Exploring The Value of Energy Disaggregation Through Actionable Feedback

Nipun Batra
IIIT Delhi
Email: nipunb@iiitd.ac.in

Amarjeet Singh
IIIT Delhi
Email: homer@thesimpsons.com

Kamin Whitehouse
University of Virginia
Email: whitehouse@virginia.edu

***Abstract*—Over the past few years, dozens of new techniques have been proposed for more accurate energy disaggregation, but the jury is still out on whether these techniques can actually save energy and, if so, whether higher accuracy translates into higher energy savings. In this paper, we explore both of these questions. First, we develop new techniques that use disaggregated power data to provide actionable feedback to residential users. We evaluate these techniques using power traces from 240 homes and find that they can detect homes that need feedback with as much as 84% accuracy. Second, we evaluate whether existing energy disaggregation techniques provide power traces with sufficient fidelity to support the feedback techniques that we created and whether more accurate disaggregation results translate into more energy savings for the users. Results show that feedback accuracy is very low even while disaggregation accuracy is high. These results indicate a need to revisit the metrics by which disaggregation is evaluated.**

I. Introduction

Over the past years, new techniques have been proposed for more accurate energy disaggregation, but the jury is still out on whether these techniques can actually save energy and, if so, whether higher accuracy translates into higher energy savings. In this paper, we explore both of these questions. NILM research is broadly motivated by the philosophy of Lord Kelvin: "If you can't measure it, you can't improve it," but is the opposite also true? Disaggregation techniques can be used to help users identify the major energy consumers in their home, and some users will pour over this data to achieve significant energy savings [2]. However, several studies have shown that just providing energy data does not necessarily translate to long-term energy savings. After the novelty wears off, users experience a *rebound effect* as unsustainable energy-saving actions unwind themselves and as users tire of sifting through too much data [1]. Other studies have shown more sustainable effects by providing more targeted feedback or recommendations for simple actions [5]. Can disaggregation techniques be used to generate the types of actionable feedback that would produce sustainable energy savings?

In this paper, we present the exploration of two research questions that are highly relevant to the research community interested in energy disaggregation. First, we explore whether disaggregated power data can be used to provide actionable feedback to residential users, and whether that feedback is likely to save energy. We focus on feedback about refrigerators and HVAC. We develop a model that breaks the power trace of a refrigerator into three parts: baseline (when no one is using the fridge), defrost (energy consumption when the fridge is in defrost mode) and usage (energy consumption due to fridge usage). Then, we develop techniques to identify users with 1) much more energy due to fridge usage than the norm 2) much more energy due to defrost than the norm, or 3) fridges that are malfunctioning or misconfigured, even during baseline operation. We evaluate our model using a dataset with power traces from 95 refrigerators. Results indicate that our model can break down fridge usage into its three components with only 4% error. Additionally, the three types of feedback could help users save up to 23%, 25% and 26% of their fridge energy usage, respectively. These techniques provide targeted feedback with specific actions, e.g. fix or repair the fridge, and so we expect this energy savings to be sustainable. Similarly, we develop new techniques to differentiate homes with and without setback schedules on the HVAC system based on their HVAC power traces and outdoor weather patterns. This information can be used to give feedback to install a programmable thermostat. We evaluate these techniques with power traces from 58 homes and results indicate that our techniques can classify homes with 84% accuracy. Based on these results, we conclude that disaggregation does indeed have the potential to provide targeted, actionable feedback that could lead to sustainable energy savings.

Second, we explore whether existing energy disaggregation techniques provide power traces with sufficient fidelity to support the feedback techniques that we created, and whether more accuracy disaggregation results translate into more energy savings for the users. To do this, we re-evaluate the feedback techniques above using power traces produced by disaggregation algorithms instead of those produced by direct submetering. We use three benchmark algorithms provided in an open source toolkit called NILMTK [3]. The feedback techniques that we developed become almost completely ineffective when using the disaggregated energy traces. In some cases, they failed to identify over 70% of the homes that should be getting feedback and falsely flagged 14% homes of additional homes that should not receive feedback.

