NILM Redux: The Case for Emphasizing Applications over Accuracy

Sean Barker, Sandeep Kalra, David Irwin, and Prashant Shenoy
University of Massachusetts Amherst

***Abstract*—Non-Intrusive Load Monitoring (NILM) has recently experienced a rebirth due to the expanding deployment of network-connected smart meters by utilities and the increasing availability of Internet-enabled consumer-grade power meters. While many dimensions of the problem have been well-studied over the past 25 years, we argue that prior work has placed too much emphasis on incremental improvements in accuracy and not enough on designing novel NILM applications. As a result, the basic NILM problem and its primary application—a simple appliance-level breakdown of home energy usage—has remained unchanged since its inception. We believe a renewed focus on NILM applications could help steer future research in novel directions by exposing new problem variants, data analysis techniques, and evaluation metrics. In this paper, we summarize our own application-centric research agenda, which focuses on online applications that generate results in real time as smart meters produce data. As we discuss, our focus on applications has led us to consider efficiency and performance issues not addressed in prior work, which typically targets offline data analysis.**

I. Introduction

Non-Intrusive Load Monitoring (NILM), or energy disaggregation, has been an active research area for 25 years, starting with Hart's original work in 1989 [1], [2]. As recent surveys show [3], [4], researchers have proposed novel NILM algorithms for a broad spectrum of problem variants, which differ based on their data type (e.g., real/reactive power, current, voltage, etc.), data acquisition hardware, data resolution, scale (i.e., number of devices), diversity (i.e., type of devices), deployment length, and time lag (i.e., the delay in returning results). In recent work, Armel et al. categorize 18 different algorithms from prior work that effectively target 18 different variants of the problem [3]. To further complicate matters, in many cases, authors do not fully specify their problem variant. As one example, four of the 18 authors above did not specify their data resolution, which may vary from 100,000,000 samples per second to 1 sample per hour and has a strong impact on the efficacy of any NILM method [3]. Surprisingly, there are actually few NILM algorithms for any given variant, and often just one. For instance, Armel et al. cite only two approaches that disaggregate the 1Hz real power data common to popular consumer-grade, Internet-enabled power meters, such as the TED [5] and eGauge [6]. As a result, proposed common standards for evaluating and comparing different NILM algorithms (e.g., REDD [7]), while laudable, may have limited value, since algorithms targeting widely different problem variants are fundamentally incomparable.

In addition to the many existing problem variants, a wide variety of accuracy metrics for NILM also exist, many of which are incomparable across different algorithms for the same problem variant. For example, some algorithms might detect when devices either turn "on" or "off" and use binary classification metrics (e.g., precision, recall, MCC, etc.) to quantify accuracy, while others might assume devices vary their energy usage continuously and quantify accuracy based on the total energy correctly assigned to each device. As a result, algorithms that use the former approach to quantify accuracy could show "high" precision and recall (for some definition of "high"), but still incorrectly assign much of the energy usage if each device's energy use varies widely when "on," as recent work shows is often the case [8]. The converse may also be true, with algorithms capable of showing a "low" precision and recall, but correctly assigning much of the energy usage. In addition, NILM algorithms may exhibit highly variable accuracy (for some metric) on a per-load basis. For example, techniques that model devices as having a small number of discrete power states will likely perform well for devices that actually have discrete power states, and less well for devices that continuously vary their power usage in complex patterns. As a result, any technique's accuracy for a given dataset will depend on the set of devices within the data and their characteristics. Each accuracy metric also implicitly values certain behavior in a NILM algorithm. For example, using the "total energy correctly assigned" for an entire home as the accuracy metric prioritizes accurately assigning energy to the devices that consume the most energy, while discounting high accuracy for low-power devices. However, as we discuss, some applications may value high accuracy for some important but low energy devices. For example, many interactive devices consume relatively little energy, but are useful for activity recognition and provide the most insight into user behavior.

