function to map sets of input data onto a set of appropriate outputs. MLPs consisting of three or more layers (input, output, and one or more hidden layers) is called a deep neural network, where each node in one layer connects with a certain weight w_{pq} to every node in the following layer. The error in output of a node *q* in the n^{th} training data point is represented as $e_q(n) = d_q(n) - y_q(n)$, where *d* is the target value and *y* is the value

produced by the perceptron. The calculated error for each training data point is used to make corrections to the weights of the node as $\xi(n) = \frac{1}{2}\sum_q e_q^2(n)$, which in turn minimizes the error in the entire output of the ANN. Change in each weight during an epoch is calculated as $\Delta w_{qp}(n) = -\eta \frac{\partial \varepsilon(n)}{\partial v_q(n)} y_p(n)$, where y_p is the output of the previous neuron and η is the learning rate.

In the learning phase of our MLP execution, for each epoch we input power usage history of last three weeks recorded in 5 minute intervals. Outdoor temperature for the corresponding interval and day of the week is also fed in each epoch (Figure 6). The output of the ANN is a structured object (Y) containing multiple possible $\phi_{day_j}^{h_k}$ for next day. Given the next day's temperature forecast and day of the week is known, the structure object is parsed for the matching $\phi_{day_j}^{h_k}$. More details about the MLP specifications used in our simulation experiments can be found in chapter 8.

7.2 Prediction at CH

The purpose of using prediction at SM is to remove specific load signatures (such as spikes and plateaus) form $A_{day_j}^{h_k}$. Although the missing spikes and plateaus from the SM of one household represent a minuscule amount of energy for the grid, spikes and plateaus from multiple households in a cluster can add up to a significant amount of unpredicted energy, which can endanger proper functioning of the electricity grid. Thus, we introduce another level of statistical prediction at the CH based on historical load profile of the cluster, while also factoring in individual predictions from all SMs in the cluster $\{\phi_{day_j}^{h_1}, \phi_{day_j}^{h_2}, \phi_{day_j}^{h_3}, ...\}$. The algorithm (Protocol 4) uses average of difference between past load predictions and actual loads of the entire cluster $(\Lambda_{day_d} = \{\lambda^{(\tau_1)}, \lambda^{(\tau_2)}, ..., \lambda^{(\tau_n)}\})$, in order to complement missing loads. The output of the algorithm $\Psi_{day_j} = \{\psi^{(\tau_1)}, \psi^{(\tau_2)}, ..., \psi^{(\tau_n)}\}$ is the prediction for the whole cluster reported to CH, where $\psi^{(\tau_i)} = \delta^{(\tau_i)} + \sum_k p^{(\tau_i)}$ and $\delta^{(\tau_i)} = \frac{\sum_{d=j-m}^{d=j-1} \{\lambda_{day_j}^{(\tau_i)} - \sum_k p_{day_d}^{(\tau_i)}\}}{m}$. Although trivial, the algorithm can achieve high accuracy.

1: **Prediction Function (for day *j*)**
2:
3: Define new $\Psi_{day_j} = \{\psi^{(\tau_1)}, \psi^{(\tau_2)}, ..., \psi^{(\tau_{288})}\}$
4:
5: **for** *k*=1 to *K* (*K* households in the cluster) **do**
6: $\sum_k p_{day_j}^{(\tau_i)}$
7: **end for**
8: **for** *i* = 1 to 288 (5 minutes time intervals for 24 hours) **do**
9: **for** *d* = 1 to *m* (*m* days to historical data) **do**
10: $\delta^{(\tau_i)} = \lambda_{day_d}^{(\tau_i)} - \sum_k p_{day_d}^{(\tau_i)}$
11: **end for**
12: $\delta^{(\tau_i)} = \frac{\delta^{(\tau_i)}}{m}$
13: $\psi^{(\tau_i)} = \delta^{(\tau_i)} - \sum_k p^{(\tau_i)}$
14: **end for**
15: Report Ψ_{day_j} to CH

Protocol 4: *Prediction Algorithm Executed by CH.*

7.3 Seer Grid Prediction Mechanism

Similar to the traditional SGN architecture, Seer Grid also consists of a hierarchical three-level network (Figure 4). At the lowest level are the SMs, physically located in households. At the middle level, each neighborhood has a CH, and SMs reports predicted energy consumption patterns to CH. At the higher level is the EC, to which all CHs report aggregated and re-predicted energy load forecast of their respective neighborhood. The predicted load forecast from all CHs aids EC in optimizing generation and distribution of electricity.

A distributed model of SMs is used in our SGN model, where the first level prediction is performed independently on all SMs belonging to the SGN. The prediction algorithm running at the SM of a household h_k locally stores a small database (Figure 7), containing actual consumption patterns A^{h_k} and outdoor temperature measurements OT_i from last *m* days. Each day, the A^{h_k} and OT_i values are used to train the MLP and predict the $\Phi_{day_j}^{h_k}$ for next (*j*-th) day. Also, at the end of a day, the day's actual consumption pattern $A_{day_{j-1}}^{h_k}$ is inserted in to the queue of the database and the oldest actual consumption pattern $A_{day_{j-m+1}}^{h_k}$ is removed. As mentioned before, $\Phi_{day_j}^{h_k}$ is computed and reported only once (before beginning of) each day. All communications for reporting $\Phi_{day_j}^{h_k}$ are assumed to be point-to- point and symmetrically encrypted, for example, using AES. CH accumulates all $\Phi_{day_j}^{h_k}$ in the cluster, adds the calculated $\delta^{(\tau_i)}$ to $\sum_k p^{(\tau)}$ for each time interval (τ), and

reports the resulted pattern Ψ_{day_j} to the EC. CH also stores a database of past Λ and $\sum_k p^{(\tau)}$ values from last *m* days, which is updated at the end of each day (Figure 8).

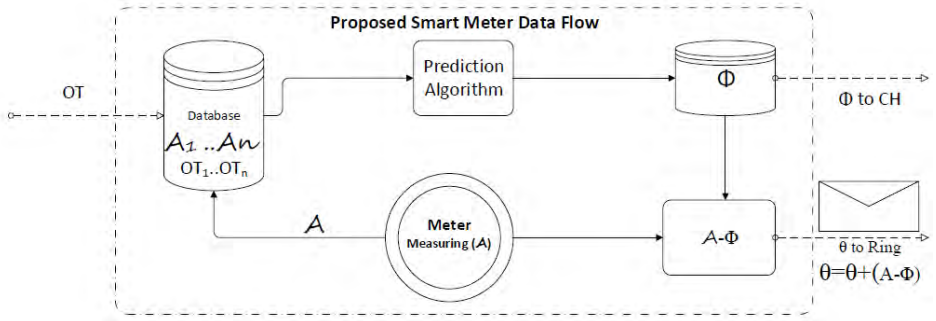


Figure 7. *Proposed Smart Meter Data Flow.*

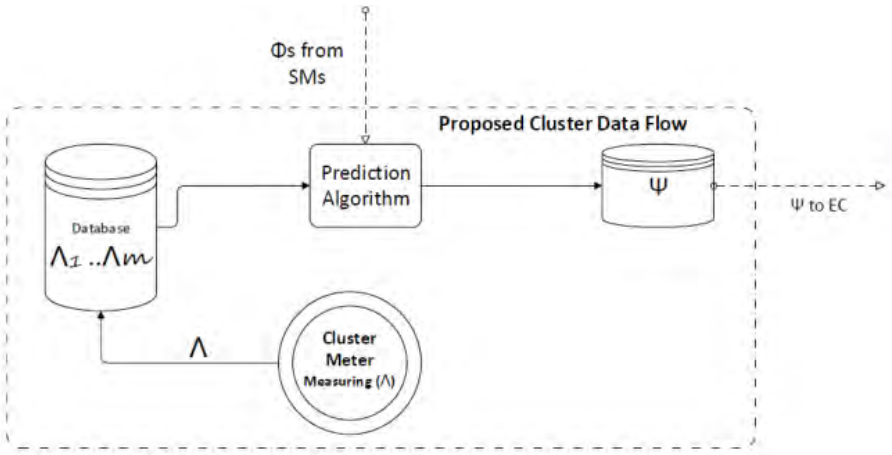


Figure 8. *Proposed Cluster Head Data Flow.*

The cluster level predicted pattern Ψ_{day_j} is a refined estimate of next day's energy consumption compared to individual SM predictions $\Phi^{h_k}_{day_j}$, but it is still not a definite future. There may occur unexpected events which are not captured by the input variable of our prediction system. To ensure proper functioning of SGN in such case of an unexpected power demand, we incorporate a real-time reporting system in our architecture to measure the difference in actual and predicted energy consumption of all households. But, directly reporting difference in actual and predicted energy consumption pattern to CH defeats our goal of privacy, because CH can add back the difference to predicted pattern to obtain the actual pattern of individual SMs. So, the real-time reporting system uses a token ring mechanism to aggregate the difference in actual and predicted energy consumption pattern for all SMs in the cluster. The token ring calculates

the difference in actual and predicted energy consumption $\theta^{(t)} = \sum_{k} (p^{(\tau)} - a^{(\tau)})$ in each time interval τ. The final token value containing the aggregated difference in actual and predicted energy consumption of the cluster is reported to EC (via CH) for regulating generation and distribution, if necessary. Figure 7 illustrates how each SM adds their difference in actual and predicted energy consumption to the token, and Figure 4 show how the token ring is circulated across all SMs in the cluster for aggregation of difference in actual and predicted energy consumption. To protect the token ring against eavesdropping attacks, all SMs symmetrically encrypt (and decrypt for addition) the token using a shared secret, obtained using, say a, authenticated Diffie-Hellman key exchange protocol.

8. Seer Grid: Empirical Evaluation

In order to validate the benefits of our Seer Grid architecture, we conduct extensive simulation experiments using real SM data. In this chapter, we present our experimental setup followed by an overview of the simulation results.

8.1 Experimental Setup

We use real SM data from East Midlands, UK from the year 2008. The data is collected from residences equipped with BS EN62053-21002003 smart meter, which measures true active power in five-minute time intervals. The fabricated cluster we consider for evaluation consists of 5 household, each having one smart meter. Due to limited memory of SMs, we limit the use of historical data in our experiments to last three weeks, i.e, m = 21. Longer training period not only takes more storage space, but also makes less significant contribution in the prediction because of changing temperature conditions throughout the year. The ANN prediction algorithm is trained with data from past 21 days to predict the energy consumption for 3 test days. The training data consists of nine variables: interval number and target date as indexing variable, 3 power usage measurements in the interval from last three weeks, and 3 outdoor temperature measurements in the interval from last three weeks. More specific details of the parameters applied to train the ANN can be found in Table 2. We interpret our results by using the classical squared correlation coefficient (R^2) to measure the strength of relationship between predicted and actual energy consumption patterns at each level of Seer Grid. Lower R^2 implies more privacy, whereas higher R^2 implies more utility. We also calculate the normalized relative entropy between the actual energy consumption of a test day and the mean of actual energy consumption during training days of corresponding test day. This information will help us understand how training data with properties different from the test data affect the prediction.

8.2 Results and Observations

The experiments (with 21 training and 3 test days) were performed over four seasons: winter (January 1 to 24), spring (April 1 to 24), summer (July 1 to 24), and fall (October 1 to 24). The results, averaged over the 3 test days, are presented in Table 3. The squared correlation between actual and predicted energy consumption patterns of SM vary between 51.07% and 80.09%, and averages at 62.10% across all 5 SMs. As an example, Figure 9(a) shows the actual and predicted energy consumption pattern for SM3 on 22nd January, and Figure 9(b) shows the squared correlation between them. The squared correlation between actual and predicted energy consumption pattern for CH vary between 89.95% and 91.15%, and averages at 90.60%. Figure 9(c) shows the actual and predicted energy consumption pattern for CH on 22nd January, and Figure 9(d) shows the

squared correlation between them. Evidently, SM prediction is less correlated than CH prediction by a clear margin, as seen in Figure 10.

Table 2. *Neural Network Training Parameters*

Parameter	Value
Number of SMs in cluster (assumed neighborhood)	5
Training period	3 weeks (21 days)
Testing period	3 days
Number of predicted data point a day	288
Number of ANN Inputs	9
ANN Proto	50
Number of ANN hidden layers	3
Number of nodes in each hidden layer	10
Number of ANN output	1
ANN Learn Rule	Ext DBD
ANN transfer mode	Sigmoid
Epoch	288*21=6048
Number of iterations	10^6

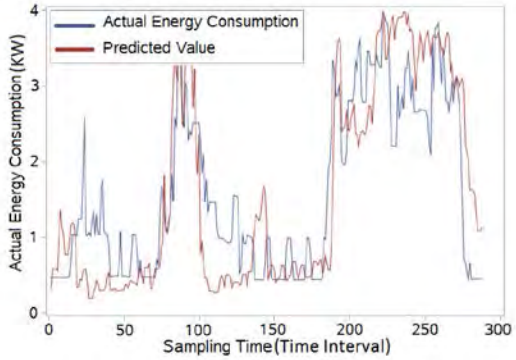
Another interesting observation is about how difference between training and test data affects the SM prediction. Intuitively, having an approximately fixed daily schedule should ease predicting energy load for test day, and high uncertainty in daily schedule should make prediction harder. However, our experimental results show that higher relative entropy between test data and mean of training data leads to more accurate prediction, and thus less privacy. This observation can be associated with the fact that ANN models converge at a faster rate when there is higher variation in the training data. In case we have longer training period, the convergence in learning may be more uniform for all SMs. But we decided to restrain ourselves to only 3 weeks because of reasons discussed before. We also discuss in the following section how the convergence in learning can be made more uniform across SMs by utilizing training data characteristics in regulating the learning rate.