To conclude, we discuss why feedback accuracy is low even while disaggregation accuracy is high: accurate *energy breakdown* feedback (i.e. "Your fridge accounts for 8% of your energy bill") can be given even if the power traces have many errors as long as those errors average out over time. Our results indicate that the disaggregation community needs to revisit the metrics by which it measures progress. Part of this process will be to look through the lens of applications, to find the aspects of power traces that are most important.

II. Data sets

We now describe the two data sets that we will be using throughout the rest of this paper. We use the Dataport data set [7], the largest publicly available dataset containing submetered and aggregate electricity consumption. The first release of the data set contains minutely power readings across different appliances from 240 homes in Austin, Texas from January through July 2014. Since our fridge work predates the latest release, we use the first release made available in NILMTK [3] format consisting of data from 240 homes for fridge analysis.

The data set contains power data logged every minute for 172 fridges. Of these, we filtered out 77 fridges that had data collection problems such as missing data and multiple appliances on the same sensor. We use the remaining 95 fridges for evaluation of our proposed techniques. We use the 58 homes having both the temperature setpoint and power data information in our analysis.

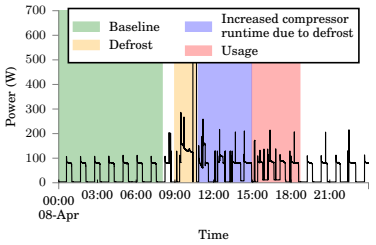

Fig. 1. Breakdown of fridge energy consumption

We also collected data from four identical fridges operated in identical ambient conditions across four floors of our CS building. We put Hobo loggers to collect power data at 1 Hz frequency from these four fridges. For one of the fridge to which we had easy access to, we collected door status for both doors and the freezer unit and internal temperature data at 1 Hz frequency, in addition to the power data. We collected data for two weeks.

III. Appliance Energy Modelling

The key idea behind the energy models is to extract features from the power data which serve as the basis for the energy feedback methods that we later describe in Section IV.

A. Fridge energy modelling

A fridge is a compressor based appliance where the motor duty cycles to maintain the fridge at a set temperature. When the compressor is ON, the refrigerant transfers heat from inside the fridge to the outside. The compressor turns ON and OFF at a small offset temperature above and below the set temperature. Since the fridge is operated at a lower temperature than the surroundings, there is always heat leakage from the outside into the inside of the fridge, which is proportional to the temperature difference between the fridge setpoint and ambient temperature. In the absence of fridge usage (such as opening fridge door), the compressor typically duty cycles at the same rate, shown as the **baseline** compressor usage in Figure 1 which occurs in the early morning hours of the shown fridge. Each time the fridge is opened, the leakage from the ambient environment increases and the compressor has to run longer to remove this extra heat. The addition of items in the fridge also causes the compressor to run longer due to the increased thermal mass. Both these factors cause an increase in the duty percentage of the fridge. The increased compressor ON and decreased compressor OFF durations are shown as **usage** in Figure 1. For efficient running of the fridge, fridges defrost periodically to get rid of frost developed on the cooling coil. **Defrosting** is done via the defrost heater and introduces heat into the system, which is removed in the next few compressor cycles having higher duty percentage. These cycles can be seen in Figure 1.

Thus, the fridge energy consumption can be broken down into three components: usage, defrost and baseline. We now describe the procedure for breaking down fridge energy into these three components:

1. Finding baseline duty percentage: Duty percentage of a fridge cycle (c) is given by the ratio of the compressor ON duration to the total fridge cycle. Or,

$$\text{Duty percentage (c)} = \frac{\text{ON duration(c)}}{\text{ON duration(c)+OFF duration(c)}}$$

Baseline duty percentage is found as the median of the duty percentage during early morning hours (1 to 5 AM) over the duration of the dataset. Using median overcomes the cases when a home may have high fridge usage on some days.

