INTERACTING PARTICLE-BASED MODEL FOR MISSING DATA IN SENSOR NETWORKS: FOUNDATIONS AND APPLICATIONS

Farinaz Koushanfar[1], *Negar Kiyavash*[2], *Miodrag Potkonjak*[3]

[1] ECE Department, Rice University
[2] ECE Department, University of Illinois, Urbana-Champaign
[3] CS Department, University of California, Los Angeles

ABSTRACT

Missing data is unavoidable in sensor networks due to sensor faults, communication malfunctioning and malicious attacks. There is a very little insight in missing data causes and statistical and pattern properties of missing data in collected data streams. To address this problem, we utilize interacting-particle model that takes into account both patterns of missing data at individual sensor data streams as well as the correlation between occurrence of missing data at other sensor data streams. The model can be used in algorithms and protocols for energy efficient data collection and other tasks in presence of missing data.

We use statistical intersensor models for predicting the readings of different sensors. As a driver application, we address the problem of energy efficient sensing by adaptively coordinating the sleep schedules of sensor nodes while we guarantee that values of nodes in the sleep mode can be recovered from the awake nodes within a user's specified error bound and probability of missing data at awake nodes is less than a given threshold. The sleeping coordination is addressed by creating the maximal number of subgroups of disjoint nodes, each of whose data is sufficient to recover the data of the entire network in presence of missing data. On simulated and actually collected data for temperature and humidity sensors in Intel Berkeley Lab, we show that by using sleeping coordination that considers missing data, we reduce the typical 40% missing data rate of traditional sleeping techniques to less than 7%.

1. INTRODUCTION

Missing data is unavoidable in sensor data collection. Recovery of missing data is a canonical task in sensor networks and can be used for a variety of applications, including compression, fault and attack detection and calibration. In order to characterize properties of missing data, we analyzed data streams collected at Intel Berkeley Lab where 54 MICA-2 motes sampled light, temperature, and humidity sensors, each 30 seconds. The radios on the MICA-2 motes have an outdoor transmission range of around 300m. Even though the radio range decreases in the indoor environment, the transmission range of the radios are still more than the distances of the nodes deployed and their distances to the server. For the purposes in this paper we assume that all sensor nodes can directly communicate to the server.

Our starting point for addressing properties of missing data is statistical and simulation model of missing data. The model takes into account not only patterns and frequencies of missing data in each stream, but also the mutual cross-correlations between the different node streams. Nevertheless, the model is conceptually simple and computationally fast. We believe that there are three main causes for missing data: lossy links [1], collision of data at the MAC layer during collection of data in direct one hop communication from each node to the gateway [2], and transient malfunctioning of the data collection and communication software due to nested interrupts [3].

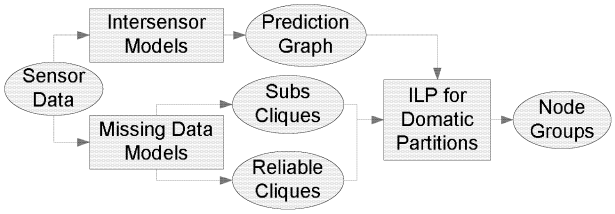

Fig. 1. Global flow of the approach.

We also use intersensor models that quantify the relationship between the sensor measured value at different sensors. We have developed intersensor models for all pairs of nodes such that one node can be used to predict readings of an another. Given a time series of data measurements from two sensors, it is natural to ask whether the values sensed by one sensor can be predicted the other, i.e., can sensor Y can be predicted via some function of sensor X's data, $Y = f(X)$. Regression analysis uses data samples from both X and Y to find the function f. For this task we use new combinatorial isotonic regression technique, that outperforms the standard parametric and nonparametric regressions [4].

Using the intersensor prediction models, we build a graph,

called a *prediction graph*, in which a directed edge from sensor node i to node j exists only if sensor node i can predict the value that node j senses to within a target error rate. Using the interacting particle models for missing data, we find two types of node groups in the networks. The first type of node group is denoted as *reliable clique* and has the property that at least one node from the clique is present at each measurement epoch with a probability of more than $\rho\%$. The second type of node group is denoted as *substitute clique*, where each clique is substituting a particular node. It has the property that when its corresponding node is not present, the nodes in the clique could recover the missing data from that node with more than $\rho\%$ probability.