8.3 Discussion

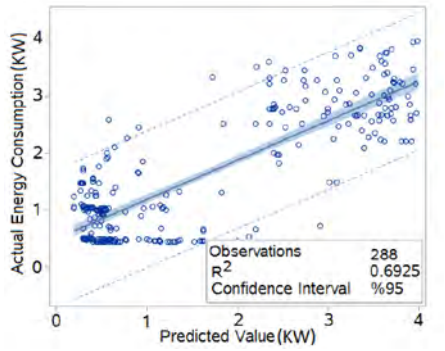
Prediction Parameters: In our experiments, we took a heuristic approach for determining the prediction parameters for the ANN used by the SMs. The parameters were chosen in such way that it satisfies our goal of optimizing both privacy and utility of SM data. From the experiment results we observe that the correlation between actual and predicted energy consumption pattern varies moderately across households and seasons. This is primarily because of different characteristics of the training data (actual energy consumption for last 21 days) leading to differently converged prediction model in each SM. In future, we plan to develop a unified prediction framework for the SMs which will analyze characteristics of the training data, and accordingly govern learning rate such that

prediction accuracy remains below a privacy preserving threshold with high likelihood. Unlike this work, where all SMs use the same prediction parameters, the unified framework will adapt to the characteristics of local training data of individual SMs.

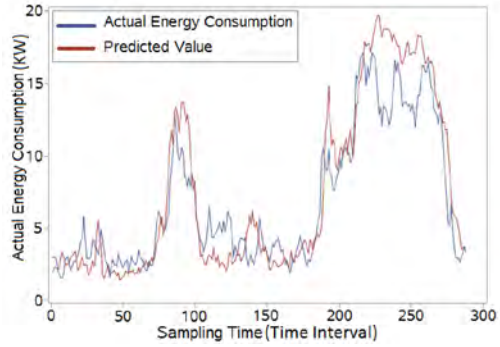
Larger Cluster: We consider a very small scale cluster in our experiments, and yet achieve considerably high prediction accuracy at the cluster level. As evident form previous cluster level prediction schemes [64], accuracy tend to dramatically improve with increasing number of customers in the cluster. Thus, we think our results are highly encouraging for large scale implementation.



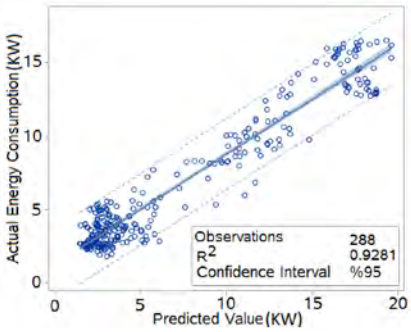
(a) Actual and predicted energy consumption pattern for SM3.



(b) Correlation between actual and predicted energy consumption pattern for SM3.



(c) Actual and predicted energy consumption pattern for CH.



(d) Correlation between actual and predicted energy consumption pattern for CH.

Figure 9. *Exemplary results from 22nd January 2008, showing the correlation between actual and predicted energy consumption patterns at different levels of Seer Grid.*

Table 3. *Squared correlation coefficient (R2) between predicted and actual energy consumption patterns for each SM and CH, and normalized relative entropy between the actual energy consumption of a test day and the mean of actual energy consumption during training days of corresponding test day. All values are average of the 3 test days.*

Season		SM1	SM2	SM3	SM4	SM5	CH
Winter	R^2: Actual vs Predicted	**0.5715**	**0.5529**	**0.7793**	**0.6421**	**0.5772**	***0.9098***
	Relative Entropy: Actual vs Past	0.1785	0.1853	0.2397	0.2069	0.1871	0.1265
Spring	R^2: Actual vs Predicted	**0.5107**	**0.5627**	**0.8009**	**0.6687**	**0.5799**	***0.9115***
	Relative Entropy: Actual vs Past	0.1670	0.1852	0.2415	0.2112	0.1814	0.1224
Summer	R^2: Actual vs Predicted	**0.5888**	**0.5341**	**0.6322**	**0.6439**	**0.6528**	***0.8985***
	Relative Entropy: Actual vs Past	0.1880	0.1682	0.2002	0.1961	0.2047	0.1341
Fall	R^2: Actual vs Predicted	**0.6195**	**0.6025**	**0.6477**	**0.6450**	**0.6072**	***0.9041***
	Relative Entropy: Actual vs Past	0.1956	0.1953	0.2113	0.2053	0.1937	0.1349

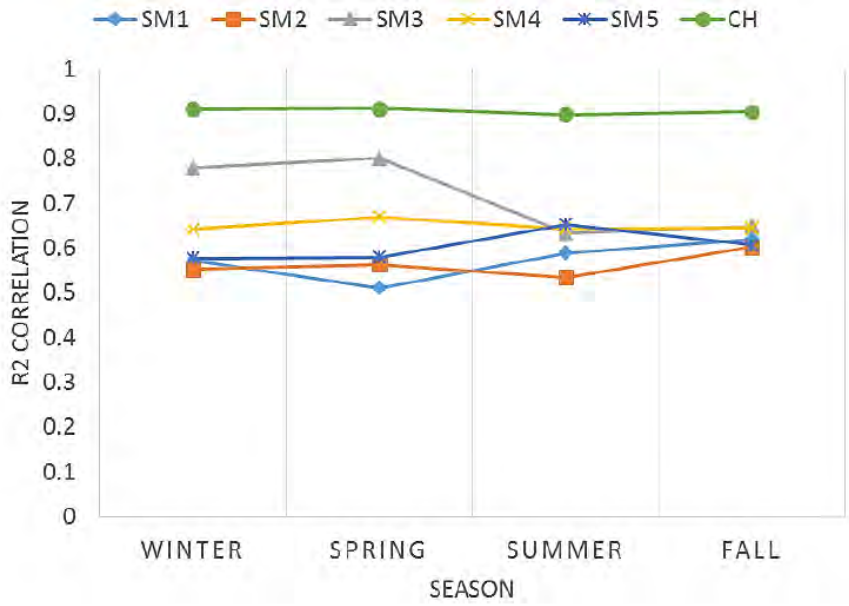


Figure 10. *Correlation between actual and predicted energy consumption patterns for SMs and CH over four seasons (data from Table II). The correlation at CH level is clearly higher than all SMs.*

9. Conclusion

In Chapters 3-5, we presented a new approach for securing data aggregation in smart metering systems. Our proposed approach uses a coding theory approach by utilizing salient properties of spread spectrum communications on a CDMA channel to securely aggregate sensitive power consumption data from smart meters. Our analysis showed that, provided appropriate parameters are chosen, our proposed technique imposes lesser delay and overhead on SGNs as compared to cryptographic approaches. The proposed method uses code division multiplexing to enable simultaneous transmissions, and also results in a reduced bit error rate and interference. As part of future work, we are planning to implement the proposed scheme in a real test-bed of SMs to analyze its security and efficiency in practice.

We introduce Seer Grid method in Chapters 6-8, an alternate SGN architecture aimed to minimize the privacy-utility trade-off faced by SMs. As a result of two-level energy load prediction in Seer Grid, there exists high correlation between predicted and actual energy consumption patterns at cluster level, which indicates excellent utility preservation. However, the correlation between predicted and actual energy consumption patterns of individual SM is weak, which indicates strong privacy preservation.

References

[1] Blackout 2003, http://www.ieso.ca/imoweb/EmergencyPrep/ blackout2003, 2012.

[2] U.S. Department of Energy. "Smart Grid/ Department of Energy", 2012.

[3] I. Rouf, H. Mustafa, M. Xu, W. Xu, R. Miller, and M. Gruteser, "Neighborhood Watch: Security and Privacy Analysis of Automatic Meter Reading Systems," CCS '12, USA, Oct 2012.

[4] A. Barenghi, and G. Pelosi, "Security and Privacy in Smart Grid Infrastructures," DEXA'11, France, Aug 2011.

[5] Z. Erkin, J. R. Troncoso-Pastoriza, R. L. Lagendijk, and F. Pérez-González, "Privacy-Preserving Data Aggregation in Smart Metering Systems: An overview," IEEE Signal Processing Magazine'13, Mar 2013.

[6] A. G. van Engelen and J. S. Collins, "Choices for smart grid implementation," HICSS'10, 2010.

[7] A. Bose, "Smart transmission grid applications and their supporting infrastructure," IEEE Trans. on Smart Grid, 2010.

[8] F. R. Yu, P. Zhang, X. Weidong, and P. Choudhury, "Communication systems for grid integration of renewable energy resources," IEEE Network Magazines'11, 2011.

[9] R. Amin, J. Martin, and X. Zhou, "Smart Grid Communication using Next Generation Heterogeneous Wireless Networks," SmartGridComm'12, Taiwan, Nov 2012.

[10] R. D. Tabors, G. Parker, and M. C. Caramanis, "Development of the smart grid: Missing elements in the policy process," HICSS'10, 2010.

[11] R. Schuler, "Electricity markets, reliability and the environment: Smartening-up the grid," HICSS'10, 2010.

[12] D. Niyato, P. Wang, and E. Hossain, "Reliability Analysis and Redundancy Design of Smart Grid Wireless Communications System for Demand Side Management," IEEE Wireless Communications Magazine, 2012.

[13] B. Falahati, Y. Fu, and L. Wu, "Reliability Assessment of Smart Grid Considering Direct Cyber-Power Interdependencies," IEEE Transaction on Smart Grid, 2012.

[14] K. Moslehi, and R. Kumar, "A Reliability Perspective of the Smart Grid," IEEE Transaction on Smart Grid, Jun 2010.

[15] S. Shao, M. Pipattanasomporn, and S. Rahman, "Demand Response as a Load Shaping Tool in an Intelligent Grid With Electric Vehicles," IEEE Transaction on Smart Grid, Dec 2011.

[16] S. Shao, M. Pipattanasomporn, and S. Rahman, "Grid Integration of Electric Vehicles and Demand Response with Customer Choice," IEEE Transaction on Smart Grid, Mar 2012.

[17] I. C. Paschalidis, B. Li, and M. C. Caramanis, "Demand-Side Management for Regulation Service Provisioning Through Internal Pricing," IEEE Trans. on Power Systems, Aug 2012.
[18] F. Li, B. Luo, and P. Liu, "Secure Information Aggregation for Smart Grids Using Homomorphic Encryption," IEEE SmartGridComm'10, USA, Oct 2010.
[19] Y. Yan, Y. Qian and H. Sharif, "A Secure Data Aggregation and Dispatch Scheme for Home Area Networks in Smart Grid", IEEE Globecom'11, USA 2011.
[20] A. Bartoli, J. Hernández-Serrano; M. Soriano; M. Dohler, A. Kountouris, and D. Barthel, "Secure Lossless Aggregation for Smart Grid M2M Networks," SmartGridComm'10, USA, 2010.
[21] H. S. Fhom, and K. M. Bayarou, "Towards a Holistic Privacy Engineering Approach for Smart Grid Systems," IEEE TrustCom'11, China, Nov 2011.
[22] M. B. Line, I. A. Tondel, and M. G. Jaatun, "Cyber Security Challenges in Smart Grids," ISGT'11, UK, Dec2011.
[23] X. He, M. Pun, and C. -C. J. Kuo, "Secure and Efficient Cryptosystem for Smart Grid Using Homomorphic Encryption," IEEE ISGT'12, USA, Feb 2012.
[24] M. Lisovich and S. Wicker, "Privacy concerns in upcoming residential and commercial demand-response systems," Clemson University Power Systems Conference, 2008.
[25] M. A. Lisovich, D. K. Mulligan, and S. B. Wicker, "Inferring personal information from demand-response systems," IEEE Security and Privacy, 2010.
[26] A. Molina-Markham, P. Shenoy, K. Fu, E. Cecchet, & D. Irwin, "Private memoirs of a smart meter," ACM BuildSys, 2010.
[27] P. McDaniel and S. McLaughlin, "Security and privacy challenges in the smart grid," IEEE Security and Privacy, 2009.
[28] F. Cohen, "The Smarter Grid," IEEE Security & Privacy, 2010.
[29] Computerworld magazine, "Stuxnet renews power grid security concerns", Jul 2010.
[30] R. Lagendijk, Z. Erkin, M. Barni, "Encrypted signal processing for privacy protection," IEEE Signal Process. Mag., 2013.
[31] R. Rivest, A. Shamir, and L. Adleman, "A method for obtaining digital signatures and public key cryptosystems," Communications of the ACM, Feb 1978.
[32] C. Gentry, "Fully homomorphic encryption using ideal lattices," STOC'09, USA, 2009.
[33] M. V. Dijk, C. Gentry, S. Halevi, and V. Vaikuntanathan, "Fully homomorphic encryption over the integers," ACM EUROCRYPT, 2010
[34] P. Paillier, "Public-key cryptosystems based on composite degree residuosity classes," ACM EUROCRYPT, 1999.
[35] E.-J. Goh, "Encryption Schemes from Bilinear Maps," Department of Computer Science, Stanford University, 2007.

[36] F. D. Garcia and B. Jacobs, “Privacy-friendly energy-metering via homomorphic encryption,” STM, 2010.

[37] K. Kursawe, G. Danezis, and M. Kohlweiss, “Privacy-friendly aggregation for the smart-grid,” HotPETs'11, Canada, 2011.

[38] Z. Erkin and G. Tsudik, “Private computation of spatial and temporal power consumption with smart meters,” ACNS, 2012.

[39] G. Ács and C. Castelluccia, “I have a DREAM! (Differentially PrivatE smart Metering),” ACM IH’11, May 2011.

[40] R. Lu, X. Liang, X. Li, X. Lin, and X. Shen, "EPPA: An Efficient and Privacy-Preserving Aggregation Scheme for Secure Smart Grid Communications," IEEE Transactions on Parallel and Distributed Systems, 2012.