Ultimately, the variety of NILM variants and accuracy metrics combined with its 25 year history of prior research makes it difficult to determine NILM's important open problems, or even if it has any important open problems. The plethora of continuing research in the area indicates that researchers believe NILM is not a solved problem, i.e., new techniques will result in more than just incremental improvements in accuracy over existing methods (at least for some problem variants). However, since problem variants differ widely, the open problems for each variant (and their associated accuracy metric) are likely also different. So, is there a "right" problem variant and accuracy metric for NILM? Given some accuracy metric, what is a sufficiently "high" level of accuracy? How do we determine if a new technique's improvement in accuracy over an existing technique is significant or just incremental? Of course, the "correct" answer to each of these questions depends on NILM's target application. For instance, using NILM as part of a recommendation engine that pushes energy-efficiency suggestions to users' smartphones in real time differs substantially from using it to determine which home appliances consume the most energy over a month. The former values accuracy at each point in time and has real-time performance requirements, while the latter values accuracy over the course of an entire month and permits offline analysis.

Unfortunately, in most cases, prior NILM research does

not explicitly mention or target a specific application. Often, the implicit application is computing a simple appliance-level energy breakdown over some time interval, e.g., a day or month. We believe NILM research needs to move past computing appliance-level energy breakdowns. Energy breakdowns, themselves, are not a particularly compelling application, since they do not directly lead to quantifiable improvements in energy-efficiency. Instead, to demonstrate its value to outsiders, the research community should emphasize designing new and novel applications using NILM, rather than seeking incremental improvements in various accuracy metrics for slightly different problem variants. We argue that NILM research should be application-centric: the primary results of any new technique should demonstrate how it enables a novel application or enhances an existing application. A focus on applications will also have the side-effect of putting evaluations of accuracy in proper context by demonstrating if the technique was accurate enough to support the application. Likewise, future work can focus on showing how improving the technique improves the application, which also serves the purpose of putting any accuracy improvements in context.

In this paper, we outline our own application-centric research agenda, including ongoing and future research on NILM, as well as a broader set of energy data analytics. This focus on applications has led us to value *online and scalable analytics*: online NILM (and other online and scalable analytics) computes results in real time soon after a meter generates new data. As a result, such analytics must be efficient, enabling them to scale to massive grid-sized data sets including tens of thousands of customers. Many novel applications require such real-time and efficient analytics to be useful. Since most prior NILM variants target offline analysis, ostensibly to compute energy breakdowns, they are often computationally expensive. As a result, they are neither online (as they cannot generate new results as meters generate data) nor readily scalable (as they consume significant computing resources per home).

II. Example Online Applications

Before discussing our own research agenda and approach, we first briefly describe a handful of online applications that we consider interesting and potentially novel. While these applications are not strictly dependent on NILM, they could each benefit from an accurate NILM algorithm.

- **Virtual Power Sensor.** The foundational application that an online approach to NILM should be able to provide is that of a *virtual power sensor* that mimics having a network-connected energy meter attached to each device. As recent research shows [9], large-scale sensor deployments remain problematic due to their expense ($\sim$\$40-80 per device), invasiveness (many hard-wired devices are difficult to directly meter), and unreliability (sensors often fail to report data). Thus, an online NILM-based system that creates virtual power sensors by analyzing data from a single smart meter to infer each device's power usage is highly attractive. Of course, to be an effective power sensor replacement, the system should report power usage at similar resolutions in real time, e.g., every few seconds for a typical networked plug meter. A virtual power sensor application would be useful for effectively any application that requires deploying power sensors to monitor the power of one or more devices. Prior research is replete with examples of such sensing-based applications—we list a few possibilities below.
- **Device Scheduling.** There are a variety of policies a home might employ to programmatically schedule when background devices, such as refrigerators, freezers, air conditioners, and heaters, operate. For example, utilities might incentivize homes to schedule devices to reduce their peak demand [10], or homes with local renewable energy deployments might schedule devices to better align their power usage with renewable generation (thereby decreasing battery capacity requirements) [11]. Of course, such device scheduling requires knowing when each device is using power and how much power is used. Thus, a virtual power sensor as described above would eliminate the need to install and maintain a multitude of per-device power sensors to collect this data in real time.
- **Recommendation Engine.** Another novel online application is a recommendation service that monitors the power usage of various devices, and then issues real-time alerts to users that notify them of immediate actions they can take to optimize their energy usage. In recent work, Banerjee et al. design such a recommendation service for an off-grid home that relies on solar energy and battery-based energy storage for power [12]. Such homes must carefully regulate their power usage over time to align with renewable generation as closely as possible. Thus, the recommendation service advises users when high-power appliances should be run, and notifies them of energy conservation opportunities. The service uses power sensors to report fine-grained power usage for high-power devices, but, as the authors note, could benefit from a more non-intrusive method of data collection.
- **Demand Response Capacity Estimation.** Finally, we propose a new online application to aid utility demand response programs. Currently, these programs typically enable a utility to schedule the operation of specific devices, e.g., air conditioners, for participating consumers in exchange for lower electricity rates [13]. Thus, the demand response capacity—the amount of power the utility can shed—varies over time based on how many of these devices from participating consumers are operating. Today, utilities have no way to monitor the total demand response capacity in real time. However, utilities with access to smart meter data could use the virtual power sensor application above to monitor each home's air conditioner. This application also highlights the importance of computational efficiency: a utility may need to perform NILM across tens of thousands of homes in its customer base to perform this estimation. While, in a single home, dedicating a server for online NILM analysis does not seem onerous, dedicating tens of thousands of servers for utility-scale analysis is likely cost-prohibitive.