We seek to find subgroups (or partitions) of nodes such that each subgroup can accurately predict the sensed values for the entire network while the percentage of missing data in the subgroup is less than $1 - \rho\%$. We propose the idea of choosing these groups to be disjoint dominating sets that are extracted from the prediction graph using an ILP-based procedure. Each dominating set has the property that at least one reliable clique associated with each node in the set is included. Also, for each node outside the set, at least one substitute clique should be included. The ILP-based procedure yields mutually disjoint groups of nodes called *domatic partitions*. The energy saving is achieved by having only the nodes in one domatic set be awake at any moment in time. The different partitions can be scheduled in a simple round robin fashion. If the partitions are mutually disjoint and we find K of them, then the network lifetime can be extended by a factor of K. The global flow of the approach we have just described is depicted in Figure 1.

2. INTERACTING PARTICLE MODEL FOR MISSING DATA IN MULTIPLE SENSOR STREAMS

Our first step is development of models that capture statistics and time-dependent dynamic patterns of missing data. Figure 2(a) shows a histogram of the number of nodes for a specified level of missing data shown on the x-axis. We see that the majority of nodes have around 50% of data missing. Figure 2(b) show the histogram of probability of missing data in all epochs (time intervals within each nodes is sampled). We see that there is a significant variation in the percentage of the available data at different nodes and epochs.

Figure 3 present boxplots of number of node pairs (n_i, n_j) for different conditional probability of missing data (x) at one node n_j when data at node n_j is available (o) and missing (x) respectively. The boxplots are shown for all node pairs. The key observation is that the conditional probabilities have significantly higher ranges than probabilities of individual missing data. The missing data for a pair of nodes can be both positively and negatively correlated. Figure 4 shows the distribution of intervals where for one epoch, the consecutive data collection was always successful or unsuccessful (missing) for the node n_1. therefore, to capture properties of missing data in sensor streams, one has to simultaneously consider both time dependencies of missing data within each stream as well the dependencies of missing data among the different streams.

To address these simultaneous requirements, we have developed an interacting particle model [5, 6] for missing data. The conceptual novelty that enabled high statistical accuracy of the model is the application of non-parametric kernel smoothing techniques for modeling [7]. In the interactive particle model, each sensor is represented as a node with two states: available and missing. At each time moment the availability of data at one sensor is being modeled using the previous state of availability of data at that sensor and the previous state of availability of data at the other sensors.

Each node makes the decision weather to alter its current state using a voting mechanism. Each node in the network casts its vote using a probabilistic mechanism and the pertinent node changes its state only if majority of the votes are for the change. Each node n_j decides probabilistically its vote for node n_i by considering statistically derived conditional probability that node n_i has missing data in the next epoch if node n_j is in the pertinent missing or available data state in the current epoch. Specifically, we generate a random number in interval [0,1] with uniform probability and the node votes for change if the number is larger than the pertinent conditional probability. Because of space limitations, we will not discuss the details of interactive particle models that is used for generation of large instances and for long simulation of protocols.

Using the missing data models, we form groups of nodes, such that at each point of time, at least one measurement from the group is present with more than $1 - \rho\%$ probability. We call such groups of nodes *reliable cliques* and denote them by A_r, $r = 1, \ldots, R$. Each A_r is a vector with elements a_{ri}, $i = 1, ..., N$ where $a_{ri} = 1$ if node v_i is in the clique A_r and is 0 otherwise. We also form another set of node groups substituting each specific node. The substituting nodes have the property that at least one measurement from the group is present with more than $1 - \rho\%$ probability. We call such

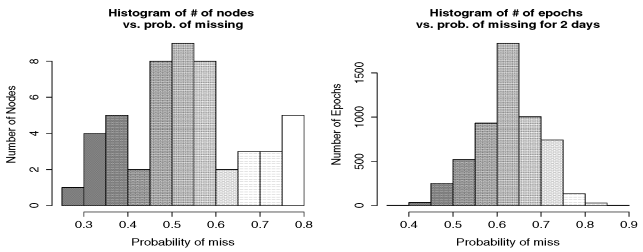

Fig. 2. Histograms of: (a) number of nodes for different missing probabilities, and (b) probability of data missing for different epochs in a 2 day period.