[41] W. He, H. Nguyen, X. Liu, K. Nahrstedt and T. Abdelzaher, “iPDA: An Integrity-Protecting Private Data Aggregation Scheme for Wireless Sensor Networks”, IEEE WCPS, 2009.

[42] H. Li, K. Lin and K. Li, “Energy-efficient and high-accuracy secure data aggregation in wireless sensor networks,” ACM Journal Computer Communications, 2011.

[43] C. Weng, M. Li and X. Lu, “Data Aggregation with Multiple Spanning Trees in Wireless Sensor Networks”, ICESS, 2008.

[44] M.B. Zanjani, R. Monsefi and A. Boustani, “Energy Efficient/highly Secure Data Aggregation Method using Tree structured Orthogonal Codes for Wireless Sensor Network,” IEEE ICSTE’10, USA, 2010

[45] M. B. Zanjani, A. Boustani, “Energy aware and highly secured data aggregation for grid-base Asynchronous wireless sensor network,” IEEE Pacrim, Canada, 2011.

[46] H. H. Chen, “The Next Generation CDMA Technologies”. John Wiley and Sons, 2007.

[47] Y. Phulpin, J. Barros, and D. Lucani, “Network Coding in Smart Grids,” IEEE SmartGridComm, 2011.

[48] Telefnica Digital, “The smart meter revolution - towards a smarter future,” Jan 2014.

[49] W. Kleiminger, C. Beckel, T. Staake, and S. Santini, “Occupancy detection from electricity consumption data,” in Proceedings of the 5th ACM Workshop on Embedded Systems For Energy-Efficient Buildings. ACM, 2013, pp. 1–8.

[50] M. Weiss, A. Helfenstein, F. Mattern, and T. Staake, “Leveraging smart meter data to recognize home appliances,” in Pervasive Computing and Communications (PerCom), 2012 IEEE International Conference on. IEEE, 2012, pp. 190–197.

[51] I. Rouf, H. Mustafa, M. Xu, W. Xu, R. Miller, and M. Gruteser, “Neighborhood watch: security and privacy analysis of automatic meter reading systems,” in

Proceedings of the 2012 ACM conference on Computer and communications security. ACM, 2012, pp. 462–473.

[52] M. A. Lisovich, D. K. Mulligan, and S. B. Wicker, "Inferring personal information from demand-response systems," Security & Privacy, IEEE, vol. 8, no. 1, pp. 11–20, 2010.

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

[54] F. D. Garcia and B. Jacobs, "Privacy-friendly energy-metering via homomorphic encryption," in Security and Trust Management. Springer, 2011, pp. 226–238.

[55] A. Rial and G. Danezis, "Privacy-preserving smart metering," in Proceedings of the 10th annual ACM workshop on Privacy in the electronic society. ACM, 2011, pp. 49–60.

[56] C. Efthymiou, G. Kalogridis, S. Z. Denic, T. A. Lewis, and R. Cepeda, "Privacy for smart meters: Towards undetectable appliance load signatures," in Smart Grid Communications (SmartGridComm), 2010 First IEEE International Conference on, 2010, pp. 232–237.

[57] S. McLaughlin, P. McDaniel, and W. Aiello, "Protecting consumer privacy from electric load monitoring," in Proceedings of the 18th ACM conference on Computer and communications security. ACM, 2011, pp. 87–98.

[58] X. He, X. Zhang, and C.-C. Kuo, "A distortion-based approach to privacy-preserving metering in smart grids," Access, IEEE, vol. 1, pp. 67–78, 2013.

[59] S. R. Rajagopalan, L. Sankar, S. Mohajer, and H. V. Poor, "Smart meter privacy: A utility-privacy framework," in Smart Grid Communications (SmartGridComm), 2011 IEEE International Conference on. IEEE,2011, pp. 190–195.

[60] L. Sankar, S. R. Rajagopalan, S. Mohajer, and H. V. Poor, "Smart meter privacy: A theoretical framework," Smart Grid, IEEE Transactions on, vol. 4, no. 2, pp. 837–846, 2013.

[61] R. Dong, A. A. C´ardenas, L. J. Ratliff, H. Ohlsson, and S. S. Sastry, "Quantifying the utility-privacy tradeoff in the smart grid," arXiv preprint arXiv:1406.2568, 2014.

[62] B. Defend and K. Kursawe, "Implementation of privacy-friendly aggregation for the smart grid," in Proceedings of the first ACM workshop on Smart energy grid security. ACM, 2013, pp. 65–74.

[63] R. Anderson and S. Fuloria, "On the security economics of electricity metering," in WEIS. Citeseer, 2010.

[64] T. K. Wijaya, S. F. R. J. Humeau, M. Vasirani, and K. Aberer, "Individual, Aggregate, and Cluster-based Aggregate Forecasting of Residential Demand," Lausanne, Switzerland, Tech. Rep., 2014.

[65] M. Chaouch, “Clustering-based improvement of nonparametric functional time series forecasting: Application to intra-day household-level load curves,” Smart Grid, IEEE Transactions on, vol. 5, no. 1, pp. 411– 419, 2014.

[66] A. Jain and B. Satish, “Short term load forecasting by clustering technique based on daily average and peak loads,” in Power & Energy Society General Meeting, 2009. PES’09. IEEE. IEEE, 2009, pp. 1–7.

[67] ——, “Clustering based short term load forecasting using support vector machines,” in PowerTech, 2009 IEEE Bucharest. IEEE, 2009, pp. 1–8.

[68] T. KU¨ C¸U¨ KDENI˙Z, “Long term electricity demand forcesting: An alternative approach with support vector machines,” ˙I¨U M¨uhendislik Bilimleri Dergisi, vol. 1, no. 1, pp. 45–54, 2010.

[69] R. Sevlian and R. Rajagopal, “Short term electricity load forecasting on varying levels of aggregation,” arXiv preprint arXiv:1404.0058, 2014.

[70] M. Ghofrani, M. Hassanzadeh, M. Etezadi-Amoli, and M. Fadali, “Smart meter based short-term load forecasting for residential customers,” in North American Power Symposium (NAPS), 2011. IEEE, 2011, pp. 1–5.

[71] R. P. Singh, P. X. Gao, and D. J. Lizotte, “On hourly home peak load prediction,” in Smart Grid Communications (SmartGridComm), 2012 IEEE Third International Conference on. IEEE, 2012, pp. 163–168.

[72] C. Efthymiou and G. Kalogridis, “Smart grid privacy via anonymization of smart metering data,” in Smart Grid Communications (SmartGrid- Comm), 2010 First IEEE International Conference on. IEEE, 2010, pp. 238–243.

[73] M. Jawurek, M. Johns, and K. Rieck, “Smart metering depseudonymization,” in Proceedings of the 27th Annual Computer Security Applications Conference. ACM, 2011, pp. 227–236.

Part IV

Incentivizing Privacy-Guaranteed Data-Sharing by Consumers Using AMI

Lalitha Sankar

Chong Huang and Jiazi Zhang, PhD Students

Arizona State University

For information about Part IV, contact:

Lalitha Sankar
School of Electrical, Computer & Energy Engineering
PO Box 87506
Arizona State University
Tempe, AZ 85257-5706
Fax: (480) 965-0745
Tel: (480) 965-4953
Email: lalithasankar@asu.edu

Power Systems Engineering Research Center

The Power Systems Engineering Research Center (PSERC) is a multi-university Center conducting research on challenges facing the electric power industry and educating the next generation of power engineers. More information about PSERC can be found at the Center's website: http://www.pserc.org.

For additional information, contact:

Power Systems Engineering Research Center
Arizona State University
527 Engineering Research Center
Tempe, Arizona 85287-5706
Phone: 480-965-1643
Fax: 480-965-0745

Notice Concerning Copyright Material

Contents

List of Figures

List of Tables

Chapter 1

Incentivizing Consumers with Access to Renewables

1.1 Introduction

Renewable energy resources, especially roof top PV systems, are getting more and more prevalent in the distribution level of the grid. As a result, there is a need to monitor energy consumption patterns at the distribution level for more efficient dispatch and better system operation. This means the utility has to incentivize households to use smart meters. However, the installment of smart meter may create potential threats to households privacy since it has much higher sampling rate and data processing capability than traditional meter.

Privacy and security, especially in power system, have become challenging issues with the development of advanced metering infrastructure and communication technology. The ability to share data has many benefits, for example, improving load forecasting and system dispatch efficiency. However, the collected information may be used by malicious users or innocent-but-curious utility companies to analyze households electricity consumption behavior and make inferences about personal habits of consumers. Thus, privacy aware households may try to mask their actual energy consumption profile, or even refuse to use smart meters so that they can have some privacy. Therefore, we study the problem of how utility companies can incentivize households to share their consumption profile through smart meter; while offering households guaranteed levels of privacy.

1.2 Background and Related Work

The increasing number of smart meters deployed in business and residential buildings has raised concerns about privacy. The adversary can monitor households's energy consumption behavior by using data collected from smart meters [1–3]. Recent work [4] provides an overview of privacy protecting technologies in smart meter. Multiple methods and metrics have been proposed to quantify and protect smart meter privacy. One common approach is to develop control policies by using battery to hide actual electricity consumption behavior [5–7]. Another approach of interest is distorting the metering data from the smart meter to preserve user privacy [8]. Moreover, the privacy of households can be protected by using anonymization of smart meter data [9]. all of these approaches can effectively provide consumers a certain level of privacy. However, none of them considers the set of all potential tradeoff between privacy and utility.

In [10], Denic *et.al.* showed that privacy preserving algorithms which use battery to mask load behavior can affect the consumer electricity demand from the grid, and thus affect electricity prices. In [11], Tan *et.al.* proposed a model which protects the smart meter privacy from an information theoretic perspective via battery and energy harvesting unit coordination. Ratliff *et.al.* [12] studied a contract based mechanism to protect consumer provacu as well as maximizing the social welfare of electric utility companies and consumers. Recent work by Yao and Venkitasubramaniam [13] proposed a MDP based model to analyze the tradeoff between privacy and energy saving when smart meters are deployed in households. They used information theoretic metric to quantify the privacy leakage when meters share consumption data with utility companies. Our work differs in that we capture privacy by the response of households to incentives offered by the utility company in order to encourage them to share some information about their electricity consumption profile. A closely related work is the utility and privacy tradeoff problem studied by Dong *et.al.* [14], in which they proposed a privacy metric based on hypothesis testing. They assumed an adversary can infer private information via sampling data from smart meters and analyzed the tradeoff between smart meter operation and household privacy by changing the sampling rate of the smart meter. However, their privacy model has limited capability to capture data privacy when there are batteries and PVs in the system. This is because even if the sampling rate is very high, the user behavior still can be masked by controlling battery and PV generation.

1.3 Contibution

In this report, we seek to address households responses to incentives from the utility company in order to encourage households to share more data. In particular, we assume that the households privacy is protected by masking the actual consumption via controlling the operation of battery and PV. Theoretically, all usage patterns can be masked by charging and discharging the battery to maintain a constant consumption profile. However, this mechanism requires the battery to have sufficiently large storage capability and charging/discharging rate; furthermore, it is very challenging to control the operation of battery to achieve perfectly constant energy consumption over all time. The utility company requires certain amount of energy consumed by households in order to perform load forecasting and maintain efficient electricity dispatch. Thus, it may offer some incentives to households to encourage them to consume energy from the grid. Each household faces the tradeoff between reducing consumption from the utility company for privacy reasons and revealing consumption patterns to the utility company to reduce the energy cost.

The main contribution of this report is to propose a novel approach to study the tradeoff between privacy and energy cost minimization under the assumption that the utility company offers incentives to households to encourage data sharing through energy consumption. In this respect, we formulate a non-cooperative game to model interactions between households and the utility company. In this game, the strategy of a household is to select the proportion of electricity it consumes from the grid to maximize its reward from the grid while maintaining certain privacy; the strategy of the utility company is the incentive price it offers to encourage consumption so that it can maximize its profit. The final incentive price results from matching the valuation of privacy from each household with the incentive price announced by the utility company. To solve the game, we propose an algorithm based on iterative best response dynamics that enables the households and the utility company to reach a Nash equilibrium.

1.4 System Model

Below we present our system model as well as optimization problems for households and the utility company.

1.4.1 Household Model

We assume that there are M households in the system denoted by the set $\mathcal{H} = \{1, 2, ..., M\}$. Each household has installed smart meter and renewable energy generation (PV) and energy storage device (battery). For household i, let $D_{i,t}$ be the actual consumption at time t, *i.e.*, it is the demand of household i at time t. By using PVs and Batteries, household i can control its energy consumption from the grid. We denote $\alpha_{i,t} \in \mathcal{A}_i$ to be the consumption coefficient, where $\mathcal{A}_i$ is the support set of $\alpha_{i,t}$. It indicates the proportion of the total amount of energy that household i consumes from the grid. Thus, at time t, household i only consumes $\alpha_{i,t}D_{i,t}$ from the grid for economic and privacy reasons.

1.4.2 Utility Company Model

The smart meter installed by the utility company samples the energy consumption at a fixed rate. It transmits the actual consumption data back to the utility company instantly. We assume that the utility company requires at least X_t amount of energy to maintain base load and perform load forecasting. If the consumption is lower than X_t, it suffers a loss $L(X_t - \sum_{i \in \mathcal{H}} \alpha_{i,t} D_{i,t})$ where $L(\cdot)$ is assumed to be a continuous, convex and increasing function.

In order to incentivize households to use the energy from the grid and share their actual consumption with the utility companies, the utility company can offer a price to compensate for the privacy loss of the consumer in the household. We first assume that there is no price differentiation between each household. To be more specific, the utility company offers an incentive rate of $\beta_t \in \mathcal{B}$ for each KW of electricity consumed by a household, where $\mathcal{B}$ is the support set of β_t. Each household has a valuation of privacy $s_{i,t}$, which denotes how much of its consumption profile it wishes to share with the grid, and is given by an arbitrary function $s_{i,t} = f_{i,t}(\alpha_{i,t}D_{i,t})$. We assume that $f_{i,t}(\alpha_{i,t}D_{i,t})$ is an invertible, continuous, and increasing function.

