Following the Electrons: Methods for Power Management in Commercial Buildings

Gowtham Bellala, Manish Marwah, Martin Arlitt, Geoff Lyon, Cullen Bash
Hewlett Packard Laboratories
firstname.lastname@hp.com

ABSTRACT

Commercial buildings are significant consumers of electricity. The first step towards better energy management in commercial buildings is monitoring consumption. However, instrumenting every electrical panel in a large commercial building is expensive and wasteful. In this paper, we propose a greedy meter (sensor) placement algorithm based on maximization of information gained, subject to a cost constraint. The algorithm provides a near-optimal solution guarantee. Furthermore, to identify power saving opportunities, we use an unsupervised anomaly detection technique based on a low-dimensional embedding. Further, to better manage resources such as lighting and HVAC, we propose a semi-supervised approach combining hidden Markov models (HMM) and a standard classifier to model occupancy based on readily available port-level network statistics.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous

General Terms

Algorithms, Measurement, Performance

Keywords

Commercial buildings, power management, meter placement, anomaly detection, occupancy modeling

1. INTRODUCTION

In the United States alone, there are an estimated five million commercial buildings. In 2010, these buildings consumed about 1.3 trillion kWh of electricity, roughly one third of the electricity generated in the country. The energy costs for commercial buildings exceeds $100 billion annually. Due to recent economic turmoil, and increased awareness of environmental concerns (e.g., global climate change), many companies want to reduce power use in their buildings. Often, they turn to consulting firms for services like building energy efficiency analyses. The work described in this paper addresses shortcomings of existing analyses of this sort.

A first challenge for improving power use in a building is understanding how much power each appliance or device in the building uses. One option is to install power meters on all electrical panels, to collect usage data near the consumers, and get a (approximate) per-appliance breakdown of power use. The main disadvantage of this approach is the cost to buy and install the meters. For companies that own a lot of buildings (e.g., Walmart has more than 10,000 stores globally), this cost becomes prohibitively expensive. Thus, one research question we address is where to place a limited set of meters in a building, while minimizing the information loss. We propose an efficient greedy algorithm that provides a near-optimal solution.

A second challenge we investigate is how to systematically monitor building energy use and automatically detect problems that arise over time. A limitation of manual consulting services is that they can only identify issues that are occurring at the time the analysis is conducted, and typically only a limited number of panels are monitored, identified by an expert based on the likelihood of energy savings. Having a consultant repeat the study on a regular (e.g., daily) basis is not cost effective, so an automated technique is highly desired. We present results from applying our unsupervised anomaly (fault) detection and ranking methods for monitoring tens of meters over a six month period.

Lastly, consulting studies will typically recommend static solutions to reduce building energy use. For example, turn on all lights only during work hours (e.g., 8am to 6pm), and turn most off otherwise. While such techniques do help reduce power use in a building, further savings are possible. One approach is to only turn on lights (or HVAC systems) in areas where people are currently in, and to turn them off when the people leave. To facilitate such dynamic resource management, we developed a semi-supervised method for occupancy modeling.

Our group is instrumenting three large commercial buildings on the HP Labs campus in Palo Alto, CA. This instrumentation will provide extensive data on the campus power use, which will establish the "ground truth" against which we will evaluate our power management methods.

The paper makes the following contributions:

- It proposes a greedy algorithm for meter placement in a building's electrical infrastructure, to maximize mutual information while minimizing the cost of metering. Besides being computationally efficient, we also show that the proposed greedy algorithm guarantees a

KDD'12, August 12–16, 2012, Beijing, China.

Figure 1: The HP Labs Palo Alto Campus.

near-optimal solution. In particular, we show that mutual information becomes submodular under a special graphical structure that arises in distribution networks such as power, water and gas.

- The meter placement results gauged by the ability of the selected meters to predict measurements of the unselected ones are better on average by about 15% over other methods considered.

- The results are demonstrated for six months of data from a large test bed (three buildings totaling 300,000 sq. ft.). The anomaly detection, and occupancy modeling techniques, described in more detail in [3], are validated over this data set.

The remainder of the paper is organized as follows. Section 2 provides an overview of the HP Labs campus, the power delivery and measurement infrastructure, and the campus power use characteristics. Section 3 describes our KDD methods. Section 4 evaluates our methods. Section 5 discusses related work. Section 6 summarizes our work and future directions.

2. CAMPUS OVERVIEW

The HP Labs campus contains six main buildings with a total footprint of 700,000 sq. ft. We are instrumenting three two-storey buildings (1, 2, 3), as highlighted in Figure 1. These three buildings have a 300,000 sq. ft. footprint and host about 500 occupants.

2.1 Power Distribution Topology

Buildings 1, 2 and 3 are powered by a single utility feed (3-phase 12.5kV). An emergency generator (3-phase 480V) maintains critical loads in the event of a utility failure. Automatic transfer switches (ATS) are used to revert from the utility to backup power. Each building has a main distribution panel (3-phase 480V) that branches to about 10 major sub-loads or sub-panels. A 135kW photo voltaic array on top of Building 3 offsets power demand during daylight hours.

2.2 Power Data Collection

To date, 33 power meters have been deployed on our campus. These include meters for building and top-level load distribution panels in Buildings 1-3. We are now instrumenting the second-tier distribution panels within each building, to obtain finer grained electrical data for our future work.

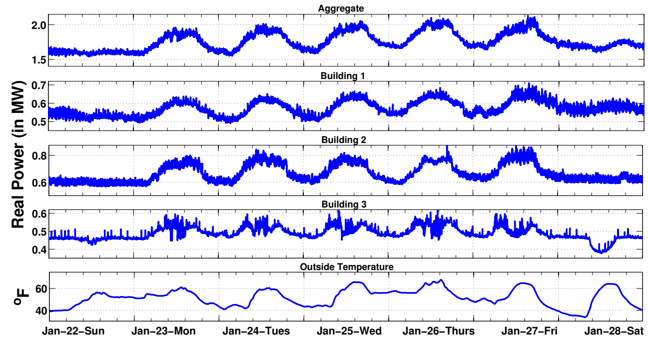


Figure 2: Campus power use and outside temperature.

