# Staggered Sampling for Energy Efficient Data Collection

Jennifer L. Wong
Computer Science Department
SUNY Stony Brook University
Stony Brook, NY 11794
Email: jwong@cs.sunysb.edu

Seaphan Megerian
Electrical & Computer Engineering Department
University of Wisconsin, Madison
Madison, WI 53706
Email: megerian@cs.ucla.edu

Miodrag Potkonjak
Computer Science Department
University of California, Los Angeles
Los Angeles, CA 90095
Email: miodrag@cs.ucla.edu

***Abstract*— Efficient and complete data collection is one of the most important tasks in wireless ad-hoc sensor networks. Additionally, the collection of the full data set should be performed in the most resource efficient way, thus prolonging the battery lifetime of the network. We introduce a new approach for energy efficient data collection through the use of staggered sampling. Staggered sampling means that at each sampling moment (epoch) only a small percentage of sensors collect (sample) data. The proposed approach leverages on statistical relationships between samples taken from different sensors and/or at different epochs for the prediction of the non-sampled sensor data.**

**The main goal of the approach is to ensure complete collection of data during a periodic cycle while minimizing the number of sensor readings collected at any point in time. Complete data collection is confirmed by ensuring that each sensor is either sampled at each epoch or the data sample can be accurately recovered though model prediction of the sampled sensors. The proposed approach consists of two main phases. First, efficient modeling of the prediction relationship between two sensors using kernel smoothing over different time lags is performed. Second, the selection of epochs at which each sensor is to sample the data is determined. A 0-1 integer linear programming formulation is used to address this NP-complete assignment problem optimally on relatively large instances. We demonstrate the effectiveness of the approach on traces from actually deployed networks for sensor of two modalities: temperature and humidity.**

## I. Introduction

One driver application for wireless ad-hoc sensor networks is environment and event monitoring. A primary requirement of this type of application is accurate and complete observability of the environment by the sensors. Simultaneously, all the observation should be collected while minimizing network resources, and therefore prolonging the overall lifetime of the system. The lifetime of the network and each individual sensor is mainly dependent on the amount of sampling and therefore generated communication traffic at the node [1], [2]. In order words, the objective is to sample sensor data at each node as rarely as possible while being able to calculate any missing samples from measurements taken at other nodes. Until now, all of the previous proposed sampling schemes [3], [4], [5] for sensor networks assumed simultaneous sampling (identical epochs) at all nodes. Our goal is to demonstrate that by relaxing this requirement and utilizing time shifted data for data recovery significant improvement in the lifetime of the network can be achieved while maintaining a user specified level of accuracy.

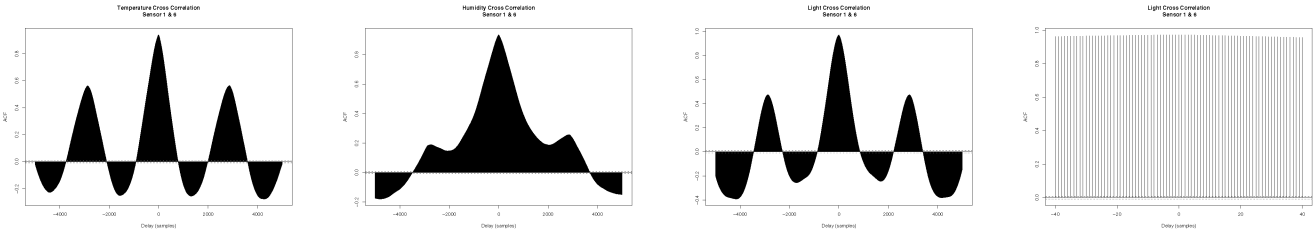
In order to create efficient protocols for staggered sampling, we have developed a two-phase approach for minimizing energy consumption by reducing the required amount of data communication and maximizing the available time for applying sleep methodologies [6], [7]. If we assume that the user specifies that data has to be collected periodically every $W$ time units with precision accuracy of $L_1$ error $\leq p\%$, staggered sampling schedules data collection for each sensor in such a way that that missing data at each sensor at each epoch can be calculated using statistical models and data collected at the same or other sensors in recent epochs. Hence, there are two main technical challenges for the effective application of staggered sampling: (i) the development of accurate prediction models; and (ii) the creation of a schedule as to when each node will sample in order to maintain the specified fidelity for all data streams.