1.4.3 The Optimization Problem

Since the consumer wishes to be compensated for sharing data with the utility company, it is intuitive to expect that the consumer will try to match its valuations to the incentive it receives from the utility. Thus, the household consumes a fraction $\alpha_{i,t}D_{i,t}$ from the utility company such that $\beta_t = s_{i,t} = f_{i,t}(\alpha_{i,t}D_{i,t})$, where $s_{i,t}$ is defined to be the clearing price. We denote $\boldsymbol{\alpha}_t = \{\alpha_{1,t}, \alpha_{2,t}, ..., \alpha_{M,t}\}$ and β_t to be the strategy profile of all households and incentive price offered by the utility company at time t, respectively. The reward function of household i for consuming electricity from the utility is given by:

$$U_i(\alpha_{i,t}) = \beta_t \alpha_{i,t} D_{i,t}. \tag{1.1}$$

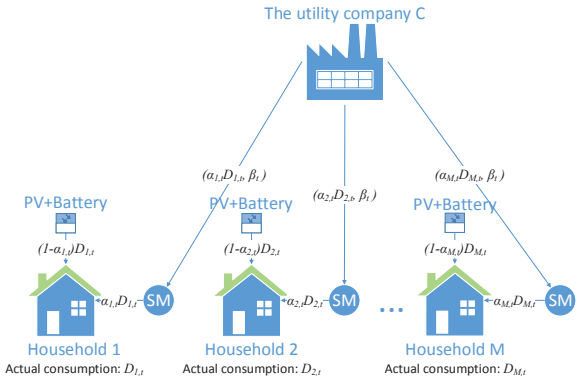


Figure 1.1: Household-utility company interaction diagram

The profit of the utility company when supplying $\sum_{i\in\mathcal{H}} \alpha_{i,t}D_{i,t}$ amount of energy to households is given by $g(\sum_{i\in\mathcal{H}} \alpha_{i,t}D_{i,t})$. We assume that $g(\sum_{i\in\mathcal{H}} \alpha_{i,t}D_{i,t})$ is an increasing, continuous and concave function. Thus, the reward of the utility company can be expressed by:

$$V_t(\beta_t) = g(\beta_t \sum_{i\in\mathcal{H}} \alpha_{i,t}D_{i,t}) - \beta \sum_{i\in\mathcal{H}} \alpha_{i,t}D_{i,t} - L(X_t - \sum_{i\in\mathcal{H}} \alpha_{i,t}D_{i,t}). \tag{1.2}$$

To maximize its return from sharing consumption data with the utility company, household i solves the following optimization problem:

$$\max_{\alpha\in\mathcal{A}_t} \quad U_{i,t}(\alpha_{i,t}) = \beta_t \alpha_{i,t}D_{i,t} \tag{1.3}$$

$$s.t. \qquad \beta_t = f_{i,t}(\alpha_{i,t}D_{i,t}). \tag{1.4}$$

The objective of the utility company is to choose the price β_t such that it will maximize its reward from incentivizing households to share at least a minimal amount of their energy consumption profile data. Thus, the utility company solves the following optimization problem:

$$\max_{\beta_t\in\mathcal{B}} \quad V_t(\beta_t) = g(\beta_t \sum_{i\in\mathcal{H}} \alpha_{i,t}D_{i,t}) - \beta \sum_{i\in\mathcal{H}} \alpha_{i,t}D_{i,t} - L(X_t - \sum_{i\in\mathcal{H}} \alpha_{i,t}D_{i,t}). \tag{1.5}$$

1.5 The Household-Utility Company Game

For a given set of energy demand $(D_{1,t}, D_{2,t}, ..., D_{M,t})$ at time t, the amount of energy $\alpha_{i,t}D_{i,t}$ that each household i decides to consume from the grid and the incentive price β_t strongly impact both the clearing price as well as the rewards for both households and the utility company. Moreover, the strategy for each household also affect other households' strategies indirectly via influencing the strategy of the utility company. To model this privacy-energy cost tradeoff, we develop a scheme based on non-cooperative game theory. The household chooses the amount of energy to consume from the grid (its strategy) at time t; on the other hand, the strategy for the utility company is to determine the incentive price to compensate for the privacy loss of the consumer. We assume that the strategies $\alpha_{i,t}$ and β_t are selected from convex, closed, and bounded sets supported on $\mathbb{R}$. Thus, this problem can be modeled as an N-player strategic game $(N = M + 1)$ with components given as follows:

- Set of players: $\{(\mathcal{H}, C)\}$ is the set of players in which households belong to set $\mathcal{H}$ and the utility company is denoted by C.
- Set of strategies: $\{(\{\mathcal{A}_i\}_{i\in\mathcal{H}}, \mathcal{B})\}$ is the tuple of strategy sets for households and the utility company, where the strategy for households i (consumption coefficient $\alpha_{i,t}$) is given in $\mathcal{A}_i$ and the strategy of the utility company C (incentive price β_t) belongs to $\mathcal{B}$.
- Payoff functions: $\{(\{U_{i,t}(\cdot)\}_{i\in\mathcal{H}}, V_t(\cdot))\}$ is the tuple of payoff functions in which we denote $U_{i,t}(\cdot)$ to be the reward for household i and $V(\cdot)$ to be the reward for the utility company C.

The resulting strategic game is written as $\{(\mathcal{H}, C), (\{\alpha_{i,t}\}_{i\in\mathcal{H}}, \beta_t), (\{U_{i,t}\}_{i\in\mathcal{H}}, V_t)\}$. Such a game has one well-studied solution which is called the Nash equilibrium (NE). The NE is a strategy tuple in which none of the players can be more profitable by unilaterally deviating from this equilibrium strategy. Formally, it is defined as follows:

Definition 1. *Consider the N-player strategic game $\{(\mathcal{H}, C), (\{\alpha_{i,t}\}_{i\in\mathcal{H}}, \beta_t), (\{U_{i,t}\}_{i\in\mathcal{H}}, V_t)\}$, a strategy tuple $(\{\alpha^*_{i,t}\}_{i\in\mathcal{H}}, \beta^*_t)$ is an NE if and only if*

$$U_{i,t}(\alpha^*_{i,t}, \boldsymbol{\alpha}^*_{-\mathbf{i},t}, \beta^*_t) \geq U_{i,t}(\alpha_{i,t}, \boldsymbol{\alpha}^*_{-\mathbf{i},t}, \beta^*_t) \qquad \forall \alpha_{i,t} \in \mathcal{A}_i, i \in \mathcal{H}$$

and

$$V_t(\alpha^*_{i,t}, \boldsymbol{\alpha}^*_{-\mathbf{i},t}, \beta^*_t) \geq V_t(\alpha^*_{i,t}, \boldsymbol{\alpha}^*_{-\mathbf{i},t}, \beta_t) \qquad \forall \beta_t \in \mathcal{B}.$$

where the vector $\boldsymbol{\alpha}_{-\mathbf{i},t}$ denotes the strategies of all other households.

The NE in our context defines a strategy tuple in which neither the households nor the utility company can be more profitable by unilaterally deviating from the equilibrium choice of other households and the utility company. It presents a stable outcome of the interaction between households and the utility company.

Theorem 1. *There exists at least one Nash equilibrium for the above household-utility company game.*

Proof. The reward function is continuous and concave for both households and utility company. Also, the strategy set is convex, closed and bounded. Thus, the above household-utility company game is a concave N-person game, where $N = M + 1$ in our context. Therefore, the existence of Nash equilibrium is guaranteed by continuous and concave reward functions and convex feasible set for each player [15]. □

To find the NE for our household-utility company game, we first introduce the notion of best response. The best response is a function which captures the behavior of each player by making other players strategies fixed. By Definition 1, in every NE, each player plays the best response with respect to other players strategies.

Definition 2. *The best response* $R_{i,t}(\boldsymbol{\alpha}_{-\mathbf{i},t}, \beta_t)$ *of a household* i *to all the other players strategies is a set of strategy such that* $R_{i,t}(\boldsymbol{\alpha}_{-\mathbf{i},t}, \beta_t) = \{\alpha_{i,t} \in \mathcal{A}_i | U_{i,t}(\alpha_{i,t}, \boldsymbol{\alpha}_{-\mathbf{i},t}, \beta_t^*) \geq U_{i,t}(\alpha'_{i,t}, \boldsymbol{\alpha}_{-\mathbf{i},t}, \beta_t), \forall \alpha'_{i,t} \in \mathcal{A}_i\}$. *Similarly, we define the best response of the utility company to be a set of price* β_t *which is given by* $R_{C,t}(\boldsymbol{\alpha}_{\mathbf{i},t}) = \{\beta_t \in \mathcal{B} | V_t(\boldsymbol{\alpha}_{\mathbf{i},t}, \beta_t) \geq V_t(\boldsymbol{\alpha}_{\mathbf{i},t}, \beta'_t), \forall \beta'_t \in \mathcal{B}\}$.

Given above assumptions on payoff functions, the reward of household i is increasing with respect to $\alpha_{i,t}$. Thus, the best response of household i given other players strategies fixed can be expressed as

$$\alpha_{i,t} D_{i,t} = f_{i,t}^{-1}(s_{i,t}) = f_{i,t}^{-1}(\beta_t). \tag{1.6}$$

The objective of the utility company is to maximize its reward from incentivizing households to share their consumption profile data. Thus, its best response is given as

$$\beta_t = \arg\max_{\beta_t \in \mathcal{B}} V_t(\beta_t) \tag{1.7}$$

$$= \arg\max_{\beta \in \mathcal{B}} g(\beta \sum_{i \in \mathcal{H}} \alpha_{i,t} D_{i,t}) - \beta \sum_{i \in \mathcal{H}} \alpha_{i,t} D_{i,t} - L(X_t - \sum_{i \in \mathcal{H}} \alpha_{i,t} D_{i,t}). \tag{1.8}$$

Since the existence of NE in the household-utility company game mentioned above is guaranteed by Theorem 1, the strategies of the utility company and households reach an NE if and only if all of their best responses intersect at the same strategy $(\boldsymbol{\alpha}_t^*, \beta_t^*)$. Assume that the first order derivatives for $g(\sum_{i \in \mathcal{H}} \alpha_{i,t} D_{i,t})$, $L(X_t - \sum_{i \in \mathcal{H}} \alpha_{i,t} D_{i,t})$, and $f_{i,t}^{-1}(\beta_t)$ exists. The NE is given by solving the following equations:

$$\begin{cases} \alpha_{i,t} = \frac{f^{-1}(\beta_t)}{D_{i,t}} \\ \frac{\partial V(\beta_t)}{\partial \beta_t} = 0 \end{cases}, \tag{1.9}$$

where $V_t(\beta_t)$ is defined in (1.2).

Next, we propose an iterative algorithm to compute the NE. During the n^{th} iteration, household i selects its best response with respect to the strategy of the utility company. Meanwhile, the utility company chooses its incentive β_t^n to maximize its reward function given the response of all households. This iterative process continues until β_t^n converges. In general, best response dynamics-based algorithms have been proven to converge to an NE for many classes of non-cooperative games [16]. However, for general non-cooperative games, the existence of an NE is not always

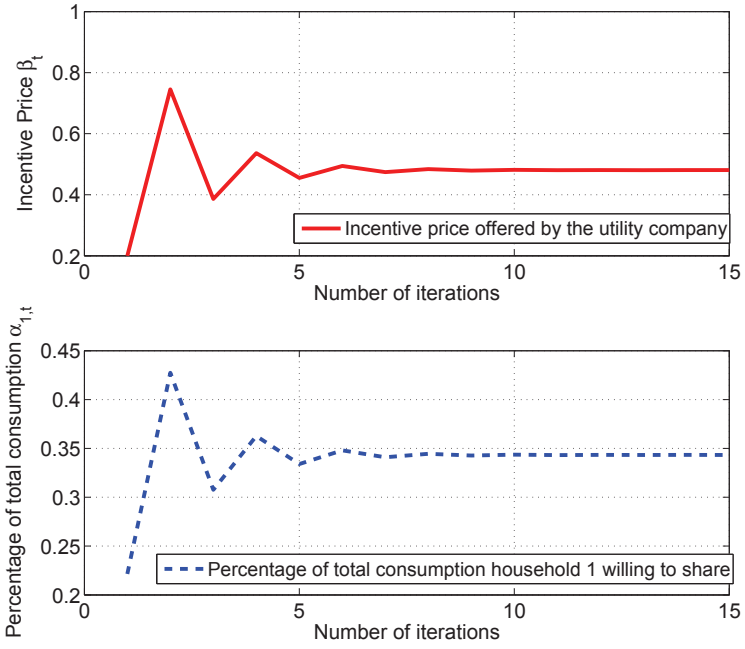


Figure 1.2: Best response of the utility company and household 1

guaranteed. In the proposed household-utility company game, the privacy valuation function $f_{i,t}(\alpha_{i,t}D_{i,t})$, which is based on households subjective behavior, might not be continuous. Thus, it can introduce a discontinuity. As a result, it is difficult to prove the existence of the NE analytically [15]. In cases of non-existence, the utility company announce a final clearing price β_t based on the observed best responses. Also, in general, there may be multiple NE solutions; however, we only focus on one of the NE since it is a practical outcome of the interaction between households and the utility company.

Algorithm 1 Iterative process to find Nash equilibrium

Begin

Initialization : Set $\beta_t^0 = \beta_0 > 0$.