The installed electrical meters are commercial (3-phase) devices from Schneider Electric (`www.schneider-electric.com`). Data is retrieved from each meter every 10 seconds using the MODBUS over Ethernet protocol. The data includes metrics such as line voltage, real and apparent power, power factor, current and frequency. The data is stored in a PI-Server from OSIsoft (`www.osisoft.com`).

2.3 Campus Power Use Characteristics

As further motivation for the challenges addressed in this paper, we briefly examine some characteristics of the campus power use.[1] The top graph in Figure 2 shows the total power use for Buildings 1-3 over a one week period (from Jan. 22 through Jan. 28, 2012). The peak load is nearly 2 MW; understanding how to reduce this would translate directly to operational savings for the company. The base load is roughly 1.5 MW. An implication of this is very little insight on what is responsible for campus power consumption can be gleaned (e.g., via disaggregation techniques like [9] since none of the algorithms scale up to handle hundreds to thousands of loads present in commercial buildings, and further, none of the methods disaggregate base load) from the aggregate power. This means that more meters must be installed. Our meter placement algorithm addresses the issue of how many meters are needed and where they are needed, to minimize the cost while maximizing the information obtained.

The middle three graphs in Figure 2 show the total power demand for Buildings 1, 2 and 3, respectively. The bottom graph shows the outside temperature. Comparing this graph to the others reveals a correlation between outside temperature, occupancy (i.e., work hours) and power use. This motivates our investigation in Section 3.3 of occupancy modeling, to reduce the use of heating or cooling in areas of the building that people are not actively using.

3. METHODS

Figure 3 shows the overall framework of our approach. The meter placement algorithm forms the basis for instrumenting a building power infrastructure. For building power management, we propose an unsupervised anomaly detection and ranking method based on low dimensional embed-

[1]We expect these characteristics to exist with many other commercial buildings as well.

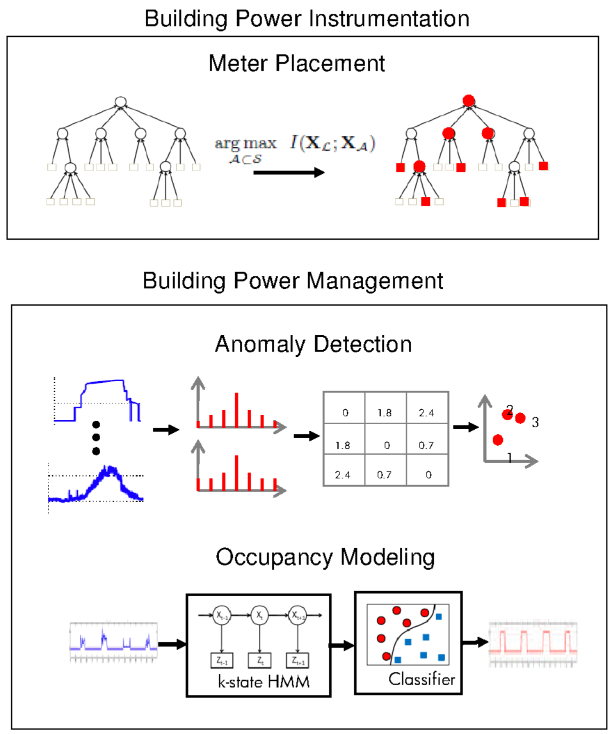


Figure 3: Overview of the methods used.

ding and k-nearest neighbors. For dynamic control of lighting and HVAC resources, we describe a semi-supervised approach that uses network port statistics to model occupancy.

3.1 Meter Placement

As noted in Section 2.3, the significant base load in the power demand of these buildings and the sheer number of devices in a large commercial building make known load disaggregation methods unreliable, thus requiring extensive metering of different electrical panels in each of these buildings for fine-grained power monitoring. However, one issue with this approach is that the total number of panels that could potentially be monitored can be very large, to the extent that meter deployment at all these locations is not economical (the cost of each power meter ranges between \$900 and \$3,000. This raises an interesting research question as to how and which panels should be selected for power meter deployment.

There are several criteria one could use to choose the panels for meter deployment. They include the total energy consumption of a panel, variability in the energy consumption, number of sub-panels/loads, predictability of panel power demand, or an information-theoretic measure. We choose mutual information, an information-theoretic measure that in a loose sense chooses panels that are highly unpredictable in terms of their power demands. As we show in Section 4.1, the panels selected using this criterion are superior to those selected using criterion such as the total energy consumption or variability in energy consumption.

Next, we demonstrate how this problem can be formulated as an optimization problem with the goal of choosing the set of panels with maximum information content.

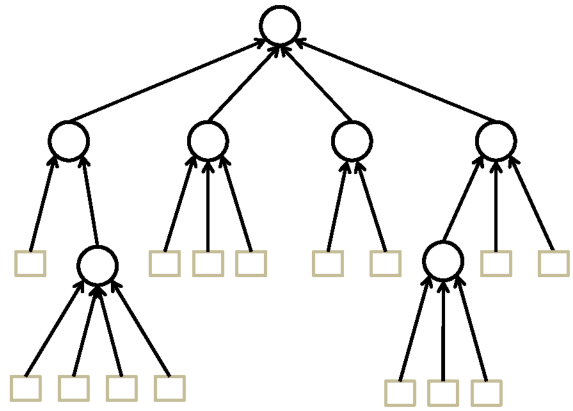


Figure 4: Example topology of building electrical panels where meters can be installed.

3.1.1 *Problem Formulation*

Before we formulate the problem, we need to introduce some notation. The panels at different locations on the site are related in a topological manner that can be represented by a tree, as shown in Figure 4. Each node in this tree denotes a panel, where the leaf nodes (denoted by squares) correspond to panels that directly feed either a single load (e.g., a chiller or a compressor) or a set of loads (e.g., lighting load). The remaining nodes in the tree topology (denoted by circles) correspond to panels that feed other panels given by their child nodes in the tree.