2. Finding defrost energy: Defrost energy comprises of two parts: energy consumption when the fridge is in the defrost state and the extra energy consumed in the regular compressor cycles that follow the defrost state. We assume that a defrost cycle causes an impact on the next D compressor cycles. For these D cycles, the extra energy consumed is found by the additional duty percentage over the baseline of the compressor cycles following the defrost cycle as:

$$\begin{aligned}\text{Extra compressor energy due to defrost} &= \sum_{c=1}^{D}\big(\text{Duty percentage (c) - Baseline duty percentage}\big) \\ &\times \big(\text{ON duration(c) + OFF duration(c)}\big) \\ &\times \text{Fridge compressor power consumption}\end{aligned} \tag{1}$$

Energy consumption when fridge is in the defrost state can be trivially calculated.

3. Finding usage energy: As a prerequisite to finding usage energy, we need to first find *usage cycles*, which we define as fridge cycles that are affected by fridge usage. After removing the defrost cycles and the subsequent D cycles, we look for cycles having duty percentage that is $P\%$ more than the baseline duty percentage. The intuition behind choosing a parameter P is that fridges may show some inherent variation in duty cycle percentage independent of usage. We assume that this variation is within $P\%$ of the baseline duty percentage. After finding these U usage cycles, the usage energy can be calculated as:

$$\begin{aligned}\text{Usage energy} &= \sum_{c=1}^{U}\big(\text{Duty percentage (c) - Baseline duty percentage}\big) \\ &\times \big(\text{ON duration(c) + OFF duration(c)}\big) \\ &\times \text{Fridge compressor power consumption}\end{aligned} \tag{2}$$

4. Finding baseline energy: All the cycles that are not affected due to defrost or usage contribute towards baseline energy and their energy consumption can be summed to find baseline energy.

1) Evaluation of fridge model

We now evaluate the accuracy of our fridge modelling approach. We use our collected data from our CS building for this evaluation as the Dataport data set does not have labels for fridge usage. Using door sensor data, we manually annotated 3 days for usage cycles from the fridge for which we had instrumented in our data set. We found that the defrost cycle impacts the next 3 cycles, and we thus chose D=3. It should be noted that choosing a slightly different value of D is only going to change marginally the usage and defrost energy numbers since defrost cycles are easily outnumbered by regular cycles. The other parameter in our evaluation, percentage threshold (P) for labelling usage cycles is more important due to the expected high number of usage cycles.

We now define three metrics to evaluate our fridge model:

1) % Usage energy error for fridge, which suggests how accurately our model captures the energy usage when a fridge is being actively used:
$\frac{|\text{Predicted fridge usage energy - Actual fridge usage energy}| \times 100\%}{\text{Actual fridge usage energy}}$
2) Precision on fridge usage cycles:
$\frac{|\text{Correctly predicted fridge usage cycles}|}{\text{\# Predicted fridge usage cycles}}$
3) Recall on fridge usage cycles:
$\frac{|\text{Correctly predicted fridge usage cycles}|}{\text{\# Total fridge usage cycles}}$

Figure 2 shows the usage energy error, precision and recall on usage cycles as they vary with P. At a P of 11-16%, the usage energy error is less than 2%. Usage energy error remains below 4% for P between 9 and 24, showing that the prediction remains useful within a wide percentage threshold. A precision of 1 is not observed until P = 17% due to the presence of a single fridge cycle having a high duty percentage despite being unrelated to usage. This is due to the fact that rare cycles may show an inherent deviation from the regular duty percentage. At P = 11%, the recall drops from 1.

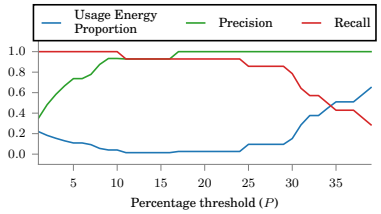

Fig. 2. Our model for breaking fridge energy into usage, baseline and defrost is accurate to within 4% energy error for a wide range of percentage threshold above baseline duty percentage.

This is due to a usage cycle which shows less than 10% deviation from baseline duty percentage. We can conclude that our model is applicable even within a broad range of parameters.