Do

$\alpha_{i,t}^n D_{i,t} = f_{i,t}^{-1}(\beta_t^{n-1})$

$\beta_t^n = \arg\max_{\beta \in \mathcal{B}} g(\beta \sum_{i \in \mathcal{H}} \alpha_{i,t}^n D_{i,t}) - \beta \sum_{i \in \mathcal{H}} \alpha_{i,t}^n D_{i,t} - L(X_t - \sum_{i \in \mathcal{H}} \alpha_{i,t}^n D_{i,t})$

Until

$\beta_t^n - \beta_t^{n-1} < \epsilon$ for $\epsilon > 0$ and sufficiently small.

Set

$\beta_t = \beta_t^n$

$\alpha_{i,t} = \frac{f_{i,t}^{-1}(\beta_t)}{D_{i,t}}$

The NE is given by $(\boldsymbol{\alpha}_t, \beta_t)$.

End

To illustrate the performance of the proposed threshold policy, we assume the average roof top PV size in each household is $35m^2$ with 12% conversion efficiency

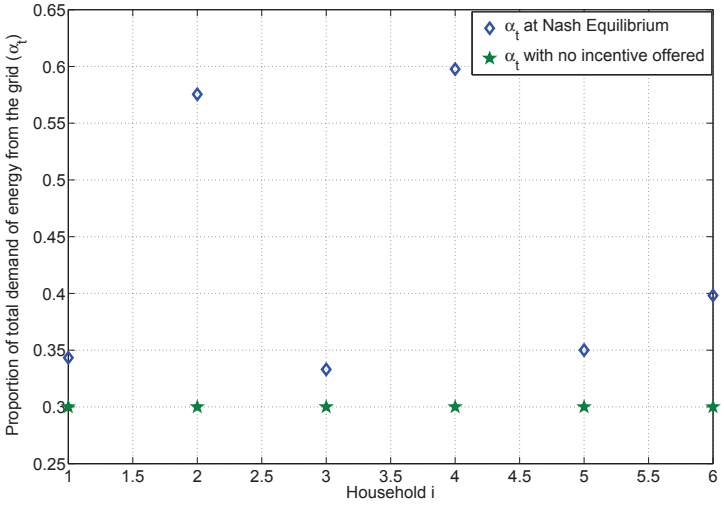


Figure 1.3: Proportion of total energy demand consumed from the grid at a Nash equilibrium

and $1,800kWh/m^2$ per year of available sunlight energy [17]. The average household energy in US is $909KWh/month$ [18]. Therefore, the PV system can provide approximately 70% of total energy consumption of a household. Furthermore, we assume

$$V_t(\beta_t) = \delta\sqrt{\beta_t \sum_{i\in\mathcal{H}} \alpha_{i,t} D_{i,t}} + \gamma - \theta * (X_t - \sum_{i\in\mathcal{H}} \alpha_{i,t} D_{i,t})^2 - \beta_t \sum_{i\in\mathcal{H}} \alpha_{i,t} D_{i,t} \qquad (1.10)$$

and

$$f_{i,t}^{-1}(\beta_t) = \eta_i \sqrt{\beta_t} \qquad \forall i \in \mathcal{H} \qquad (1.11)$$

, where δ, γ, η and, θ are system parameters determined by the utility company and households. The demand $D_{i,t}$ from each household is uniformly drawn from $[2, 4]$ and $X_t \in \mathcal{N}(2M, M)$The best response dynamics of household 1 and the retailer is depicted in Figure 1.2. The households and the utility company reaches the NE after 13 iterations. We assume that the size of the battery is large enough. Thus, the privacy preserving algorithm [7] can achieve perfect privacy by making the consumption profile flat. Since the PV system can only supply 70% of total energy consumption, the household have to consume the other 30% from the grid. Thus, $\alpha_{i,t} = 0.3$ if $\beta_t = 0$. Figure 1.3 shows the strategies of different households at a Nash equilibrium. One can observe that for the same incentive price offered by the utility company, households with low valuation on privacy (Household 2 and 4) tend to increase their consumption more than household with high valuation on privacy.

1.6 Conclusion and Future Work

In this project, we introduced a novel approach to study the tradeoff between privacy and energy cost minimization under the assumption that the utility company offers incentives to households to encourage data sharing through energy consumption. We formulated a non-cooperative game to model interactions between households and the utility company. We proved that the existence of a Nash equilibrium is guaranteed if the strategy sets and payoff functions satisfy certain concavity properties. To solve the household-utility company game, we have proposed an iterative best response algorithm that leads to a Nash equilibrium. Our simulation results have shown that the consumption of each household increases when the utility company offers an incentive price to households. Our work suggests several interesting future directions: one straightfoward extension of this work is to develop dynamic game models to capture the interaction between households and the utility company over a certain time period. Another interesting problem is to study how will this game theoretic model work when households can implement privacy preserving control policies.

Chapter 2

Sharing AMI Data with Limited Statistical Inferences

2.1 Introduction

Smart meters have been used in both homes and enterprises for energy usage collection and monitoring. Both utility company and households benefit from smart meter. However, the data collected from smart meter may give rise to serious privacy concerns since an adversary can make inference on user behavior. For instance, to decide the best time for burglary, a burglar has to monitor activities in the target household for a long time to obtain living pattern of people in that house. However, by monitoring the energy consumption from the target household, the burglar can easily infer living patterns. The data collected by the utility also may be more than they needed for system monitoring.

In this project, we seek to model behavior of consumers with privacy concerns in the smart grid and study the tradeoff between cost saving and privacy. To be more specific, we study the tradeoff from using battery/PV to achieve energy cost savings versus using them specifically for retaining a certain measure of privacy. We assume the privacy feature that the consumer wants to hide from the utility company to be a binary variable, i.e., a yes/no inference on a certain feature (e.g., consumer at home or not). We use information theory as a privacy metric that captures consumer's concerns, e.g., can the data collector infer specific personal habits? Furthermore, we try to develop optimal policy for consumers to minimize cost while retaining a certain measure of privacy.

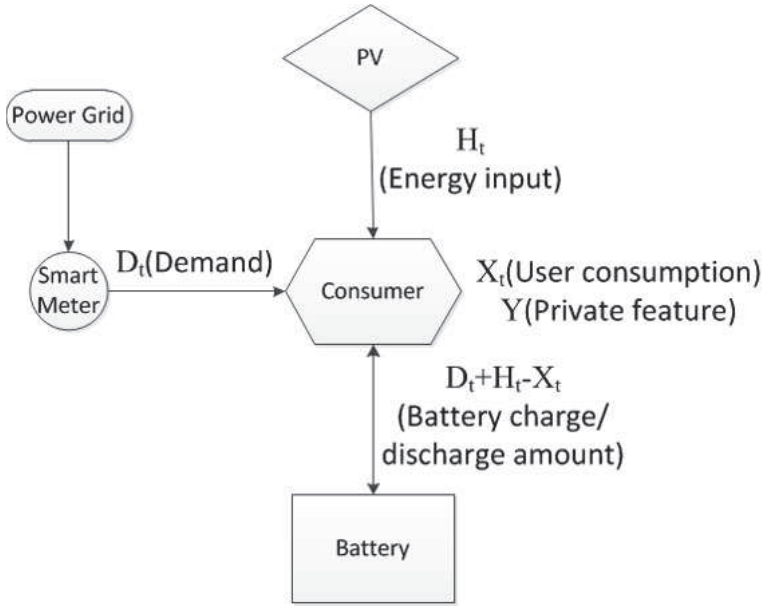


Figure 2.1: System diagram

2.2 System model

As shown in Figure 2.1, the electricity of a household is supplied by coordination of energy from the power grid, alternative energy sources and the battery. Appliances reveal user activity by coming ON or going OFF. User activity can then be correlated with specific appliance detected (from signature) and personal habits/preferences. We assume Y to be the private feature that the consumer wants to hide, e.g., appliance ON and OFF state. The actual electricity that the household consumes at time t is given by an i.i.d. random variable X_t. The energy provided by an alternative energy resource (PV) and the power grid is denoted by i.i.d. random variables H_t and D_t, respectively. We assume that the capacity of the battery is large enough to serve the load. Thus, the energy provided by the battery is given by $D_t + H_t - X_t$. The energy consumption from the grid can be written as follows:

$$D_t = F(X_t, y, H_t) \tag{2.1}$$

Since the battery size is large enough, asymptotically, for any battery control policy, the demand must satisfy:

$$\mathbb{E}(D_t|Y) + \mathbb{E}(H_t) = \mathbb{E}(X_t|Y). \tag{2.2}$$

We assume a strong adversary who has knowledge of $P_{X|Y}(X_t|Y)$, $P_H(H_t)$ and $P_{X,H|Y}(X_t, H_t|Y)$ but only have access to D_t in real time. Let y_0 and y_1 denote that the privacy feature is on or off, respectively. The adversary can infer the private action Y by making a guess $\hat{Y}$ based on the Neyman-Pearson hypothesis testing approach (Figure 2.2). For a group of sampled smart meter data $d^N = (d_1, ..., d_N)$ with length N, the optimal region of making a guess $Y = y_0$ is given by

$$\mathcal{A}_{y_0} = \{d^N : \frac{P_{D^N|Y}(d^N|Y = y_0)}{P_{D^N|Y}(d^N|Y = y_1)} \geq T\}, \tag{2.3}$$

where T is the threshold set by the adversary. By asymmetrical equipartition property of relative entropy, the above region is equivalent to

$$\mathcal{A}_{y_0} = \{d^N : \frac{P_{D^N|Y}(d^N|Y = y_0)}{P_{D^N|Y}(d^N|Y = y_1)} \to 2^{ND(P_{D^N|Y}(d^N|Y=y_0)||P_{D^N|Y}(d^N|Y=y_1))} \geq T\} \tag{2.4}$$

If $Y = y_1$, the consumer wants to use D_t to make the adversary thinks $Y = y_0$. Therefore, the privacy of the consumer is preserved if the decision of the adversary falls into "Not detected" in Figure 2.2. Thus, the probability that the consumer's privacy has been preserved can be written as:

$$p^{ud} = \sum_{d_t \in \mathcal{A}_{y_0}} P_{D^N|Y}(d^N|Y = y_1), \tag{2.5}$$

where $\mathcal{A}_{y_0}$ is the set of demand in which the adversary's estimation of Y is y_0. Define the false alarm probability to be

$$p^{fa} = \sum_{d_t \in \mathcal{A}_{y_0}^C} P_{D^N|Y}(d^N|Y = y_0), \tag{2.6}$$

		True state of nature	
		y_0	y_1
Conclusion from observation	y_0	Correct	Not detected
	y_1	False alarm	Correct

Figure 2.2: Neyman-Pearson hypothesis testing

where $A_{y_0}^C$ denotes the region of d_t in which the adversary makes a guess that $Y = y_1$. By Chernoff-Stein Lemma, for $0 < \epsilon < \frac{1}{2}$, define

$$\beta = \min_{d_t \in \mathcal{A}_{y_0}^C, p^{fa} < \epsilon} p^{ud}. \tag{2.7}$$

Then, $\lim_{N\to\infty} \frac{1}{N} log\beta = -D(P_{D^N|Y}(d^N|Y = y_0)||P_{D^N|Y}(d^N|Y = y_1))$. Therefore, as $N \to \infty$, we have $\beta \to 2^{-N(D(P_{D^N|Y}(d^N|Y=y_0)||P_{D^N|Y}(d^N|Y=y_1)))}$.

We assume the energy cost $C(d^N)$ is a convex function. The consumer's objective is to minimize its expected energy consumption cost, while retain certain level of privacy. Thus, the optimization problem is formulated as follows:

$$\begin{aligned} &\min_{p_{D|Y}} \mathbb{E}(C(d^N)) \\ &s.t. \quad p_{D|Y} \in \Pi_{D|Y} \\ &\qquad \Pi_{D|Y} = \{p_{D|Y} : \mathbb{E}(D_t|Y) + \mathbb{E}(H_t) = \mathbb{E}(X_t|Y)\} \\ &\qquad V2^{-ND(P_{D^N|Y}(d^N|Y=y_0)||P_{D^N|Y}(d^N|Y=y_1)} \geq R, \end{aligned} \tag{2.8}$$

where $D(\cdot)$ is the Kullback-Leibler distance, V is the loss incurred by privacy leakage (adversary successfully infer the private feature), and R is the level of privacy that the consumer wants to retain. Since $\{p_{D|Y} : V2^{-N(D(P_{D^N|Y}(d^N|Y=y_0)||P_{D^N|Y}(d^N|Y=y_1))} \geq R\}$ and $\Pi_{D|Y}$ are convex sets, the above problem is a convex optimization problem. This problem can be interpreted as a cost minimization problem given the minimum expected privacy utility is guaranteed to be above some value R.

2.3 Expected Results

The solution of the above optimization problem will give the consumer a randomized control policy of the energy needed from the grid. The proposed control policy assigns probabilities to each action based on energy from PV, consumer consumption and private feature. The resulted demand profile guarantees the consumer certain level of privacy while minimizing the energy cost to the consumer. Meanwhile, the utility company gets precise knowledge of the amount of energy needed from the power grid (Good for monitoring and pricing). However, based on the smart meter readings, the adversary has limited capability to infer the private feature.

Chapter 3

Effect of AMI Cyber-Attacks at the Transmission Level

3.1 Introduction

The electric grid is a complex physically distributed and inter-connected network managed by a large number of agents/entities (*e.g.*, systems operators, utilities, transmission operators) to ensure reliable transmission, generation, and distribution of power. Sustained and reliable operation with dynamic situational awareness in the grid requires continued interactions and data sharing amongst the grid entities. In fact, it has been noted that while a number of physical factors are responsible for local outages, lack of automated communications and coordination between distributed operators in the grid contributes significantly to the lack of *global situational awareness* leading to runaway cascading failures [19, 20]. The grid is fast converging towards a Smart Grid characterized by (a) vastly expanded data acquisition, (b) highly variable environments due to integration of renewables, and (c) distributed processing and control. In this new paradigm, timely and controlled information exchange is critical not only to ensure reliability and stability but also to thwart cyber attacks that could potentially bring down the entire grid with one or more local outages.