Let $\mathcal{S}$ denote the entire set of panels, i.e., all the nodes in a given tree, and let $\mathcal{L} \subset \mathcal{S}$ denote the set of leaf nodes. For any node $i \in \mathcal{S}$, let X_i be a random variable denoting the power consumption recorded at panel i. Then, for any set of nodes $\mathcal{A} \subset \mathcal{S}$, we denote by $\mathbf{X}_\mathcal{A}$ the random variables associated with the nodes in $\mathcal{A}$.

Given a constraint on the number of meters that can be afforded (k), we use mutual information as a criteria to choose the best panels for meter deployment, which can be formulated as an optimization problem shown below.

$$\begin{aligned} &\arg\max_{\mathcal{A}\subset\mathcal{S}} \; I(\mathbf{X}_\mathcal{L};\mathbf{X}_\mathcal{A}) \\ &\quad \text{s.t.} \quad |\mathcal{A}| \le k, \end{aligned} \tag{1}$$

where $I(\mathbf{X}_\mathcal{L};\mathbf{X}_\mathcal{A})$ denotes the amount of information conveyed by monitoring power consumption at panels in $\mathcal{A}$ about the power consumption at panels in $\mathcal{L}$. Note that when $k \ge |\mathcal{L}|$, i.e., in the scenario where one could afford to deploy meters at each of the leaf nodes that directly feed a bunch of loads, the mutual information is maximized by choosing $\mathcal{A}$ to be the set of all the panels in $\mathcal{L}$. On the other hand when $k < |\mathcal{L}|$, the above optimization problem attempts to find the best set of panels that provide maximum information about power consumption at each of these leaf nodes (the last layer of panels in the hierarchy).

3.1.2 *Proposed Solution*

Mutual Information is given by

$$I(\mathbf{X}_\mathcal{L};\mathbf{X}_\mathcal{A}) = H(\mathbf{X}_\mathcal{L}) - H(\mathbf{X}_\mathcal{L}|\mathbf{X}_\mathcal{A}),$$

which corresponds to the reduction in uncertainty about power consumption at panels in $\mathcal{L}$ given the power consumption information at panels in $\mathcal{A}$, where

$$H(\mathbf{X}_\mathcal{L}) = -\sum_{\mathbf{x}_\mathcal{L}} \Pr(\mathbf{X}_\mathcal{L} = \mathbf{x}_\mathcal{L}) \log_2 \left(\Pr(\mathbf{X}_\mathcal{L} = \mathbf{x}_\mathcal{L})\right), \quad \text{and}$$

$$H(\mathbf{X}_\mathcal{L}|\mathbf{X}_\mathcal{A}) = -\sum_{\mathbf{x}_\mathcal{L}, \mathbf{x}_\mathcal{A}} \Pr(\mathbf{X}_\mathcal{L} = \mathbf{x}_\mathcal{L}, \mathbf{X}_\mathcal{A} = \mathbf{x}_\mathcal{A}) \log_2 \left(\Pr(\mathbf{X}_\mathcal{L} = \mathbf{x}_\mathcal{L}|\mathbf{X}_\mathcal{A} = \mathbf{x}_\mathcal{A})\right).$$

Unfortunately, the optimization problem in (1) is NP-hard. Hence, we propose a greedy approach to optimize the given problem. The greedy algorithm chooses panels for meter deployment in a sequential manner, where given the set of panels that have already been chosen by the algorithm (denoted by $\mathcal{A}$), the next best panel is chosen to be the one that maximizes the gain in mutual information, i.e.,

$$j^* = \arg\max_{j \notin \mathcal{A}} I(\mathbf{X}_\mathcal{L}; \mathbf{X}_{\mathcal{A}\cup j}) - I(\mathbf{X}_\mathcal{L}; \mathbf{X}_\mathcal{A}).$$

The solution obtained using the above greedy algorithm is not necessarily an optimal solution for the optimization problem in (1). However, we show below that the obtained greedy solution is guaranteed to be near-optimal.

3.1.3 Near-optimality of greedy solution

To show that the solution obtained using the above greedy approach is near-optimal, we rely on the theory of submodularity introduced by Nemhauser *et al.* [21] and popularized by the work of Krause *et al.* [14, 13].

Krause *et al.* [14] study budgeted maximization problems of the form

$$\begin{aligned} &\arg\max_{\mathcal{A}\subset\mathcal{S}} F(\mathcal{A}) \\ &\text{s.t.} \quad |\mathcal{A}| \leq k, \end{aligned}$$

where $\mathcal{S} = \{1, \cdots, N\}$ is a set of elements and $F : \mathcal{A} \to \mathbb{R}$ is a function that maps the set of elements to the real line. A greedy solution to this problem is to select elements sequentially according to

$$j^* = \arg\max_{j\notin\mathcal{A}} F(\mathcal{A}\cup j) - F(\mathcal{A}).$$

Krause *et al.* [14] show that the solution obtained using this greedy approach is near-optimal in the following sense

$$F_{\text{greedy}} \geq \left(1 - \frac{1}{e}\right) F_{\text{opt}},$$

iff the objective function F is submodular, where submodularity is defined below.

Definition 1. **(Submodularity)** *Let F be a function that maps from a set of elements $\mathcal{S}$ to the real line $\mathbb{R}$. Then, F is said to be submodular iff $\forall \mathcal{A} \subseteq \mathcal{B} \subseteq \mathcal{S}$ and for any $j \notin \mathcal{B}$,*

$$F(\mathcal{A}\cup j) - F(\mathcal{A}) \geq F(\mathcal{B}\cup j) - F(\mathcal{B}).$$

In our optimization problem, the objective function is mutual information, which unfortunately is not submodular, except for some known special cases [13, 16, 17]. However, as we will show below, in our problem setting, mutual information turns out to be submodular, thus guaranteeing near-optimality of the above greedy algorithm.