B. HVAC energy modelling

By optimising the HVAC setpoint schedule, upto 30% of HVAC energy can be saved [6]. Giving homes feedback on their setpoint schedule is likely to have a big impact. Thus, we try to build an HVAC model to predict setpoint temperature from HVAC energy data. Since HVAC energy usage is highly dependent on external weather conditions, we incorporate weather data into our HVAC model. While we explain our model for the cooling season (summers, when HVAC is used for cooling), it is equally applicable to the heating season. Our model is based on the following assumptions:

1)HVAC energy is impacted by weather conditions such as wind speed and temperature.

2)HVAC energy consumption is proportional to the difference in external temperature and home setpoint temperature.

3)Programmable thermostats use the following four setpoint times: night hours from 10 PM to 6 AM; morning hours from 6 AM to 8 AM; work hours from 8 AM to 6 PM; evening hours from 6 PM to 10 PM. These times are as per the schedule times reported by EnergyStar.gov.

3)HVAC energy during an hour is zero if the HVAC was not used during this hour

Based on the first assumption, we have: HVAC energy $\propto$ humidity; HVAC energy $\propto$ wind speed. Based on the second assumption, we have HVAC energy $\propto$ (External temperature- internal temperature setpoint). Based on the third assumption, we have four different temperature setpoints during the day. We use four proportionality constants (a_1 through a_4) corresponding to these four setpoint times, describing how strongly the temperature delta between external and setpoint temperature affects HVAC energy consumption. To convert our HVAC model into a regression model, we add a binary variable (is it n^{th} hour) which is 1 if the data is from the $n^{th} hour$ and 0 otherwise. We also use a binary variable indicating if HVAC was used during the n^{th} hour based on the fourth assumption. Our non-linear model has a total of 10 parameters: a_1 through a_6 and four setpoint temperatures.

We now evaluate our HVAC model on its ability to learn the temperature setpoints. We calculate hourly HVAC energy usage for the 58 homes containing both HVAC power and setpoint information. We download hourly weather data from Forecast.io web service and use linear interpolation to fill missing readings, similar to earlier work. Finally, we used non-linear least squares minimisation to estimate the 10 parameters in our model. We also constrain learnt setpoints to be within 60 and 90F. A key takeaway which we see later in section IV-C is that these learnt parameters are useful in providing feedback to homes for setpoint optimisation.

IV. Energy feedback methods

We now develop and demonstrate some examples of how NILM could be used to provide feedback to users to reduce their energy usage based on the appliance energy modelling we previously discussed. These are only examples, and the analysis presented later in this paper would apply to any applications of NILM.

A. Fridge usage feedback

In this section, we target homes based on fridge usage, where the potential feedback could be to reduce interactions with fridge, increase temperature setpoint, etc. We use robust estimator of covariance based outlier detection to detect such homes. The outlier detection method is applied on two dimensions: usage energy% and proportion of usage cycles. We apply this outlier detection method on the 95 homes from the Dataport data set. We divide this two dimensional home data into four quadrants through the medians on usage energy% and proportion of usage cycles. 13 outlier homes can save upto 23% of their fridge energy usage.

B. Fridge defrost feedback

Our method for providing feedback based on defrost is similar to the method of providing feedback based on usage. High defrost energy could be indictive of a broken fridge seal. We use outlier detection methods on two dimensions: defrost energy% and number of defrost cycles per day and give feedback to the homes lying in the first and the second quadrant. Number of defrost cycles per day is more interpretable and relatable than proportion of defrost cycles (which is going to be a very small floating point number). 17 homes can save up to 25% of their fridge energy if they fix their fridge defrost.