Note that, in this paper, we specifically target consumer-grade power meters, such as the TED [5], eGauge [6], and BrulTech, which commonly provide a sampling resolution of one average (real) power reading per second, e.g., $\tau = 1$ second. While today's utility-grade smart meters provide, at most, minute-level sampling (e.g., a reading every five minutes is common), there are indications the next generation of meters will provide second-level sampling. For example, a U.K. sub-

committee defining future smart meter specifications released a report advocating a five second sampling resolution [14].

III. Existing Approaches

Most prior NILM algorithms are not appropriate for the online applications above, in large part, because they are designed for offline analysis and are too computationally expensive [2], [15], [7]. These algorithms generally model devices using a small number of discrete power states, i.e., one or more "on" states and an "off" state with minimal or no power usage. Each device is then represented by a small state machine that transitions between power states based on either user behavior or some internal device control algorithm. Using this approach, a building's power usage is then simply a larger state machine that transitions between a large number of discrete power based on the state of its constituent devices. Such state-based approaches are attractive to computing researchers because they admit a variety of off-the-shelf techniques already developed for analyzing state machines.

For instance, much prior work maps building state machines to Hidden Markov Models (HMMs), and applies HMM-based techniques to determine which loads are on in each state [7]. For these techniques, using only a few power states per load is advantageous, since it minimizes the number of distinct power states for the entire building and reduces the complexity of analyzing the resulting state machine. However, even with only two power states per load, the number of building power states is still exponential in the number of loads, i.e., 2^n for n loads. Thus, precise analysis is still intractable for large values of n, requiring enumerating an exponential number of states. This approach is practical for applications that only focus on accurately computing the total energy correctly assigned to each device, since the value of n can often be substantially reduced because, typically, only a small number of devices, e.g., HVAC, refrigerator, dryer, washing machine, etc., contribute the vast majority of a home's energy usage. However, the approach is less practical for fully disaggregating the 30-100 individual, and often small, electrical devices found in a typical U.S. home.

In general, state-based approaches do not work well online applications that value computational efficiency and monitoring low energy devices, e.g., for activity recognition. For example, based on our benchmarks of a state-of-the-art NILM algorithm [7] based on Factorial Hidden Markov Models, after training the models on data, the algorithm takes 86 seconds on a dedicated 2.4GHz Xeon server to disaggregate all 25 circuits in a typical home (where we model each circuit as having four power states). As a result, such an algorithm would have at least an 86 second time lag before reporting each set of results. Further, for utility-scale applications, the approach would either require a dedicated server per home or exhibit much worse performance, i.e., a much longer time lag.