Lemma 1. *Given the tree topology described in Section 3.1, let $\mathcal{S}$ denote the set of nodes in the tree and $\mathcal{L}$ the set of leaf nodes. Then, $\forall \mathcal{A} \subseteq \mathcal{B} \subseteq \mathcal{S}$, and for any $j \notin \mathcal{B}$,*

$$I(\mathbf{X}_\mathcal{L}; \mathbf{X}_{\mathcal{A}\cup j}) - I(\mathbf{X}_\mathcal{L}; \mathbf{X}_\mathcal{A}) \geq I(\mathbf{X}_\mathcal{L}; \mathbf{X}_{\mathcal{B}\cup j}) - I(\mathbf{X}_\mathcal{L}; \mathbf{X}_\mathcal{B}) \tag{2}$$

Proof. From the definition of mutual information, we have

$$\begin{aligned} I(\mathbf{X}_\mathcal{L}; \mathbf{X}_{\mathcal{A}\cup j}) - I(\mathbf{X}_\mathcal{L}; \mathbf{X}_\mathcal{A}) &= H(\mathbf{X}_\mathcal{L}|\mathbf{X}_\mathcal{A}) - H(\mathbf{X}_\mathcal{L}|\mathbf{X}_{\mathcal{A}\cup j}) \\ &= H(X_j|\mathbf{X}_\mathcal{A}) - H(X_j|\mathbf{X}_\mathcal{L}, \mathbf{X}_\mathcal{A}), \end{aligned} \tag{3}$$

where the second equality follows from the first by expanding the entropy terms and by simple manipulation of the resulting terms. Note that the second term in (3) is equal to 0, i.e., $H(X_j|\mathbf{X}_\mathcal{L}, \mathbf{X}_\mathcal{A}) = 0$, as given the power consumption at all the leaf nodes, the power consumption at any panel upstream is completely deterministic.

Hence, the relation in (2) reduces to showing

$$H(X_j|\mathbf{X}_\mathcal{A}) \geq H(X_j|\mathbf{X}_\mathcal{B}),$$

which follows from the principle of "information never hurts" in information theory [5]. Thus, proving the submodularity of mutual information under the given tree topology. □

3.1.4 Use of Granger Causality

Another strategy to select meters would be by applying Granger causality, which considers the *direction* of flow of information unlike mutual information. Note that this could be remedied by the use of Transfer entropy, which is a version of mutual information that can detect the direction of information flow [24], however, transfer entropy is currently restricted to bivariate situations.

Granger Causality (or G-causality) test, which was initially introduced in the field of economics [8], is a statistical hypothesis test for determining whether one time series is useful in forecasting another. It is normally tested in the context of linear regression models. For example, let $X(t)$ and $Y(t)$ be two time series. Consider the following two auto-regressive models for predicting $X(t)$

$$X(t) = \sum_{j=1}^{p} a_j X(t-j) + e_1(t)$$

$$X(t) = \sum_{j=1}^{p} a_j X(t-j) + \sum_{j=1}^{p} b_j Y(t-j) + e_2(t),$$

where p is the maximum number of lagged observations included in the model, and $e_1(t), e_2(t)$ are the prediction errors (residuals) for the two regression models. If the variance in the prediction error is reduced by the inclusion of $Y(t)$ in the model, then Y is said to G-cause X. In other words, Y is said to G-cause X if the coefficients in $\{b_j\}_{j=1}^p$ are jointly significantly different from zero.

This test could potentially be used to reveal any hidden causal relationships between the loads (leaf nodes in the tree topology). Incorporating these relationships could further lead to a better choice of panels for meter deployment.

3.2 Anomaly Detection

Anomaly (or fault) detection is useful in detecting abnormal behavior in the power usage data collected from a building. Note that an anomaly indicates an irregular usage pattern and may not always correspond to a component failure or faulty operation. Since labeled data is difficult to obtain, we propose an unsupervised cluster-based algorithm that detects anomalous points via a low-dimensional embedding of the power data. This algorithm takes as input the power time series of a meter over a period of time, and outputs the probability of a particular day being anomalous. The probability scores can be used to rank the days in terms of anomalousness, providing a building administrator with a prioritized list of data points that require further inspection.

We refer to power data measured by a single meter over a 24 hour period (i.e., one day) as one observation or as a single power-time curve. As mentioned, we use an unsupervised approach where we cluster the power-time curves of each meter. The intuition behind this approach is that the data points that exhibit normal behavior will form tight clusters while anomalous points will lie outside these clusters.

To compare two power-time curves, we use the standard Euclidean distance measure, or the l_2 norm, between the frequency spectrum of the two power-time curves. While the frequency spectrum consists of two components - magnitude and phase, we restrict our attention to the magnitude of the frequency spectrum as it contains all the necessary information regarding the power consumption behavior.

Our algorithm consists of five steps. First, missing values in a power-time curve are imputed. Second, the frequency spectrum of the imputed power-time curve is computed. Third, the standard Euclidean distance measure is used to determine the dissimilarity between the power consumption profiles of any two days. Fourth, a low dimensional embedding of the power-curves. The fifth step uses this low dimensional embedding to compute the probability score of each observation being anomalous. We achieve this using a k-NN (nearest neighbor) density estimation algorithm. A more detailed explanation of our algorithm is provided in [3].

3.3 Occupancy Modeling

Occupancy modeling forms another important component for efficient power management in buildings. Many commercial buildings employ either a fixed time L-HVAC (lighting, heating ventilation and cooling) schedule or a fixed temperature set point schedule. This often leads to unnecessary conditioning of the building, especially when the actual occupancy is low. Hence, some recent work has suggested occupancy-based L-HVAC scheduling for efficient power management. However, most of this work assumes the availability of occupancy sensors, whose installation and maintenance may be prohibitive on a large campus.