C. HVAC setpoint feedback

We use machine learning methods on the parameters learnt previously from the HVAC model, for finding homes needing HVAC feedback. We calculate an HVAC efficiency score for the 58 homes in the Dataport data set on a scale of 0 to 4 based on recommended setpoint temperature from EnergyStar as follows: 1)Morning score = 1 if morning setpoint temperature >78F, 0 otherwise; 2) Evening score = 1 if evening setpoint temperature >78F, 0 otherwise; 3) Work hours score = 1 if work hours setpoint >85 F, 0 if setpoint $\leq$ 78, (85-setpoint)/7 otherwise; and 4) Night score = 1 if setpoint >82F, 0 if setpoint $\leq$ 78F, (82-setpoint)/4 otherwise. We decide that 34 homes that have an overall score of 2 or less can be given feedback to optimise their HVAC setpoints. In addition to the 10 parameters of the HVAC model, we add additional features such as total energy used in work, morning, night and evening hours and the number of minutes HVAC system was on during these times to our machine learning methods We use 2-fold cross validation and a grid search on the feature space to find that the feature <**a_1, a_3, Energy in evening hours, Mins HVAC usage in morning hours**> used by the Random Forest classifier give the optimal accuracy of 84.4%.

V. Evaluation of NILM for feedback

We now evaluate how accurately do current NILM approaches match these feedback. We now describe the experimental setup for evaluating NILM performance on the Dataport data set.

A. Experimental setup

We use NILMTK [3] to perform our NILM experiments. We use the 3 reference implementations made available in NILMTK: combinatorial optimisation (CO), factorial hidden Markov model (FHMM) and Hart's steady state algorithm. These 3 algorithms cover both classes of disaggregation algorithms- CO and FHMM are non-event based, while Hart's algo is event based. While FHMM is a recent development, Hart's is the seminal work. We use the standard definition of NILM metrics as made available in NILMTK [3]: %

Dataset	Algorithm	Fridge			HVAC		
		RMSE (W)	Error Energy %	F-score	RMSE (W)	Error Energy%	F-score
REDD	Additive FHMM	-	62.5	-	-	-	-
REDD	Difference HMM	83	55	-	-	-	-
Colden	Bayesian HMM		45				
iAWE	FHMM	-	50	**0.8**	-	30	**0.9**
Data port	CO⋆	85	**19**	0.65	**600**	**15**	0.87
Data port	FHMM⋆	95	20	0.63	650	18	0.89
Data port	Hart	**82**	21	0.72	890	23	0.76

TABLE I

BENCHMARK ALGORITHMS ON THE DATAPORT DATASET GIVE COMPARABLE PERFORMANCE TO EXISTING LITERATURE.

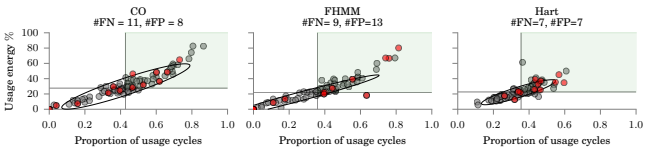

Fig. 3. NILM algorithms show poor accuracy in identifying homes which need feedback for high fridge usage energy. Red dots indicate the homes which should be getting feedback based on analysis of submetered fridge data, while these algorithms would give feedback to all homes in the green region outside the elliptical boundary.

Error in Energy, Root Mean Squared Error (RMSE) Power, and F-score, to evaluate these NILM models.

We now present the results of NILM evaluation on the Dataport data set. We also compare our results with the state of the art. From Table 1, we can see that for both fridge and HVAC, the benchmark algorithms we use are comparable in performance to existing literature.

B. Fridge usage feedback

We now see how accurate fridge usage feedback we can provide with the disaggregated power trace. Figure 3 shows that all three NILM algorithms have poor accuracy in identifying homes that need feedback for high fridge usage. False negatives (FN) are those homes that should be getting feedback but are not getting, and false positives (FP) are those homes that would wrongly get feedback. We now explain the reasons for the poor accuracy of the used NILM algorithms.

During the night hours when typically only background appliances such as fridge are running, Hart's algorithm has good disaggregation accuracy. Due to this, Hart's algorithm closely matches the baseline duty percentage computed on submetered data. However, Hart's algorithm is susceptible to detection of false events and missing true events, especially during active hours when appliances similar in magnitude to the fridge may be operating. Thus, Hart's algorithms underpredicts and overpredicts fridge compressor cycle durations during the day creating a deviation in fridge usage. The median baseline duty percentage found by CO and FHMM are higher than the median baseline duty percentage on submetered data. Owing to higher baseline duty percentage, usage energy in these homes is lower than submetered, thereby explaining the high false negative rate. The reason behind CO and FHMM finding a high baseline duty percentage is that the objective function in both these algorithms includes minimising the difference between aggregate power and sum of power for predicted appliances. To satisfy this objective, these