In this report, we focus on a class of topology-targeted man-in-the-middle (MitM) communication attacks aimed at limiting information sharing between adjacent areas, particularly when one or both areas experience topology changes (*e.g.*, line outages). While wide-area monitoring and information sharing has been proposed by the Federal Energy Regulatory Commission based on observations that lack of seamless data sharing is an important factor in cascading failures, real-time data sharing in the grid is still done in an ad hoc manner between connected areas. For example, in the Northeast blackout of 2003 [19], [20], a line out in one area (Ohio) was not conveyed for a sufficient period of time to neighboring regions leading to convergence failure of the state estimator and other cascading problems. Furthermore, the mode, amount, and granularity of data shared is not standardized; for example, two connected areas may only share limited topology information such as low granularity network equivalent models which in turn are insufficient to capture the complexity of the electric grid and ensure wide-area reliability (*e.g.*, the Yuma-Southern California outage of 2011 [21]). In fact, changes in the grid topology are often communicated via human operators and not in an automated manner which adds to communication delays and errors. In the light of such limitations, a smart adversary can limit information sharing in a number of ways. We seek to understand the effects of such limited data sharing scenarios (both adversarial and otherwise) on the electric power system real-time operations.

We introduce a class of distributed communication attacks wherein an attack on the Energy Management System (EMS) of one area prevents the sharing of topology changing information with the other area (in automated systems where topology may be shared real-time or frequently, this can be achieved via MitM attacks). We assume that the attacker is either involved in bringing down a line remotely (breakers can be remotely tripped in some cases) or is aware of a line out (again possible via presence of software trojans in the EMS). The attacker, therefore, is assumed to have some knowledge of the network topology.

There has been much recent interest on cyber attacks on the grid, in particular false data injection (or integrity) attacks, where as the name suggest false data is

introduced in specific measurement and computing units of the EMS such as state estimation (SE) (*e.g.*, [22, 23]), automatic generation control (*e.g.*, [24]), generator frequency control (*e.g.*, [25]), topology processing (*e.g.*, [26]), as well as attack consequences on markets (*e.g.*, [27, 28]). However, the consequences of such attacks on the system are yet to be demonstrated. While changes in locational marginal prices could demonstrate the effect on prices of unobservable attacks, an important question that remains to be addressed is whether serious damage such as instability, cascading failures, and potential blackouts, which can cripple society and the economy, can be caused by cyber attacks on the grid.

To understand the broader consequences of MitM attacks on measurements or shared data, we develop a layered systems model that enables the modeling of the time progression of attacks. In [29], Liang *et al.* introduced a time progression based modeling to demonstrate how an unobservable false data injection attack on AC SE, by a sophisticated attacker who is assumed to be capable of performing AC SE in a small subgraph of the network, can lead to a physical generation dispatch when none was needed. In this report, we focus on a distributed two-area (managed by two operators and EMSs) setting to demonstrate the consequences of limited information sharing. Specifically, we focus on attacks that create or exploit outages in one area and limit information sharing via a communication attack thereby affecting the power flow solutions and dispatch in a connected area that has incorrect topology information. Specifically, we modeled the tie-lines connected the two areas under two conditions: (a) in normal operation, the tie-line interchange is fixed according to the day-ahead contract between areas, we simulate with only 10% variation on tie-line power flow interchange; and (b) under contingencies, the tie-line power flow can vary any values under the tie-line capacity, we then simulate with no tie-line interchange variation limitation. Our results demonstrate that such an attack in a distributed power network leads to a range of possibilities; these include actual physical line overloads that are not observable from the cyber measurements but can eventually cause line overheating and cascading outages; false overload alert in cyber layer while the physical system operates in normal condition; progressively severe lack of convergence of OPF in both areas; relatively benign oscillations in the power flow solutions between the two areas that eventually fix themselves; and line overloads in both physical can cyber layers. Our time progression based attack model allows us to capture the major computational components of EMSs including AC SE, optimal power flow (OPF) including generation dispatch, and power flow calculation unit which adjusting dispatch mismatch between areas. Based on our observations, we also present countermeasures for such attacks.

3.2 System Model

We consider a two-area network model in which each area uses its measurements to evaluate the state of the system, compute the optimal power flow, and determine generation dispatch. It is worth noting that there are many control and actuation functions that operate at multiple timescales in the EMS and not all of them are captured by our model. Our choice of functions is driven by a specific time-scale that focuses on topology processing and state estimation, power flow computation, and generation dispatch. It is assumed that the computations are performed at a local control center as shown in Figure 3.1, and henceforth, when we refer to the two areas sharing information, it implies that information is exchanged between the control centers. We make the following assumptions about the information shared between the two areas.

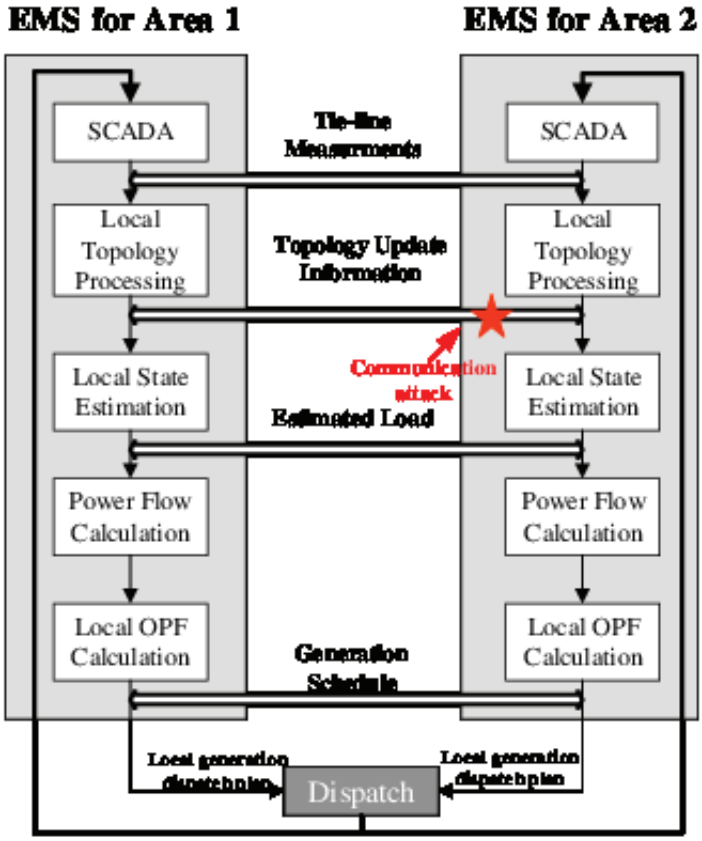


Figure 3.1: Computational Units and Data Interactions between the Two Areas of the Network.

3.2.1 Information Sharing Model

To illustrate the distributed effects of a communication attack, we assume that the two areas share as much information as relevant and based on current practices. The primary assumption is that each area performs its own computations with some data (depending on the computation block) obtained from the other side. Our assumptions are as follows:

- Static topology information: The static topology information is shared among all areas of the interconnected power system.
- Dynamic topology information: Each area is assumed to communicate the topology changing information among the whole system in real-time. Thus, once a topology error is found, the local operator should send this information to other areas immediately, which allows them to update the whole system topology information in time.
- Generation: The generation schedule of each unit is shared among areas in real-time.
- Measurements: The tie-line measurements are shared between adjacent areas in real-time. In general, more measurements can be shared but we assume that each area does its own local SE (as is often the case in practice).
- Estimated load: The estimated load data is shared among areas in real-time.
- Network models for power flow: Each area computes its own AC OPF. In practice, each area uses a network equivalent model of its connected areas to simplify the OPF computation. However, since we seek to understand the effect of a communication cyber-attack on dispatch and power flows (line overloads often contribute to outages), we choose the best case network model, *i.e.*, we assume each area uses the complete network model of the other side in computing its OPF. However, each area can only dispatch its own generators, and thus, computes the OPF while keeping the dispatch for the other area fixed according to the generation data sharing model.

3.2.2 Computational Models

We briefly outline the mathematical model for each of the computational units we consider here. The different computational units and their interactions across the two areas is illustrated in Figure 3.1.

State Estimation

Each area estimates its system state (complex voltages) using the measurements from meters in its area as well as tie-line measurements. As is the norm, we assume the use of a weighted least-squares (WLS) AC state estimation to calculate the voltage angles and magnitudes (assuming a voltage angle reference). The objective of the estimation process is to minimize the sum of the squares of the weighted deviations of

the estimated variables from the actual measurements. The non-linear measurement model is given by

$$z = h(x) + e \tag{3.1}$$

where z, e and x are $M \times 1$ measurement, $M \times 1$ noise, and $N \times 1$ state vectors with entries z_i, e_i and x_k, respectively, for $i \in \{1, ..., M\}$ and $k \in \{1, ..., N\}$. The function $h(.)$ is a vector of nonlinear functions describe the relationship between states and measurements, and e_i is assumed to be independent and Gaussian distributed with 0 mean and σ_i^2 covariance such that the measurement error covariance matrix is given by $R = diag(\{\sigma_i^2\}_{i=1}^{M})$. The commonly obtained measurements in the grid are the active and reactive power flows and node injections. In AC state estimation, the state variables are solved as a least square problem with the following objective function [30]:

$$\min J(x) = (h(x) - z)^T R^{-1} (h(x) - z), \tag{3.2}$$

the solution to which satisfies

$$g(\hat{x}) = \frac{\partial J(\hat{x})}{\partial x} = H^T(\hat{x}) \cdot R^{-1} \cdot (h(\hat{x}) - z) = 0 \tag{3.3}$$

where $H = \frac{\partial h(x)}{\partial x} \mid_{x=\hat{x}}$. The WLS solution for this non-linear optimization problem can be solved iteratively.

Power Flow Calculation

Each area uses Newton's method to solve the power flow (PF) problem which involves solving for the set of voltages and flows in a network corresponding to the estimated loads and generation schedule obtained from OPF in previous time period. This unit is to adjust the overall load and generation mismatch caused by joint dispatch of two areas.

Optimal Power Flow

Assuming perfect network equivalent models (*i.e.*, complete sharing of neighboring network graphs for OPF), area i, $i = 1, 2$, runs its OPF with the dispatch for area j, fixed around values that were shared from the previous time period that area j ran its own OPF. The resulting OPF problem can be viewed as each area performing a centralized power flow problem but with the capability to only dispatch local units.

Let B and Br denote the set of buses and branches in the entire two-area network, and B_i and B_j denote the set of buses in area i, $i = 1, 2$, and area j, $j = 1, 2$, $j \neq i$, respectively. Futhermore, let G_n denotes the set of generators at bus n, $\{G_n\}_{n \in B_i}$ denotes the set of generators in area i, $i = 1, 2$. Let $c_g(\cdot)$ to denote the generation cost function for generator g. The OPF for each area can be formulated as the following optimization problem for area i, $i = 1, 2$:

$$\min \sum_{g \in \{G_n\}_{n \in B_i}} c_g(P_g) \tag{3.4}$$

$$s.t. \sum P_g + \sum_{\forall k(n,;)} P_k - \sum_{\forall k(;,n)} P_k = P_{dn}, \quad \forall n \in B, \tag{3.5}$$

$$\sum_{g \in G_n} Q_g + \sum_{\forall k(n,;)} Q_k - \sum_{\forall k(;,n)} Q_k = Q_{dn}, \quad \forall n \in B, \tag{3.6}$$

$$P_k = V_n^2(g_{sn} + g_{nm}) - V_n V_m(g_{nm} cos(\theta_n - \theta_m) + b_{nm} sin(\theta_n - \theta_m)), \ k \in Br \tag{3.7}$$

$$Q_k = -V_n^2(b_{sn} + b_{nm}) - V_n V_m(g_{nm} sin(\theta_n - \theta_m) - b_{nm} cos(\theta_n - \theta_m)), \ k \in Br \tag{3.8}$$

$$P_k^2 + Q_k^2 \leq (S_k^{max})^2 \quad \forall k \in Br \tag{3.9}$$

$$P_g^{min} \leq P_g \leq P_g^{max} \quad \forall g \in \{G_n\}_{n \in B_i} \tag{3.10}$$

$$Q_g^{min} \leq Q_g \leq Q_g^{max} \quad \forall g \in \{G_n\}_{n \in B_i} \tag{3.11}$$

$$V_n^{min} \leq V_n \leq V_n^{max} \quad x \in \{\theta, V\}, \forall n \in B \tag{3.12}$$

$$\hat{P}_g - \triangle \bar{P}_g \leqslant P_g \leqslant \hat{P}_g + \triangle \bar{P}_g \quad \forall g \in \{G_n\}_{n \in B_j} \tag{3.13}$$

$$\hat{Q}_g - \triangle \bar{Q}_g \leqslant Q_g \leqslant \hat{Q}_g + \triangle \bar{Q}_g \forall g \in \{G_n\}_{n \in B_j} \tag{3.14}$$

where $c_g(.)$ is the cost function for generator g, b_{nm} and g_{nm} are the susceptance and conductance, respectively, of line k from bus n to bus m, b_{sn} and g_{sn} are the shunt branch susceptance and conductance, respectively, of bus n, $k(n,;)$ is the set of lines k with bus n as its receiving bus, and $k(;,n)$ is the set of lines k with bus n as its sending bus, G_n is the set of generators at bus n, P_g is the active power output of generator g with maximum and minimum limit P_g^{max} and P_g^{min}, Q_g is the reactive power output of generator g with maximum and minimum limit Q_g^{max} and Q_g^{min}, $\hat{P}_{gj}$ and $\hat{Q}_{gj}$ are the fixed active and reactive power outputs with $\triangle \bar{P}_g$ and $\triangle \bar{Q}_g$ deviations of generator g in area j, respectively, P_k and Q_k are the active and reactive power flows, respectively, on line k with line capacity limit S_k^{max}, P_{dn} and Q_{dn} are the active and reactive power demands, respectively, at bus n, θ_n is the voltage angle for bus n, and V_n is the voltage magnitude for bus n with maximum and minimum limits V_n^{max} and V_n^{min}, respectively.