Melfi et al. [20] propose an alternative method that uses existing network infrastructure to estimate occupancy. They studied the use of Dynamic Host Control Protocol (DHCP) logs and other explicit ways such as monitoring PC activity in estimating occupancy. We instead developed an implicit occupancy sensing procedure, where we use traffic statistics associated with layer 2 network ports in each cubicle to build occupancy models. Network switches typically maintain counters (for each physical layer 2 port) for the amount of in and out flowing traffic. We retrieve these statistics from the switches in the buildings every 30 minutes. The estimated occupancies at cube level are then used to estimate occupancy of a zone (e.g., multiple cubicles), which

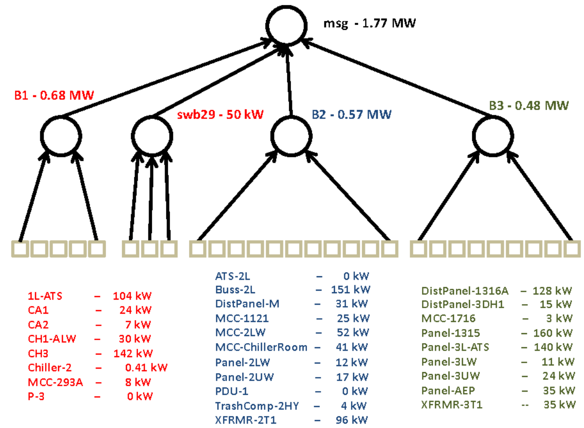


Figure 5: Tree topology of the 33 meters

can further be used for occupancy-based L-HVAC scheduling of that zone.

A primary challenge with using network data to estimate occupancy is the lack of labeled data. To address this, we consider two different approaches. The first is an unsupervised approach where we use hidden Markov model (HMM) [6] to estimate occupancy from network data. The second is a two-stage semi-supervised approach. Its first stage involves unsupervised learning using HMM, while the second trains a classifier using minimal labeled data. The two approaches are described in more detail in [3].

4. EXPERIMENTAL RESULTS

4.1 Meter Placement

To assess our meter placement method, we used the greedy algorithm described in Section 3.1 to select the most informative power meters among the 33 installed in our campus. The tree topology corresponding to these power meters is shown in Figure 5. The figure also shows their average power consumption values. For simplicity, we assumed in our experiments that the random variables corresponding to the power consumption at different panels are Gaussian.

Next, we greedily selected the meters in a sequential manner using the greedy algorithm described in Section 3.1 under three different criteria: mutual information, total power consumption and variability in power consumption. Table 1 shows a ranked list of the meters in the order in which they are selected using these three criterion. Figure 6 shows plots comparing the three ranked lists. The diagonal line (dotted red line) in these plots correspond to the scenario where the two ranked lists are exactly equal.

Given this ranked list and a budget ($t < 33$) on the number of panels that can be metered, the top t panels from the ranked list can be chosen to be metered. The power consumption measured at these metered panels can then be used to predict the power consumption at the remaining panels. We use this predictive ability as a measure of goodness of the selected meters. We compare the above three criteria based on the predictive ability of the selected panels. A random selection is included as a baseline. Figure 7(a) shows the average RMS (root mean squared) prediction error over all non-metered panels as a function of the number of panels metered (t). Similarly, Figure 7(b) shows the average normalized RMS prediction error, where the nor-

#	Total Power Consumption	Variability in Power Consumption	Mutual Information
1	msg	swb29-main	swb29-main
2	swb29-main	msg	msg
3	b1-main	b1-main	b2-Buss-2L
4	b2-main	b2-main	b1-main
5	b3-Panel-1315	b3-Panel-AEP	b1-CH1-ALW
6	b1-CH-3	b1-CH-3	b2-main
7	b3-Panel-3L-ATS	b2-DistPanel-M	b3-Panel-AEP
8	b3-DistPanel-1316A	b1-CH1-ALW	b1-CH-3
9	b1-1L-ATS	swb29-Chiller-2	b2-DistPanel-M
10	b2-XFRMR-2T1	b2-MCC-2LW	b2-MCC-2LW
11	b2-Buss-2L	b3-Panel-3UW	b1-CA-2
12	b2-MCC-2LW	b2-Panel-2UW	b3-Panel-1315

Table 1: Selection order of first 12 meters

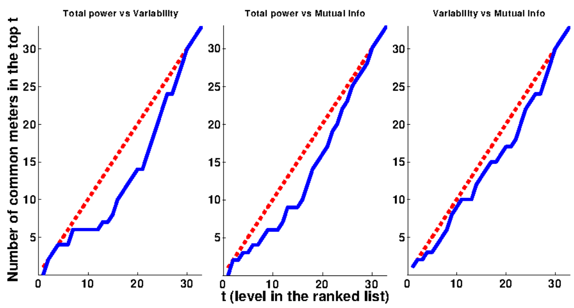


Figure 6: Comparison of ranked list similarity

malization is based on the average power consumption of a meter. These two figures reveal that the proposed mutual information based meter deployment outperforms those based on the total power consumption and the variability in power consumption by an average of about 15%. The curve corresponding to the random selection is averaged over 100 different random orderings of the 33 meters.

Figure 8 shows the top 12 panels selected using mutual information. This solution can be verified intuitively. For example, panel b3-main is not selected as it is completely deterministic given the aggregate meter (msg) and the other three main meters (b1-main, swb29-main and b2-main). Similarly, most of the selected leaf node panels consist of loads that are less predictable given the others. For example,"b3-Panel-AEP" directly measures the power generated by the photo-voltaic array installed on building 3. This panel's output is less predictable and hence can be considered as a good choice for meter deployment.

4.1.1 *Discussion and Extensions*

One limitation of our current implementation is the Gaussian assumption on the distribution of the random variables. This is not a strictly valid assumption, as most panels have at least two distinct operating states, one with higher power consumption during business hours and the other with a baseload power consumption during non-business hours. Hence, a more accurate approach would be to model these random variables as a mixture of Gaussians. However, one limitation of using a Gaussian mixture model is that there is no closed form expression for entropy of a Gaussian mixture density, and hence one would need to approximate it.