IV. Model-driven Approach

To overcome the problems mentioned above, our own work is taking a different strategy that decouples the NILM problem into multiple distinct and independent subproblems, some of which are performed offline and some online. The general NILM problem consists of at least three distinct parts: determining what devices are in a home, building a model of the devices to capture their behavior, and using those models to determine when the device runs. Many NILM techniques conflate these issues by assuming that nothing is known about the home *a priori*, and attempt to concurrently solve each problem. In contrast, our work considers the first two problems as offline, one-time "configuration" tasks for each building and device and the last problem as an online task designed to run as a smart meter generates new data. We discuss the challenges associated with each of these NILM subproblems in turn.

A. Device Discovery

Device discovery is the process of determining the set of electrical devices in a home. Existing NILM algorithms that generate their models based on training data also require a device discovery phase to know how many devices to model. In general, these existing algorithms simply assume training data from each device in the home (or a similar device in another home) is available to generate a device-specific model. However, we do not leverage training data in our approach for at least two reasons: i) in practice, it is usually not available from the home in question (since its presence would eliminate the reason for performing NILM) and ii) it often embeds usage-specific information, i.e., how someone uses the device, that diminishes the value of using training data across homes.

Rather than use training data, our approach requires knowledge of specific device models, down to their particular brand and model number. Such specificity is necessary, since as we show in recent work [8], different devices of the same type might use power in different ways, e.g., refrigerators from Maytag and GE might use power in different ways and would thus need different models. One challenge with this approach is that performing manual device discovery, while possible, imposes a significant burden on the user. We see multiple possibilities for addressing this challenge. For example, mobile apps, such as Amazon PriceCheck, are capable of automatically indexing a specific type of brand and model of a device from a photograph, and could be modified for this purpose. Alternatively, future devices may include RFID tags that enable users to index them with an RFID reader. Finally, for external users without access to the home, device discovery could be done directly from the home's smart meter data by searching for identifiable power signatures—specific sequences of changes in power—from a large database of known signatures. Such device discovery itself represents an interesting NILM-related problem.

B. Device Modeling

For each device in a home, our approach requires an accurate model of the device. Rather than generate models for each home's devices using training data, as we mention above, we view model derivation as a one-time offline "configuration" task. Ideally, manufacturers would derive and publicly release power models for their devices as part of the manufacturing process. Many manufacturers often already include similar types of detailed power usage information in device technical manuals. Since these models are developed offline, they are only able to capture usage-independent characteristics of a device's power usage when on, i.e., they can capture how a toaster uses power when turned on but not how often the

toaster turns on or how long. Such models could also be crowd-sourced by individuals. For example, The Power Consumption Database already provides crowd-sourced information on maximum and idle power for a wide range of specific loads, indexed by type, manufacturer, brand, and model number [16].

In recent work [8], we present a simple and accurate modeling methodology based on whether a device is resistive, inductive, or non-linear load (or some combination thereof) in an alternating current (AC) system. Since each type of AC load exhibits similar characteristics, we design a few classes of parameterized models that capture the power usage of loads in each class. For example, high-power resistive heating elements exhibit a exponential decay in their power usage after startup as they heat up and their heating element's resistance decreases. As another example, non-linear switched mode power supplies exhibit rapid, random fluctuations from a stable power state. Our work defines four basic model classes—on-off (for low-power resistive devices), exponential growth/decay (for high-power resistive devices and inductive devices), stable min-max (for switched mode power supplies and electronic controllers), random range (for other types of electronics)—and two compound classes—cyclic (for devices with automated controllers that repeat at a regular, well-defined interval), and composite (for devices that operate multiple simpler internal devices using the above models, in sequence, parallel, or both).

We show that our model classes apply to nearly every household device (since every device is either resistive, inductive, non-linear, or a combination of them), and is more accurate than models based on a limited number of discrete power states. The primary challenge with our modeling approach is deriving a device's class and constructing its model programmatically. Currently, constructing a device's model is a manual and time-consuming process. Some composite devices operate a collection of resistive, inductive, and non-linear devices in complex patterns that requires independently modeling each of the constituent devices and combining them together. Ideally, models and their specific parameters could be derived empirically using traces of device power usage. In ongoing work, we are exploring techniques to automatically identify model features in per-device power data to both classify the type of device and construct its model.