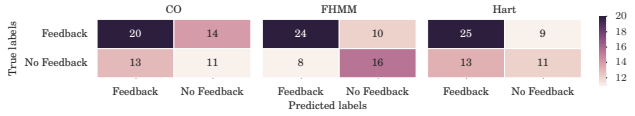

Fig. 4. Classification of homes into those with setback schedules decreases from 84% with submetered power traces to 53%, 69%, and 62% respectively with power traces produced by the three NILM algorithms.

algorithms predict fridge to be ON longer than actual during the night hours when typically few loads are used.

C. Fridge defrost feedback

We find that the our approach of breaking down fridge energy into baseline, defrost and usage is unable to find even a single defrost cycle when fed the disaggregated power data. This is due to the inadequacy of the used NILM methods in effectively learning and disaggregating the defrost state. CO and FHMM rely on KMeans and Expectation Maximisation algorithms respectively for learning the different states of an appliance. Due to defrost events being rare in comparison to regular usage, these algorithms are not able to accurately associate a cluster with the defrost state.

D. HVAC setpoint feedback

Figure 4 shows that the classification of homes into those with setback schedules decreases significantly for all NILM algorithms. We now explain the low classification accuracy based on the features used by Random Forest classifier. Of the four features used, a_1 and a_3 are hard to interpret, and thus we provide an explanation based on *Mins HVAC usage during morning hours*. Most of the HVAC usage in the data set occurs during the night hours. Thus, NILM accuracy is likely to be highly dependent on night time HVAC disaggregation. Since, only HVAC and fridge would be typically used in the night, and, HVAC has a distinct much higher power signature than the fridge, NILM accuracy for HVAC is decent (as per Table 1). However, during the morning hours, when typically there is more activity in the home, NILM accuracy for HVAC is expected to be lesser.

VI. CONCLUSIONS

In this paper, we show that disaggregated power data has the potential to provide targeted, actionable energy feedback to homes. However, we found that the state-of-the-art NILM accuracy isn't effective in enabling such feedback. We believe that the community needs to revisit the metrics for gauging NILM performance, and our work is a step in that direction. We finally conclude that- "If you can measure it, you may not necessarily be able to improve it."

VII. NOTE

This paper is an abridged version of a paper [4] that appeared in Buildsys 2015.

REFERENCES

[1] W. Abrahamse, L. Steg, C. Vlek, and T. Rothengatter. A review of intervention studies aimed at household energy conservation. *Journal of environmental psychology*, 25(3):273–291, 2005.

[2] K. C. Armel, A. Gupta, G. Shrimali, and A. Albert. Is disaggregation the holy grail of energy efficiency? The case of electricity. *Energy Policy*, 52:213–234, 2013.

[3] N. Batra, J. Kelly, O. Parson, H. Dutta, W. Knottenbelt, A. Rogers, A. Singh, and M. Srivastava. Nilmtk: An open source toolkit for non-intrusive load monitoring. In *Proceedings of the 5th international conference on Future energy systems*, pages 265–276. ACM, 2014.

[4] N. Batra, A. Singh, and K. Whitehouse. If you measure it, can you improve it? exploring the value of energy disaggregation.

[5] S. Darby. The effectiveness of feedback on energy consumption. *A Review for DEFRA of the Literature on Metering, Billing and direct Displays*, 2006.

[6] J. Lu, T. Sookoor, V. Srinivasan, G. Gao, B. Holben, J. Stankovic, E. Field, and K. Whitehouse. The smart thermostat: using occupancy sensors to save energy in homes. In *Proceedings of the 8th ACM Conference on Embedded Networked Sensor Systems*. ACM, 2010.

[7] O. Parson, G. Fisher, A. Hersey, N. Batra, J. Kelly, A. Singh, W. Knottenbelt, and A. Rogers. Dataport and nilmtk: A building data set designed for non-intrusive load monitoring. 2015.