4.1.2 *Use of Granger Causality*

As described in Section 3.1.4, we investigated applying

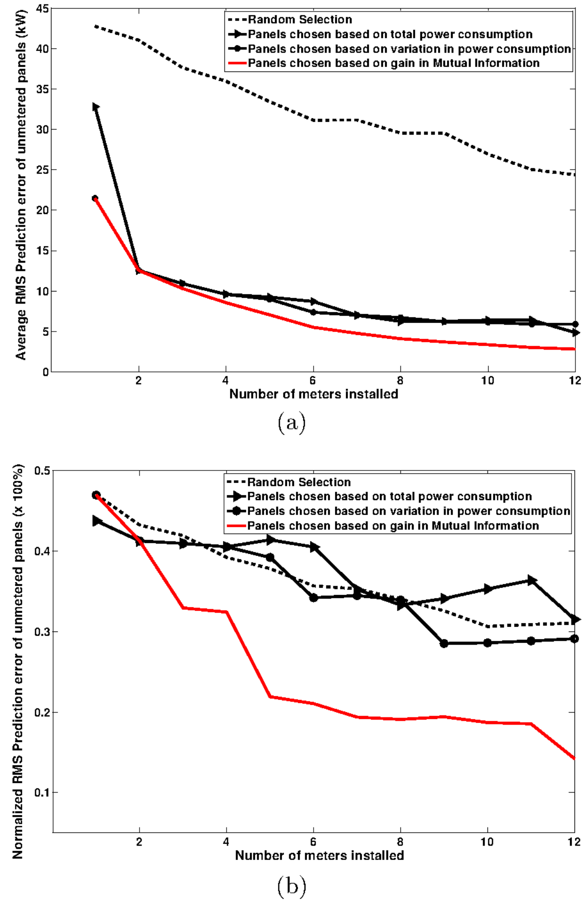


Figure 7: Comparison of the predictive ability of the panels selected for metering: (a) Average RMS prediction error (b) Average Normalized RMS prediction error

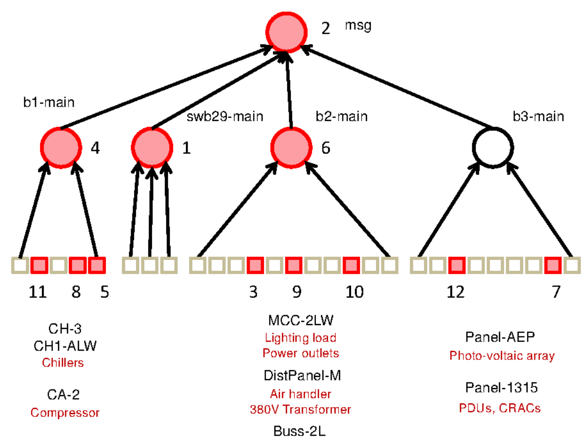


Figure 8: The panels selected by the greedy algorithm and mutual information criterion

Granger causality to a subset of meters. Figure 9 shows interesting causal relationships between 8 sub-panels in Building 1. The loads on these panels are as follows: panels b1-CA-1 and b1-CA-2 feed two different compressors, while the panels b1-CH-3, b1-CH1-ALW and swb29-Chiller-2 feed three different Chillers. The loads on the remaining 3 panels

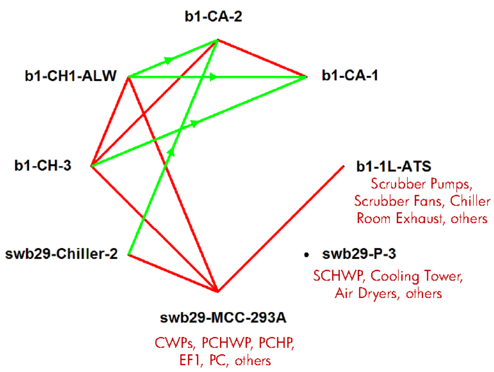


Figure 9: The causal relationship between different sub-panels in Building 1

Meter name	AUC
b1-main	0.87
b2-main	0.96
b3-main	0.99

(a)

	Category	# of Anomalies
1	High power usage	66
2	Low power usage	65
3	Irregular Shutdown	6
4	Irregular (time) usage	9
5	Oscillatory behavior	28
6	Abnormal drop/rise	29

(b)

Figure 10: Anomaly detection measurements: (a) AUC (b) Anomaly types and number.

are listed in the figure. The undirected edges demonstrate strong bi-directional causal relationships, while the directed edges demonstrate strong uni-directional causal relationship in the direction given by the arrow. Note from this figure that there is a strong causal relationship between the chillers and the compressors. For example, there is a strong causal relationship from chiller-2 to compressor-2, while chiller-3 seems to strongly G-cause compressor-1.

Furthermore, we used the G-causal test to rank the panels based on their predictive ability (results omitted for brevity). However, the mutual information based ranking performs better than the G-causal ranking with respect to the RMS prediction error on the non-metered panels. This could be due to the fact that the G-causal test is currently limited to linear regression models, whereas information theoretic measures are also sensitive to any non-linear relationships. Although extensions of the G-causal test to non-linear models exist, they are computationally less efficient and their statistical properties are not well studied [1].

4.2 Anomaly detection

We performed anomaly detection on six months of data from the 33 power meters. To validate our results, for three meters (b1-main, b2-main and b3-main) we obtained the ground truth by consulting with the building administrator, who looked at the entire time series data and marked days with potential anomalous regions. As described in Section 3.2, our algorithm assigns a probability score to each day, which can be used to obtain a ranked list of days in decreasing order of them being anomalous.

Given this ranked list, a building administrator could choose a threshold k and declare the top k points as anomalies for further inspection, and the remaining as normal, where k could vary from 0 to the maximum number of points in the input data (M). Each choice of k results in a certain number of false positives and false negatives. For example, when $k = 0$, i.e., when all the points are declared as normal, the false positive rate (FPR) is 0 while the false negative rate (FNR) is 1. On the other hand, when $k = M$, the associated FPR is 1 and FNR is 0. Varying this threshold k results in different values of FPR and FNR, leading to a receiver operating characteristic (ROC) curve. The area under the ROC curve (AUC) defines the quality of the obtained ranking. In the ideal case, where all the anomalous points are ranked at the top followed by normal points, the AUC takes the maximum value of 1. On the other hand, a random ranking achieves an AUC value of 0.5. We use AUC as a performance metric for our algorithm. Figure 10(a) shows the performance of our algorithm on 3 meters.