In the first phase of our approach, the initial set of data samples at each sensor is collected in order to form a training set. This training set is used to build smooth and monotonic prediction models of the sensor data between pairs of sensors using non-parametric kernel smoothing. The prediction models are created not only for prediction at the same epoch, but also for phased (time-shifted) prediction of the data at one sensor from another sensor or itself. In the second phase, the prediction models are used to determine the maximum phase difference between all pairs of sensors while maintaining accurate predictions. An integer linear programming (ILP) formulation is generated in order to find the epochs for sampling each sensor while ensuring all sensor data can be predicted from the sampled data with the specified accuracy.

## II. Related Work

Before we start the technical exposition, we briefly survey the most directly related work. Akyildiz et al. [8] provides an introductory survey on sensor networks research. A number of techniques have been proposed to address one of the key issues, power conservation, at all levels of the design process from communication protocols [6] to digital signal processing [9]. Willet et al. introduced backcasting in [4] where adaptive sampling is applied for efficient field estimation. In [3], an adaptive sampling approach is proposed

1-4244-0376-6/06/$20.00 




(a) Temperature (b) Humidity (c) Light (d) Light, Zoom of peak

Fig. 1. Cross correlations for sensor readings at Sensor 1 and 6.

which varies the sampling rate at each sensor and therefore adapting to the streaming-data characteristics of the sensor. The use of mobile sensor nodes are used to determine sampling density required in various environmental regions in [5]. Their Fidelity Driven Sampling actively seeks to minimize error without prior knowledge of the variable field. All of the proposed sampling schemes assumed simultaneous sampling at all nodes. Our goal is to demonstrate that by relaxing this requirement and using time-shifted data for data recovery we can improve the lifetime of the network by more than an order of magnitude while maintaining the user specified level of accuracy. The proposed approach utilizes a 0-1 ILP formulation [10] which is often used for addressing NP-complete optimization problems. A number research problems in sensor networks have been addressed using integer linear programming formulations including broadcast trees [11] and routing [12].

## III. Staggered Sensor Sampling

In this work a standard model of sensor data flow, where all data collected in the network is processed at the data sink (aka gateway or fusion center), is assumed. The data sink has unlimited energy constraints and sufficient processing resources. Additionally, the community standard, where the main component of energy consumption is communication, is also assumed. In accordance with proposed sleep mechanisms, the assumption is made that is a sensor node is not sampling data, it is in the minimal energy consumption state, ie. sleeping. Finally, the staggered sampling approach presented in this work is sensor data-driven. This means that it has to be conducted on data collected from actual deployed networks, and this data must be at least partially cross-correlated and predictable. The analysis in this work was performed using the sensor network and sensor data collected at Intel Berkeley Labs, which is a dataset which satisfies these conditions.

The staggered sampling approach is divided in two phases. First, efficient modeling of the prediction relationship between pairs of sensors using kernel smoothing over different time lags is performed. We assume that a data training set is collected and processed off-line in order to build the phase prediction models (see Section IV). The maximum phase delay for each sensor prediction pair is determined by examining the error in each phased prediction model, the model with the largest phase which still satisfies the specified user accuracy ($p\%$) for prediction is then selected as the maximum delay. In the second phase, the staggered sampling problem is formulated and addressed as an ILP problem (see Section V). Formally, the staggered sampling problem is defined as follows.

---

**Problem:** *Staggered Sampling Problem*
**Instance:** *An $ixj$ integer matrix $R$ of "maximum rephasing", a positive integer $W$ for "window size", and a positive integer $S$ of "maximum samples".*
**Question:** *Is there an assignment of each sensor $i$ to at most $S$ time steps in $W$ s.t. each sensor $i$ at each time step $t_i$ in $W$ is assigned to $t_i$ or at least one sensor $j$ at $t_k$ where $(t_k - t_i)\%W \leq R_{ij}$?*

---

The staggered sampling problem is an NP-complete problem. The Domatic Number Problem [13] is a special case of the staggered sensor sampling problem.

## IV. Modeling

In order to evaluate the suitability of the data traces for staggered sampling, we first conducted exploratory statistical analysis to evaluate the potential for time-shifted accurate statistical modeling. Figure 1 shows linear cross-correlation for readings at a typical pair of sensors. In the first figure, we can observe the diurnal trends of temperature data. Similar cross correlations were obtained for humidity. The last subfigure indicates that cross-correlations often changes at a very slow rate as the time shift increases.