The objective in (3.4) is to minimize the total active power generation cost of the whole interconnected power system. Constraints (3.5) and (3.6) represent the active and reactive power balance for each bus in the centralized system (two-area network). The constraints in (3.7) and (3.8) are the active and reactive transmission line power flow constraints for the whole system while (3.9) is the thermal limit for each transmission line. Constraints (3.10) and (3.11) are the local (for area i only) unit active and reactive power output limits while (3.12) defines the complex voltage stability limits for each bus in the whole system. Finally, (3.13) and (3.14) incorporate the unit active and reactive power output limits for area j, $j \neq i$, *i.e.*, the power output of generation units external to area i are fixed around the values shared by the other areas.

When no feasible solution (*i.e.*, a solution which satisfies (3.5)-(3.14)) can be found, the distributed OPF program fails to converge. In practice, to find a feasible solution, system operators often relax the constraints. In this report, the thermal limit constraint on the congested line is the first constraint to be relaxed. Multiple iterations

of relaxing the line limits may be needed to obtain a feasible solution; to this end, we model the relaxed limits as follows:

$$P_k^2 + Q_k^2 \leq (S_k^{\max} + u\triangle\bar{S}_k)^2$$

where line k is the congested line, $\triangle\bar{S}_k$ is the incremental value by which the line limit is relaxed in each iteration, and $u \leqslant u_{\max}$ is the iteration number. In each iteration, the thermal limit is relaxed by increasing the rating of line k by, and the OPF program is executed to check whether it converges. This process is repeated until the OPF program converges or the relaxation time reaches its maximum value. Following this, other important lines (such as those with high reactive power flow) will be relaxed using the same procedure. If both methods fail to work, then we consider the test case as a not converge case.

3.3 Attacker Model

3.3.1 Time Progression Model of Attack

We assume that the attacker has access to the data being shared between areas and can corrupt the data. Examples abound of such data corruption attacks including the oft cited Stuxnet virus attack. The attacker is assumed to either participate in creating a line outage in one area or be aware of such an outage and then act to corrupt the topology information shared with the other area. Our attack model also captures simple human errors in information sharing between connected areas, including delays and mis-communications. In the interest of understanding worst-case attacks and data sharing limitations, the area with the outage is assumed to be aware of the outage shortly after. This assumption is based on frequently seen patterns of limited data sharing that precede (and are a cause of) large blackouts.

In order to understand the effect of such an attack, we study the time progression of the attack. We consider the following time-progression of the attack and system behavior includes the following steps:

1. Event 0: *Area i*: Outage occurs in Area i, $i = 1, 2$. Area i becomes aware of outage and updates its topology and shares with Area j. Area i then performs SE, PF, and OPF.
2. Event 0: *Attack*: Attacker replaces updated topology information shared with area j, $j \neq i, j = 1, 2$, with the previous static topology information.
3. Event 0: *Area j*: Area j uses measurements with updated topology (which has been changed by attacker) to compute SE, PF, and OPF.
4. Event 1: *System*: Area i and Area j jointly dispatch according to their own OPF results.
5. Event 1: *Area i*: Area i uses measurements with updated topology to compute SE, PF, and OPF. Shares dispatch status with Area j. Attacker sustains attack.
6. Event 1: *Area j*: Area j uses measurements with updated topology (which has been changed by attacker) to compute SE, PF, and OPF. Shares dispatch status with Area i.
7. Events repeat until alarms are set off either due to repeated lack of convergence or physical line overloads. All the while it is assumed that the attacker sustains the attack.

We illustrate this time sequence in Fig 3.2 for the case in which Area 1 experiences a line outage while Area 2 does not have the real-time topology information following the outage due to a communication attack.

3.3.2 Tie-line Agreement Assumption

In real-time operation, the tie-line interchange is fixed according to the day-ahead contract between areas. Therefore, under normal operation condition, the tie-line interchange should be fixed with only a small variation. However, under contingencies,

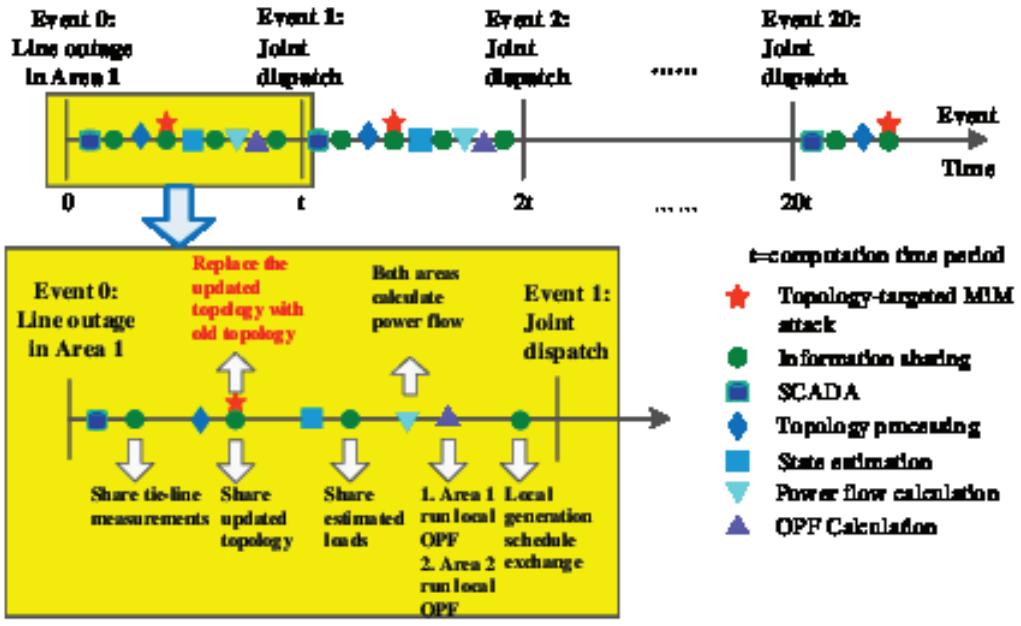


Figure 3.2: Time Sequence of Events at the Two Areas at the Time of and Following An Attack in One Area.

the highest priority is to fix the violation. Therefore, the tie-line power flow can vary up to the tie-line capacity. In this report, we model the system under both normal and contingency conditions. We first assume the attack is launched under normal condition. Under this condition, there are interchange agreement values on the tie-lines that are generally smaller than the tie-line capacities. In this model, the tie-line interchange values are allowed to vary only 10% variation of the original interchange agreement values. We then model the system under contingency with tie-line interchange varying up to the tie-line capacity. The simulation results for both system models are demonstrated in Section 3.4.

3.4 Illustration of Results

In this section, we illustrate our distributed communication attack and its consequences. We consider an IEEE 24-bus reliable test system (RTS) and decompose it into two areas (henceforth referred to as Areas 1 and 2) as shown in Figure 3.3 (the dashed red line separates the two areas) such that Area 1 and Area 2 are connected by four tie lines. Each area is assumed to have its own local control center that performs local SE with local measurements and tie-line power flow measurements shared from adjacent areas, following which it shares its estimated load information with the other area. This is followed by a PF calculation unit to make up for the load and generation mismatch caused by joint dispatch and then an OPF re-dispatch keeping the generator outputs external of the other area fixed. This process alternates between the two areas every t time units (see Figure 3.2).

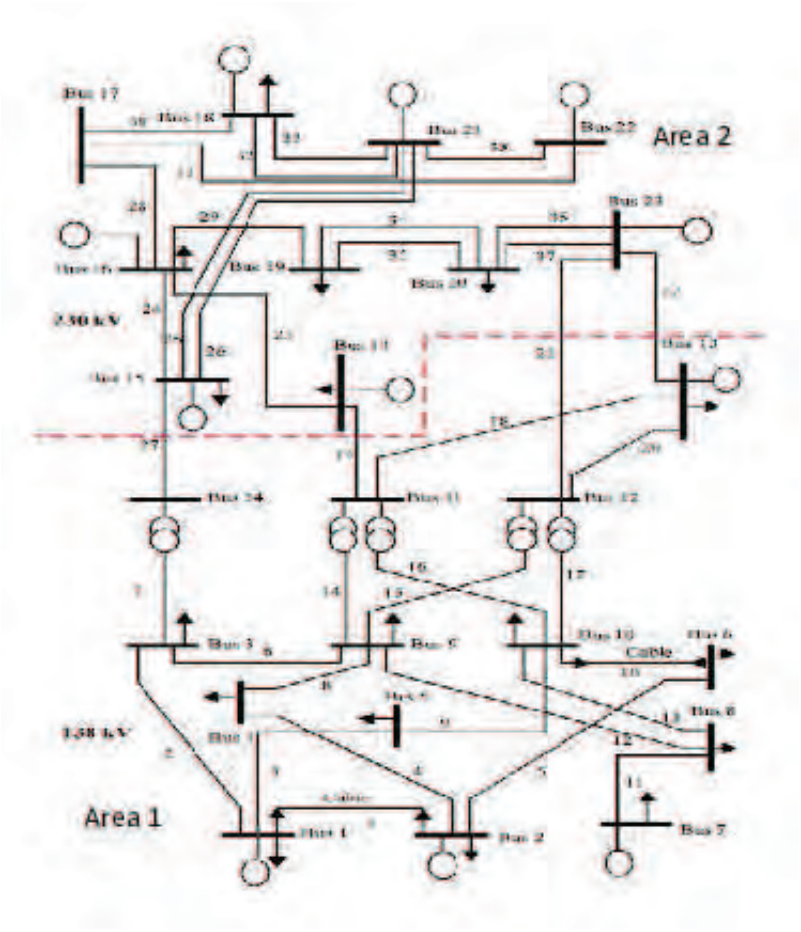


Figure 3.3: An IEEE RTS 24-bus Divided into Two Areas (Separated by Red Dashed Line).

The attack is modeled as a line outage in one area (*e.g.*, line 6 in Area 1). In order to understand the worst-case effect of the attack, the area without knowledge of the outage is assumed to have a congested line prior to the attack. The attacker, aware of

this outage in one area, compromises the topology changing communication signals such that the same static topology prior to the attack is shared. All possible choices of line outages in one area and congested lines in the other are considered exhaustively to demonstrate the effect of the possible attack cases. The system behavior is followed over $20t$ time units following the outage and over this time the two areas perform SE, PF, OPF, and dispatch. The events sequence when Area 1 has an outage and Area 2 is affected by the communication attack is shown in Figure 3.2. The time immediately after topology changing is assumed as Event 0. The attacker launches a MitM attack to block the topology changing data sharing between Area 1 and Area 2 at Event 0 and sustains such an attack during the following events. Therefore, the two areas continue re-dispatching together in the simulation time period with one area (Area 1) using correct topology to obtain optimal dispatch plan while the other (Area 2) using false topology to do so.

In this report, we focus on worst-case attacks. We assume that the area without real-time topology information has some lines at capacity, *i.e.*, congested. This is achieved in simulation by reducing the line rating to 90% of the base case power flow to create congestion. We first model the system under tight tie-line agreement, in which only 10% variation on tie-line power flow interchange is allowed, then we model the system under contingencies, in which no tie-line interchange limit is modeled. To demonstrate our simulation, we first document our results in tables and then provide the detailed analysis and plots for both tie-line agreement cases.

Table 3.1: System Behavior with Sustained Attack for IEEE 24-bus System When Tie-line Interchange Is Fixed with 10% Variation.

Feasible Case	Physical PF Overload	Cyber PF Overload	Not Converge	No Violation Case	Cyber-Physical PF Overload
540	24.82%	14.26%	30.00%	23.33%	7.59%

*PF: Power flow

Table 3.1 shows the numbers in percentage of the five possible long term (20 or more events) outcomes of an attack after Event 0 with tie-line interchange fixed. These attack consequences are quantified by comparing the cyber power flow and physical power flow in the area without the real-time topology (say Area 2) over the entire attack time duration. The cyber power flow is the OPF solution calculated by the control center in Area 2 with fixed external generation. The physical power flow, on the other hand, is the real power flow values of the system after dispatching with the most recent OPF dispatch solution with the true topology information. Therefore, for the area with false topology information, the cyber power flow values will be different from the physical power flow values. Five kinds of disparities are observed between the cyber and physical power flows following Event 0; we name them *physical PF Overload*, *cyber PF overload*, *Not converge*, *No violation case*, and *cyber-physical PF overload case*. We describe these disparities in detail below.