Further, we applied our algorithm on the remaining 30 meters, where we obtained a ranked list of anomalous days for each meter. We then manually characterized the top anomalies (for a conservative k value) by assigning them categories, as shown in Figure 10(b). Note that a particular anomaly could belong to multiple categories. Detecting these anomalies could potentially offer several benefits such as energy savings, detecting faulty equipment resulting in savings in maintenance costs, etc. Potential power savings in the high power usage and irregular time usage anomalies varied from around 50 to 2000 kWh per anomaly. In Figure 11, we demonstrate 4 of these 6 categories.

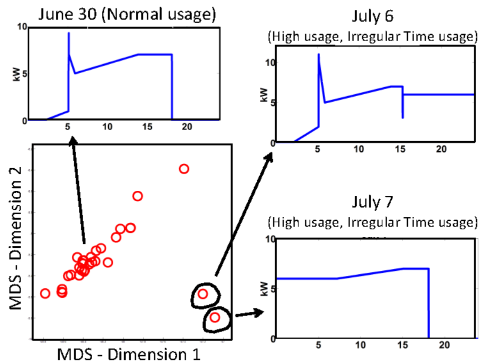
Figure 11(a) corresponds to the air handlers in a building, while Figure 11(b) corresponds to a chiller load. The low-dimensional embeddings obtained using MDS in both cases show clusters of normal behavior (June 30th and June 17th), and points (circled) that were detected as anomalous.

For the fan load, two anomalous points are seen (July 6th and July 7th) corresponding to high and irregular time usage (categories 1 and 4), where the air handlers were operating all through the night. Detecting and correcting these anomalies could have potentially saved 70 kWh over these two days. For the chiller load, three anomalous points corresponding to three consecutive days have been detected (July 5th and July 6th shown in the figure), where the chiller was abruptly shut down (categories 2 and 3) during business hours. If this was not caused due to a maintenance schedule, it could potentially correspond to a failed component.

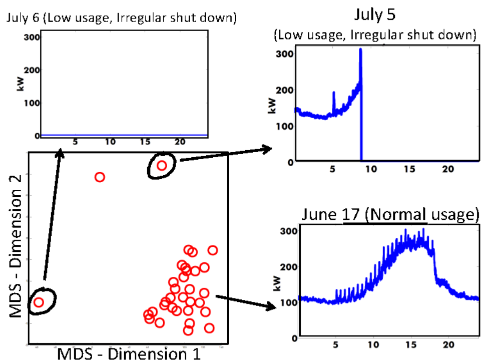
4.3 Occupancy modeling

In this section, we demonstrate the superior performance of the proposed two-stage approach for occupancy estimation. In particular, we show that the use of k-state HMM as a pre-processing step in the two stage approach leads to a better classification accuracy over directly using the network switch port statistics in the classification algorithm. We compare the performance of the 2-stage approach with that of a 1-stage approach where the classifier is trained directly based on the network switch level port statistics.

To quantify the estimation accuracy of these algorithms, we collected ground truth data from 10 occupants over a pe-

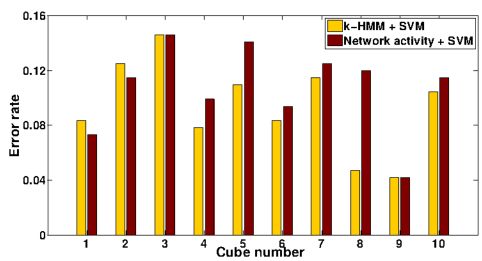


(a) Load corresponding to air handlers.

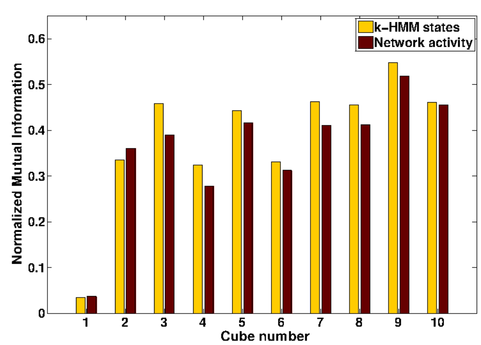


(b) Load corresponding to a chiller.

Figure 11: Examples of low dimensional embedding



(a)



(b)

Figure 12: (a) Average error rates of models. (b) Feature quality.

riod of 16 weekdays. The occupants recorded their presence in their cube at a time resolution of 30 minutes, where an occupant marks their presence only if they are present for a majority of that 30 minute period. The in and out flowing network switch statistics are also collected every 30 minutes.

Figure 12 shows the experimental results. Figure 12(a) compares the average error rate between the two approaches, where the error rate is averaged over 16 different test cases. Figure 12(b) compares the quality of the two features (k-HMM output and network statistics) in terms of estimating the occupancy states. The feature quality is measured using normalized mutual information between the feature states and the true occupancy labels. The figures show that the k-HMM output has a better feature quality resulting in an improved classification accuracy.

The estimated occupancy states for each cube are then aggregated to estimate the occupancy of a zone. We further show that the estimated zone occupancy using the proposed 2-stage approach can be used to dynamically control the lighting in that zone, resulting in an energy savings of around 9.53%. More details on use of this approach to control lighting are provided in [3].

5. RELATED WORK

Commercial buildings consume a lot of energy [25], which motivates research to improve building energy efficiency. As discussed in Section 1, there are a variety of challenges to address. The problem of selecting optimal locations for meter placement can be framed as a budgeted optimization problem. This problem has been well studied in the literature in different contexts, such as in the case of sensor placement in a water distribution network [18] or observation selection in an autonomous robotic exploration [15]. Several different criteria have been proposed in the literature for selecting these optimal locations [15]. The most popular among them being mutual information [16, 17]. However, as we discussed earlier, this constrained optimization problem with mutual information as an objective function is known to be NP-hard, and hence requires use of greedy, near-optimal strategies to solve.

Examining data for anomalies is a known approach for identifying abnormal system behavior. Catterson et al. use this approach to monitor old power transformers [4]. Their goal is to pro-actively search for abnormal behavior that may indicate the transformer is about to fail. Li et al. [19] search for anomalies in building power consumption, where they employ simple statistical tests such as the Q-test to detect time points with abnormal power usage. On the other hand, Jakkula and Cook [11] demonstrate the superiority of a clustering based approach over such simple statistical tests in the context of identifying abnormal activities in household power consumption data. Our work on anomaly detection is built on observations in these studies.