For time-shifted prediction we used an interleaved application of kernel smoothing and monotonicity. Smoothing is important in order to improve the accuracy of the model and to compensate for otherwise insufficient amounts of data for some ranges of sensor values. Monotonicity is important in order to enforce the natural requirements that sensors that are predicting each other well and are, therefore, exposed to the same set of stimuli simultaneously increase or decrease their values as the intensity of the stimuli changes. In order to simultaneously achieve both smoothness and monotonicity we iteratively and interchangeably applied kernel smoothing regression tools and monotonic regression tool.

For kernel smoothing we tried several variants of Nadaraya-Watson kernel-weighted average with Epanicechnikov, tricube, and Gaussian kernels [14] and selected the one with the smallest Akaike criterion [14]. For monotonic smoothing we used Stout's version [15] of Ayer et al. "pair adjacent violators (PAV)" algorithm that iteratively replaces each pair of adjacent segments that violate monotonicity constraint by a new single horizontal segment that is optimal with respect to the selected $L_p$ norm and removes the violation. We used an $L_1$ norm.

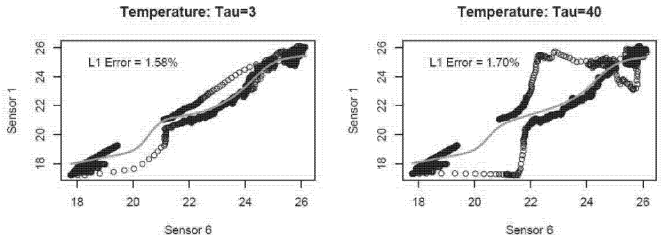



Fig. 2. Interleaved Nadaraya-Watson/PAV modeling for temperature readings of Sensor 6 predicting Sensor 1 with Delay 3 and 40.

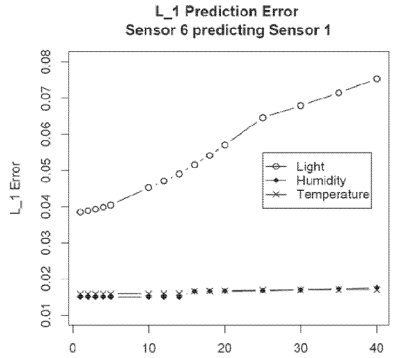



Fig. 3. $L_1$ Prediction Error using interleaved Nadaraya-Watson/PAV modeling for Sensor 6 predicting Sensor 1.

The procedure is terminated once the PAV algorithm does not induce any changes.

Figure 2 shows typical resulting smooth and monotonic models. Figure 3 shows prediction errors for sensors of three modalities obtained using our modeling approach. We see that one can predict accurately all three modalities, but humidity and temperature models are significantly more accurate and less sensitive to time shifts.

## V. Integer Linear Programming Approach

In this section we present the integer linear programming (ILP) problem formulation for addressing the staggered sensor sampling for energy efficiency problem. The problem is addressed in the scenario where the size of the periodicity window is known, along with the maximum number of allowable samples per sensor in the window. In addition to this information, it is assumed that the prediction ability of all pairs of samples is known with consideration for the delay between the predictor and predicted sensor sample. The goal is to assign the minimum number of samples per sensor in the window such that all sensor readings for all sensors at each epoch are either measured or predictable from measured samples.

ILP uses several sets of known constants. The first set of constant values, $R_{ij}$, defines the duration (delay) for which of one sensor has the ability to predict another sensor. The second constant value indicates the length of the considered periodicity window, $W$.

$$\begin{array}{rcl} R_{ij} & - & \text{max re-phasing for sensor } j \text{ to predict } i \\ W & - & \text{size of periodicity window} \end{array}$$

$$\begin{array}{rcl} x_{ik} & = & \begin{cases} 1, & \text{if sensor } i \text{ is sampled at epoch } k \\ 0, & \text{otherwise.} \end{cases} \\ p_{ik} & = & \begin{cases} 1, & \text{if sensor } i \text{ is predictable at epoch } k \\ 0, & \text{otherwise.} \end{cases} \\ l & = & \text{largest number of samples by any sensor} \end{array}$$

We use two sets of variables and a single variable. The first set, $x_{ik}$, denotes the assignment of sensor $i$ to measure a sensor reading at epoch $k$. We define a set of variables which specify if a sensor reading $i$ is predictable by any sensor at epoch $k$. In order for sensor $i$ to be predictable, at least one sensor $j$ which can predict $i$ (ie. $R_{ij} \geq 0$) must be sampled within $R_{ij}$ of epoch $k$. The final variable, $l$, is used to represent the largest number of samples taken by any sensor in the window.