Physical PF (power flow) Overload cases: For area with false topology in these cases, there is a mismatch between the cyber and physical power flows due to the false topology. Monitoring the cyber power flow cannot reflect the severity of the physical overload. The physical power flow on the previous congested line overloads

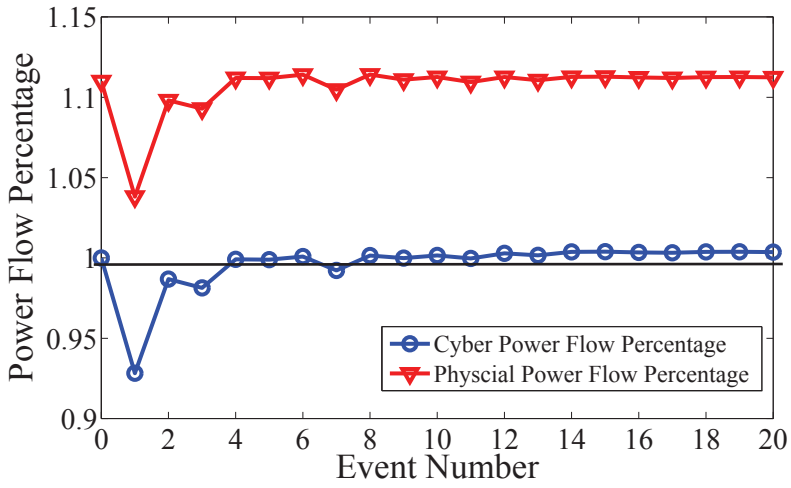


Figure 3.4: Physical PF Overload Case: Power Flow on Prior Congested Line 24 (Area 2) When Line 3 (Area 1) Is Outaged.

during the simulation time period. However, such overload problem is not shown or attenuated in cyber layer. A typical physical PF overload case (Line 3 connecting Bus 1 to Bus 5 is outaged with Line 24 connecting Bus 15 to Bus 16 congested) plot is shown in Figure 3.4. For these cases, the prior congested lines can get heated due to the dispatch of the area with false topology information. The heat accumulation may eventually cause the line to overheat and trip offline. Therefore, these cases can be viewed as *successful attack outcomes*.

Cyber PF Overload cases: For area with false topology in these cases, the cyber power flow is shown as overload during the simulation time period while in the physical layer, there is no overload happened. A typical cyber PF overload case (Line 18 connecting Bus 11 to Bus 14 is outaged with Line 29 connecting Bus 16 to Bus 19 congested) plot is shown in Figure 3.5. For these cases, the incorrect cyber overload alert can lead to wrong contingency behaviors such as throttling up other nearby sources, load shedding, or even worse, tripping transmission line or generators. We hence, view such cases as *successful attack outcomes.*

Not Converge cases: In these cases, the physical power flow overload happens in the first few events but eventually the OPF program fails obtain a dispatch plan for one or both areas. This is because for a fixed power generation from one area, there is no local dispatch plan that can satisfy all the constraints of the system even with thermal limit relaxation. In some cases, to clear the contingencies require more interchange between areas. Without the generator output changing jointly on both sides, the local center cannot find a feasible solution to solve the existing overload or stability problem. The worse operation states will continue until more serious consequences happened. Therefore, such cases can also be viewed as *successful attack outcomes*.

No violation case: For these cases, there is no overload immediately after Event 0 or a line overloaded after Event 0 can finally reduce below 100% of the rating in the simulation time period. Though the re-dispatch plan of the area with false topology

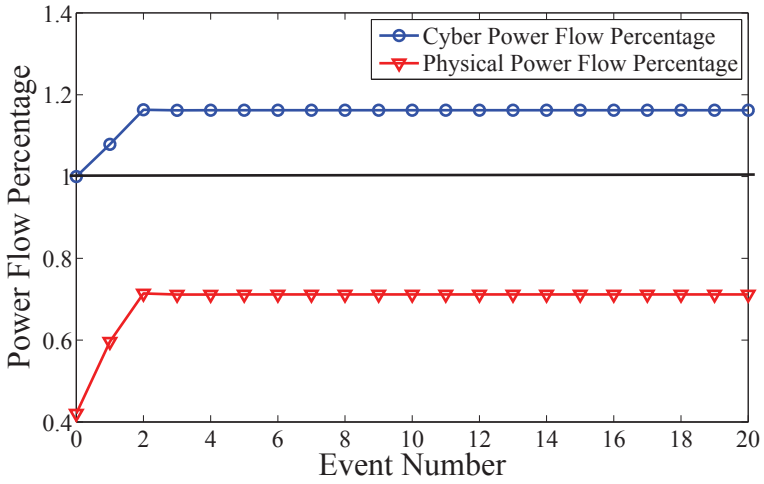


Figure 3.5: Cyber PF Overload Violation Case: Power Flow on Prior Congested Line 29 (Area 2) When Line 18 (Area 1) Is Outaged.

still give a wrong calculation values of the system, no further problem caused by the wrong plan. We, therefore, view the attacks leading to such cases as *unsuccessful attacks*.

Cyber-physical PF Overload cases: In these cases, despite there are overloads in physical layer, there is no mismatch between the physical and cyber power flow. Therefore, the control center can be aware of the overload problems and fix such problems in time. This class of cases is then viewed as *unsuccessful attack outcomes*.

We observe a total of 373 successful attack cases, *i.e.*, 69.08% of the total attack cases. We define the subclass of successful attacks for which the power flow of 105% relative to the flow following Event 0 as *critical* (successful) attacks, and note that the total number of *critical* attacks for the RTS system is 60, which is 11.11% of the total attack cases. These results demonstrate the potential vulnerability of a topology-based communication attack.

The statisics results of the long term outcomes of an attack after Event 0 without tie-line interchange limit is demonstrated in Table 3.2. Under such tie-line interchange model, we also observe the five disparities which are *physical PF Overload*, *cyber PF overload*, *Not converge*, *No violation case*, and *cyber-physical PF overload case* as introduced above. The proportion of successful attack cases is 65% of the total attack cases and that of the critical attack cases is 9.81% of the total attack cases. We can observe that with no tie-line interchange limitation, the number of not converge cases are largely reduced. However, such converged cases become physical PF overload cases or cyber overload cases. Hence, the proportion of the total successful attack cases are not changed too much.

Comparing the simulation results in Table 3.1 and Table 3.2, we can see that even under no tie-line interchange limitation, the MitM attacks can still lead to systematic problems and failures. Thus, the system is vulnerable to a topology-based communication attack under both tie-line interchange fixed condition and contingency condition. System operator should pay attention to such class of attacks.

Table 3.2: System Behavior with Sustained Attack for IEEE 24-bus System without Tie-line Interchange Limitation.

Feasible Case	Physical PF Overload	Cyber PF Overload	Not Converge	No Violation Case	Cyber-Physical PF Overload
540	35.74%	23.15%	6.11%	26.48%	8.52%

*PF: Power flow

3.5 Countermeasures and Concluding Remarks

In this report, we introduce a new class of distributed MitM attacks specifically targeting the topology sharing data between connected areas in the electric grid. We have demonstrated the time consequences of such attacks and have shown that such attacks can often lead to serious consequences if active intervention is not present. In this context, we observe that in addition to the traditional countermeasure of human operator-based data sharing (which have been shown to be error-prone and delayed too), it is essential to have more resiliency via automated data sharing mechanisms. Our attack is successful because the two areas process data largely independently except for data sharing and do not employ sanity checks for data from the other side or a more interactive distributed processing platform. This could help both areas become aware of inconsistencies over faster time-scales including: (a) create and share a list of *external contingencies* caused to other areas by an internal component outage; (b) identify the anormalies of such attacks and enable machine learning in EMS to detect such attacks. It is worth noting that, while some of these mechanisms are being considered or even used currently in the grid, it is not done in a uniform manner and this work highlights the limitations of not doing so.

References

[1] F. Sultanem, "Using appliance signatures for monitoring residential loads at meter panel level," *Power Delivery, IEEE Transactions on*, vol. 6, no. 4, pp. 1380–1385, 1991.

[2] G. W. Hart, "Nonintrusive appliance load monitoring," *Proceedings of the IEEE*, vol. 80, no. 12, pp. 1870–1891, 1992.

[3] M. L. Marceau and R. Zmeureanu, "Nonintrusive load disaggregation computer program to estimate the energy consumption of major end uses in residential buildings," *Energy Conversion and Management*, vol. 41, no. 13, pp. 1389–1403, 2000.

[4] M. Jawurek, F. Kerschbaum, and G. Danezis, "Sok: Privacy technologies for smart grids–a survey of options." *Microsoft Res., Cambridge, UK*, 2012.

[5] G. Kalogridis, C. Efthymiou, S. Z. Denic, T. Lewis, R. Cepeda *et al.*, "Privacy for smart meters: Towards undetectable appliance load signatures," in *Smart Grid Communications (SmartGridComm), 2010 First IEEE International Conference on*. IEEE, 2010, pp. 232–237.

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

[7] L. Yang, X. Chen, J. Zhang, and H. V. Poor, "Optimal privacy-preserving energy management for smart meters," in *INFOCOM, 2014 Proceedings IEEE*. IEEE, 2014, pp. 513–521.

[8] L. Sankar, S. R. Rajagopalan, S. Mohajer, and H. V. Poor, "Smart meter privacy: A theoretical framework," *smart grid, IEEE transactions on*, vol. 4, no. 2, pp. 837–846, 2013.

[9] C. Efthymiou and G. Kalogridis, "Smart grid privacy via anonymization of smart metering data," in *Smart Grid Communications (SmartGridComm), 2010 First IEEE International Conference on*. IEEE, 2010, pp. 238–243.

[10] S. Denic, G. Kalogridis, and Z. Fan, "Privacy vs pricing for smart grids," in *First IARIA International Conference on Smart Grids, Green Communications and IT Energyaware Technologies*, 2011.

[11] O. Tan, D. Gunduz, and H. V. Poor, "Smart meter privacy in the presence of energy harvesting and storage devices," in *Smart Grid Communications (SmartGridComm), 2012 IEEE Third International Conference on*. IEEE, 2012, pp. 664–669.

[12] L. J. Ratliff, C. Barreto, R. Dong, H. Ohlsson, A. Cardenas, and S. S. Sastry, "Effects of risk on privacy contracts for demand-side management," arXiv preprint arXiv:1409.7926, 2014. [Online]. Available: http://arxiv.org/pdf/1409.7926.pdf

[13] J. Yao and P. Venkitasubramaniam, "On the privacy-cost tradeoff of an in-home power storage mechanism," in *Communication, Control, and Computing (Allerton), 2013 51st Annual Allerton Conference on*. IEEE, 2013, pp. 115–122.

[14] R. Dong, A. A. Cárdenas, L. J. Ratliff, H. Ohlsson, and S. S. Sastry, "Quantifying the utility-privacy tradeoff in the smart grid," arXiv preprint arXiv:1406.2568 (2014), 2014. [Online]. Available: http://arxiv.org/pdf/1406.2568.pdf

[15] J. B. Rosen, "Existence and uniqueness of equilibrium points for concave n-person games," *Econometrica: Journal of the Econometric Society*, pp. 520–534, 1965.

[16] T. Basar and G. J. Olsder, *Dynamic noncooperative game theory*, 2nd ed. Society for Industrial and Applied Mathematics, 1995.

[17] NREL, "What is the energy payback for pv?" National Renewable Energy Laboratory, 2004. [Online]. Available: http://www.nrel.gov/docs/fy04osti/35489.pdf

[18] Energy Information Administration, "How much electricity does an american home use?" U.S. Energy Information Administration. [Online]. Available: http://www.eia.gov/tools/faqs/faq.cfm?id=97&t=3

[19] "Federal Energy Regulatory Commission (FERC): Final report on the August 14th blackout in the United States and Canada: Causes and recommendations," http://www.ferc.gov/industries/electric/indus-act/reliability/blackout/ch1-3.pdf, April 2004.

[20] "Federal Energy Regulatory Commission (FERC): Mandatory reliability standards for interconnection reliability operating limits," http://www.ferc.gov/whats-new/comm-meet/2011/031711/E-8.pdf, March 2011.

[21] "Federal Energy Regulatory Commission (FERC) and the North American Reliability Corporation (NERC): Arizona-Southern California outages on September 8, 2011," http://www.nerc.com/files/AZOutage-Report-01MAY12.pdf, April 2012.

[22] Y. Liu, P. Ning, and M. K. Reiter, "False data injection attacks against state estimation in electric power grids," in *Proceedings of the 16th ACM Conference on Computer and Communications Security*, ser. CCS '09, Chicago, Illinois, USA, 2009, pp. 21–32.

[23] O. Kosut, L. Jia, R. J. Thomas, and L. Tong, "Malicious data attacks on the smart grid," *IEEE Transactions on Smart Grid*, vol. 2, no. 4, pp. 645–658, 2011. [Online]. Available: http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=6032057

[24] S. Sridhar and M. Govindarasu, "Model-based attack detection and mitigation for automatic generation control," *IEEE Transactions on Smart Grid*, vol. 5, no. 2, pp. 580–591, 2014. [Online]. Available: http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=6740883

[25] J. Wei, D. Kundur, T. Zourntos, and K. Butler-Purry, "A flocking-based dynamical systems paradigm for smart power system analysis," in *Power and Energy Society General Meeting, 2012 IEEE*, 2012, pp. 1–8. [Online]. Available: http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=6344594

[26] J. Kim and L. Tong, "On topology attack of a smart grid: Undetectable attacks and countermeasures," *IEEE JSAC*, vol. 31, no. 7, pp. 1294–1305, 2013. [Online]. Available: http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=6547837

[27] L. Jia, J. Kim, R. J. Thomas, and L. Tong, "Impact of data quality on real-time locational marginal price," *IEEE Trans. Power Systems*, vol. 29, no. 2, pp. 627–636, 2014. [Online]. Available: http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=6657769

[28] L. Xie, Y. Mo, and B. Sinopoli, "Integrity data attacks in power market operations," *IEEE Transactions on Smart Grid*, vol. 2, no. 4, pp. 659–666, 2011. [Online]. Available: http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=6074981

[29] J. Liang, O. Kosut, and L. Sankar, "Cyber-attacks on ac state estimation: Unobservability and physical consequences," in *IEEE PES General Meeting*, Washington, DC, July 2014.

[30] A. Abur and A. G. Exposito, *Power System State Estimation: Theory and Implementation*. New York: CRC Press, 2004.