Occupants of a building contribute to its energy footprint. Unfortunately, directly tracking the number of people in a building is often more difficult than one might think. To estimate the occupancy, we create models based on data retrieved from periodic scans of the computer network in the campus. This is a similar approach to that of Newsham and Birt [22]. Erickson *et al.* [7], Agarwal *et al.* [2], and Kim et al. [12] model occupancy through a variety of means. Rice et al. develop a model of building energy consumption, to offer insights on where energy is used [23]. Hay and Rice investigate how to assign power use to individuals, who may be able to adjust their behaviors to reduce the aggregate power consumption [10]. Our work is complementary.

6. CONCLUSIONS

Commercial buildings consume significant amounts of energy. Concerns over energy prices and global climate change are motivating building operators to reduce energy consumption. In this paper, we propose and evaluate three methods to aid in this effort. Our meter placement algorithm is both efficient and effective, guaranteeing a near optimal solution to information maximization by exploiting submodularity. In comparisons with other methods, the ability of the meter set selected using our algorithm to predict the measurements of the unselected meter set were found to be superior (by an average of about 15%). Our anomaly detection method is shown to identify numerous types of unexpected consumption patterns. Lastly, our occupancy modeling approach can be used to dynamically control lighting or HVAC resources, thereby reducing their energy consumption.

We plan to extend our work in several ways. We intend to leverage occupancy modeling results for enhancing anomaly detection. Further, we plan to automate the anomaly characterization task, and extend our algorithm to incorporate feedback from a building administrator. In addition, we are increasing instrumentation on our campus, to aid in validating our methods. Finally, further methods may be developed as we evolve our test bed into a demonstrator.

7. REFERENCES

[1] `http://www.scholarpedia.org/article/Granger_causality`.

[2] Y. Agarwal, B. Balaji, J. Lyles, M. Wei, and T. Weng. Occupancy-driven energy management for smart building automation. In *ACM Workshop on Embedded Sensing Systems for Energy-Efficiency in Buildings (BuildSys)*, Zurich, Switzerland, 2010.

[3] G. Bellala, M. Manish, M. Arlitt, G. Lyon, and C. Bash. Towards an understanding of campus-scale power consumption. In *ACM Workshop on Embedded Sensing Systems for Energy-Efficiency in Buildings (BuildSys)*, 2011.

[4] V. Catterson, S. McArthur, and G. Moss. Online conditional anomaly detection in multivariate data for transformer monitoring. *IEEE Transactions on Power Delivery*, 25(4):2556–2564, 2010.

[5] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley Interscience, 1991.

[6] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. Wiley, New York, 2. edition, 2001.

[7] V. Erickson, Y. Lin, A. Kamthe, and R. Brahme. Energy efficient building environment control strategies using real-time occupancy measurements. In *ACM Workshop on Embedded Sensing Systems for Energy-Efficiency in Buildings (BuildSys)*, Berkeley, CA, 2009.

[8] C. W. J. Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 1969.

[9] G. Hart. Nonintrusive appliance load monitoring. *Proceedings of the IEEE*, 80(2):1870–1891, 1992.

[10] S. Hay and A. Rice. The case for apportionment. In *ACM Workshop on Embedded Sensing Systems for Energy-Efficiency in Buildings (BuildSys)*, Berkeley, CA, 2009.

[11] V. Jakkula and D. Cook. Outlier detection in smart environment structured power datasets. In *IEEE Intelligent Systems*, London, UK, 2010.

[12] Y. Kim, R. Balani, H. Zhao, and M. Srivastava. Granger causality analysis on ip traffic and circuit-level energy monitoring. In *ACM Workshop on Embedded Sensing Systems for Energy-Efficiency in Buildings (BuildSys)*, Zurich, Switzerland, 2010.

[13] A. Krause and C. Guestrin. Near-optimal nonmyopic value of information in graphical models. *Uncertainty in Artificial Intelligence (UAI)*, 2005.

[14] A. Krause and C. Guestrin. A note on the budgeted maximization of submodular functions. Technical report, CMU-CALD-05-103, 2005.

[15] A. Krause and C. Guestrin. Near-optimal observation selection using submodular functions. *National Conference on Artificial Intelligence (AAAI)*, 2007.

[16] A. Krause, C. Guestrin, A. Gupta, and J. Kleinberg. Near-optimal sensor placements: Maximizing information while minimizing communication cost. *International Symposium on Information Processing in Sensor Networks (IPSN)*, 2006.

[17] A. Krause, A. Singh, and C. Guestrin. Near-optimal sensor placements in gaussian processes: Theory, efficient algorithms and empirical studies. *Journal of Machine Learning Research*, 9:235–284, 2008.

[18] J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. VanBriesen, and N. Glance. Cost-effective outbreak detection in networks. *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 420–429, 2007.

[19] X. Li, C. Bowers, and T. Schnier. Classification of energy consumption of a building with outlier detection. *IEEE Transactions on Industrial Electronics*, 57(11):3639–3644, 2010.

[20] R. Melfi, , B. Rosenblum, B. Nordman, and K. Christensen. Measuring building occupancy using existing network infrastructure. *International Green Computing Conference*, 2011.

[21] G. Nemhauser, L. Wolsey, and M. Fisher. An analysis of the approximations for maximizing submodular set functions. *Mathematical Programming*, 1978.

[22] G. Newsham and B. Birt. Building-level occupancy data to improve ARIMA-based electricity use forecasts. In *ACM Workshop on Embedded Sensing Systems for Energy-Efficiency in Buildings (BuildSys)*, Zurich, Switzerland, 2010.

[23] A. Rice, S. Hay, and D. Ryder-Cook. A limited-data model of building energy consumption. In *ACM Workshop on Embedded Sensing Systems for Energy-Efficiency in Buildings (BuildSys)*, Zurich, Switzerland, 2010.

[24] T. Schreiber. Measuring information transfer. *Physical Review Letters*, 85:461–464, 2000.

[25] United States Energy Information Administration. Annual energy review 2010. `http://www.eia.gov/aer`, 2011.