Our ILP formulation uses three sets of constraints. The first set enforces a necessary constraint to calculate the maximum number of samples taken by each sensor for the objective function. Therefore, Eq. (1) ensures that variable $l$ is at least the sum of samples taken by each sensor. Additionally, at each epoch in the window $W$ each sensor must be either sampled or predictable from another sensor. This constraint is specified by Eq. (2).

$$\text{for all } i : \sum_k x_{ik} \leq l \tag{1}$$

$$\text{for all } i,k : x_{ik} + p_{ik} \geq 1 \tag{2}$$

$$\text{for all } i,k : \sum_j \sum_{t=0}^{R_{ij}} x_{j[(k-t)\%W]} \geq p_{ik} \tag{3}$$

$$Y = MIN(l) \tag{4}$$

The final constraint ensures that the predictability variable for each sensor at each epoch is assigned properly (ie. $p_{ik} = 1$ if and only if at least one sensor $j$ can predict sensor $i$ within the re-phasing value $R_{ij}$). For each sensor $i$ at each possible epoch $k$ the predictability variable, $p_{ik}$, is calculated. In the on-line case, which we are considering, the value is predicted from previously sampled sensor readings. Therefore, for each possible predictor sensor $j$ if $j$ is sampled at any epoch between $k - R_{ij}$ and $k$ then $p_{ik}$ is predictable. To formulate this constraint we calculate the sum of all sampled sensors $x_j$ which occur within the time period $t$=$(k - R_{ij}), \ldots, k$. If there is no sensor reading which can be used to predict sensor $i$ at epoch $k$ (ie. summation is zero), then $p_{ik}$ must be assigned to zero. However, if the summation is one or more, then $p_{ik}$ can be assigned to zero or one. This is acceptable because Eq. (2) will ensure that sensor $i$ is sampled at time $k$ or that it is predictable, and therefore forcing $p_{ik}$ to one if necessary.

Note that in Eq. (3) we denote the calculation of the time period for predictability of each sensor using modulus $W$. Since $W$ is the duration of the periodicity window, it is acceptable for a sensor to be predicted from a previous window under the assumption that the epoch it is within the specified re-phasing. Therefore, the modulus of time position of the sampled sensor reading is taken into account for the re-phasing.

The main goal of the problem is to determine which sensor(s) to sample at each of the epochs in the periodicity

window such that the maximum number of samples taken by any sensor is minimized. Therefore, the objective function for the problem is to minimize the largest number of sensor readings for any sensor, $l$. By optimizing the problem in this form, we ensure that no sensor is overly energy drained by the staggered sampling approach, and prolong the lifetime of all nodes in the network by distributing the sampling and prediction as equally as possible over the nodes.

There are three potential scenarios for defining the rephasing relationship. Our ILP formulation addresses the on-line case. When considering prediction from future data samples only, the only modification to Eq. (3) which is to change the subscript of $x_{j[(k-t)\%W]}$ to $x_{j[(k+t)\%W]}$ on the left side of the equation. This modification specifies that the prediction come from a sample in the future (after the current time $k$). In the off-line case, where all data samples are known, the constraint is the combination of the on-line and future only case. Specifically, the constraint becomes the double summation of $(x_{j[(k-t)\%W]} + x_{j[(k+t)\%W]})$.

## VI. Experimental Results

In our experimentation of the staggered sampling technique we used temperature and humidity samples taken from the Intel Berkeley dataset [16]. Our analysis was performed with comparison to a base case where each sample can only be predicted from other samples taken in the same time moment. All of the ILP formulations were solved using the CPLEX solver with a maximum runtime of 5 minutes, but almost all instance running within seconds. Three sets of $L_1$ errors were considered 2%, 3%, and 5%. In addition, three instances of the dataset were examined: all sensor nodes, a set of approximately two-thirds of the nodes, and a set of nodes from one third of the area according to the layout.