C. Device Tracking

Our ultimate goal is to use the models above to enable the virtual power sensors from §II that mimic a networked power sensor by tracking and reporting the power usage of a device in real time. The primary challenge in enabling an accurate and efficient virtual power sensor is distilling *models* into a set of identifiable *features* that are readily observed in aggregated smart meter data. In many cases, the models themselves are too complex and lengthy to accurately detect when embedded in aggregated data. Generally speaking, *identifiable* features are either large (i.e., high power) or very distinct and brief in relation to the aggregated data. In other words, large power consumption and distinct but brief periods of power consumption are least likely to be obscured by other devices in aggregated data. One goal of our work is to define an identifiability metric that captures these high level observations, enabling us to automatically select the most identifiable set of features from a device's model.

Once selected, the next challenge is to design efficient methods to i) detect these features from smart meter data and ii) extract the device's power consumption. Efficiently detecting a device model's features enables a virtual power sensor to know that the device is consuming power, while extracting the device's power consumption utilizes both the features and the device's model to infer device power consumption over time (even during periods where its less identifiable model features are obscured). Designing specific detection and extraction techniques for each of our model classes is ongoing work; however, we have been able to track devices using model-based approaches more than an order of magnitude faster (e.g., compared to the disaggregation result given in §III) than using traditional disaggregation techniques.

V. Conclusion

This paper argues that future NILM research should emphasize the design of novel applications that use NILM, rather than incremental improvements in accuracy for different problem variants. We detail the basis for this argument and then describe our own research driven by the development of a virtual power sensor—i.e., a type of online NILM that mimics having a networked power sensor attached to each device. We then discuss i) interesting higher-level applications of a such virtual power sensor, ii) some limitations of existing NILM research in supporting such an application, and iii) the status of own research in developing such a mechanism.

References

[1] G. Hart, "Nonintrusive Appliance Load Monitoring," *IEEE*, vol. 80, no. 12, December 1992.

[2] ——, "Residential Energy Monitoring and Computerized Surveillance via Utility Power Flows," *IEEE Technology and Society Magazine*, vol. 8, no. 2, June 1989.

[3] K. Armel, A. Gupta, G. Shrimali, and A. Albert, "Is Disaggregation the Holy Grail of Energy Efficiency? the Case of Electricity," *Energy Policy*, vol. 52, no. 1, January 2013.

[4] M. Zeifman and K. Roth, "Nonintrusive Appliance Load Monitoring: Review and Outlook," *IEEE Transactions on Consumer Electronics*, vol. 57, no. 1, February 2011.

[5] "Energy, Inc." http://www.theenergydetective.com/.

[6] "eGauge Energy Monitoring Solutions," http://egauge.net.

[7] J. Kolter and M. Johnson, "REDD: A Public Data Set for Energy Disaggregation Research," in *SustKDD*, August 2011.

[8] S. Barker, S. Kalra, D. Irwin, and P. Shenoy, "Empirical Characterization and Modeling of Electrical Loads in Smart Homes," in *IGCC*, June 2013.

[9] T. Hnat, V. Srinivasan, J. Lu, T. Sookoor, R. Dawson, J. Stankovic, and K. Whitehouse, "The Hitchhiker's Guide to Successful Residential Sensing Deployments," in *SenSys*, November 2011.

[10] S. Barker, A. Mishra, D. Irwin, P. Shenoy, and J. Albrecht, "SmartCap: Flattening Peak Electricity Demand in Smart Homes," in *PerCom*, March 2012.

[11] J. Taneja, D. Culler, and P. Dutta, "Towards Cooperative Grids: Sensor/Actuator Networks for Renewables Integration," in *SmartGridComm*, 2010.

[12] N. Banerjee, S. Rollins, and K. Moran, "Automating Energy Management in Green Homes," in *HomeNets*, August 2011.

[13] "Baltimore Gas and Electric Peak Rewards Air Conditioning Program," http://peakrewards.bgesmartenergy.com/programs/ac.

[14] "Smart Meter Implementation Programme," https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/42737/1480-design-requirement-annex.pdf.

[15] H. Kim, M. Marwah, M. Arlitt, G. Lyon, and J. Han, "Unsupervised Disaggregation of Low Frequency Power Measurements," in *SDM*, April 2011.

[16] "The Power Consumption Database," http://www.tpcdb.com.