In Table I we present the staggered sampling results for temperature. In the first column we show the number of nodes, followed by the amount of $L_1$ prediction error considered. The next two columns show the improvement for the base case where all sensors are used only for prediction of other sensors in the same epoch. The fifth column shows the amount of improvement for the staggered sampling approach over the base case. We see that as the amount of error allowable is increasing, the savings decreases. This is due to the fact that with increased model prediction error a higher number of sensors can predict other sensors over long time-shifts. The most important result is that even for very low error, 2%, the staggered sampling approach is capable of performing 20 times better than the standard base case sleeping strategy, translating into 20 times longer lifetimes for the network. Analogous results are shown for humidity in the second half of Table I. For humidity we see the same patterns, occurring as we increase the allowable error. Humidity overall has lower improvement over the base case, because humidity shows a more complex statistical relationship. Nevertheless, the staggered sampling approach still was able to achieve 12 times more energy savings over the base case.

| # | $L_1$ | Temperature | | | Humidity | | |
|---|---|---|---|---|---|---|---|
| Nodes | Error | Base | SSamp | Ratio | Base | SSamp | Ratio |
| 51 | 0.02 | 1 | 21 | 21.0 | 2 | 22 | 11.0 |
| 34 | 0.02 | 2 | 34 | 17.0 | 2 | 22 | 11.0 |
| 17 | 0.02 | 6 | 120 | 20.0 | 4 | 44 | 11.0 |
| 51 | 0.03 | 2 | 44 | 22.0 | 2 | 22 | 11.0 |
| 33 | 0.03 | 4 | 44 | 11.0 | 4 | 34 | 8.5 |
| 17 | 0.03 | 8 | 42 | 5.3 | 4 | 34 | 8.5 |
| 51 | 0.05 | 24 | 60 | 2.5 | 10 | 60 | 6.0 |
| 34 | 0.05 | 40 | 200 | 5.0 | 10 | 120 | 12.0 |
| 17 | 0.05 | 16 | 160 | 10.0 | 4 | 44 | 11.0 |

TABLE I

Experimental results for temperature and humidity.

## VII. Conclusion

We introduced a staggered sampling approach for energy efficient data sampling and collection in wireless ad-hoc sensor networks. It schedules sampling of each sensor in potentially different moments in such a way that it enables energy conservation by minimizing the amount of data traffic required by each sensor node and by enabling periods for entering a low power sleep state. The two-phase approach first builds smooth and monotonic prediction models for sensor prediction for various time lags using a training set of sampled data. Secondly, the models are used to build an ILP formulation that optimally determines the staggered sampling assignments for a time period and maximum number of samples per node such that each data point is obtainable within a specified accuracy level. The effectiveness of the model and approach for energy savings are evaluated on real-life data traces.

## References


[1] O. Kasten, "Measurements of energy consumption for digitan 2 mbps wireless lan module (IEEE 802.11/ 2mbps)," 2001.

[2] A. Boukerche *et al.*, "Energy-aware data-centric routing in microsensor networks," in *MSWiM*, 2003, pp. 42–49.

[3] A. Jain and E. Y. Chang, "Adaptive sampling for sensor networks," in *DMSN*, 2004, pp. 10–16.

[4] R. Willett *et al.*, "Backcasting: adaptive sampling for sensor networks," in *IPSN*, 2004, pp. 124–133.

[5] M. A. Batalin *et al.*, "Call and response: Experiments in sampling the environment," in *Sensys*, 2004, pp. 25–38.

[6] W. Ye *et al.*, "An energy-efficient mac protocol for wireless sensor networks," in *Infocom*, June 2002, pp. 1567–1576.

[7] C. Schurgers *et al.*, "Optimizing sensor networks in the energy-latency-density design space," *IEEE Transactions on Mobile Computing*, vol. 1, no. 1, pp. 70–80, 2002.

[8] I. Akyildiz *et al.*, "A survey on sensor networks," *IEEE Communications Magazine*, vol. 40, no. 8, pp. 102–116, 2002.

[9] A. Wang and A. Chandrakasan, "Energy-efficient DSPs for wireless sensor networks," *IEEE Sig Proc Magazine*, vol. 43, no. 5, pp. 68–78, 2002.

[10] A. Schrijver, *Theory of linear and integer programming*. New York: Wiley, 1986.

[11] A. Das *et al.*, "Minimum power broadcast trees for wireless networks: integer programming formulations," in *INFOCOM*, 2003, pp. 1001–1010.

[12] J. Chang and L. Tassiulas, "Maximum lifetime routing in wireless sensor networks," *IEEE/ACM Trans. Netw.*, vol. 12, no. 4, pp. 609–619, 2004.

[13] M. R. Garey and D. S. Johnson, *Computer and Intractability: A Guide to the theory of NP-Completeness*. W. H. Freeman & Co., 1979.

[14] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2001.

[15] Q. Stout, "Optimal algorithms for unimodal regression," *Computing Science and Statistics*, vol. 32, 2000.

[16] M. Paskin *et al.*, "A robust architecture for distributed inference in sensor networks," in *IPSN*, 2005.