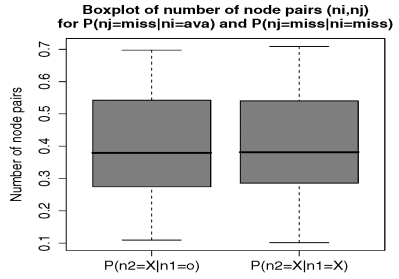

Fig. 3. Boxplots of: (a) conditional probabilities $P(n_j = missing|n_i = available)$, and (b) conditional probabilities $P(n_j = missing|n_i = missing)$ for all node pairs (n_i, n_j).

groups of nodes *substitute cliques* and denote them by B_s, $s = 1, \ldots, S$. Each B_s is a vector with elements b_{si}, where $b_{si} = 1$ if node v_i is not in the substitute clique and $b_{si} = 0$ clique. We also have a set of auxiliary variables d_{sj} where $d_{sj} = 1$ if the clique B_s substitutes node v_j and is 0 otherwise.

3. SLEEPING COORDINATION IN PRESENCE OF MISSING DATA

Placing the nodes in a network to sleep has been demonstrated to be an exceptionally effective strategy for prolonging the network's lifetime [8]. Maintaining sensing quality is ensured by strategically placing a subset of nodes in sleep mode in such a way that, from the remaining small set of awakened nodes, one can recover the data at the sleeping nodes to within a user specified target error rate while on the missing data rate at the awake nodes is less than a given probability $1 - \rho\%$. We call this problem the *sleeping coordination* problem. The problem can be formulated as follows.

Problem: Missing Data Recovery-based Domatic Partitions.

Instance: a directed graph $G = (V, E)$, where we denote the vertices as $v_i \in V$, $i = 1, \ldots, N$ and the edges by E.

Question: *Is there a partition of vertices in the graph to K disjoint sets, S_1, S_2, ..., S_K, such that for each set S_k, the subset $S_k \subseteq S$ is such that all nodes in each graph G that are not in S_k have at least one incoming edge from a node in S_k and, for each vertex $v_i \in S_k$ there is at least one reliable clique including v_i and, for each $v_j \notin S_k$ there is at least one substitute clique for v_j?*

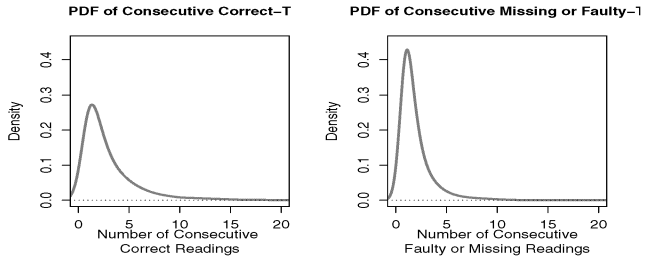

Fig. 4. The density of the number of consecutive correct measurements (middle), and of the number of consecutive missing measurements (right) for one node.

Complexity: The decision problem can be mapped to a maximization problem using a binary search. A special case of the above problem is when each reliable clique and each substitute clique include only a single vertex. This instance of the problem corresponds to the domatic number problem and is one of the classical NP-complete problems [9].

We formulate the sleeping coordination problem as an instance of integer linear program (ILP). Even though the problem is NP-complete, for many practical instances, we are able to find the solutions in very short run time (less than 1 minute). For ILP formulation, we first introduce the constants and variables. After that, we formulate the objective function and constraints.

Given: A number $K \leq (\delta + 1)$, R reliable cliques A_r, $r = 1, \ldots, R$, S substitute cliques B_s, $s = 1, \ldots, S$, and a prediction matrix $P_{\{}N \times N\}$ with elements p_{ij}, s.t.

$$p_{ij} = \begin{cases} 1, & \text{If } |\epsilon(\hat{v_j} = f(v_i))| \leq |\varepsilon| \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

Where $|\epsilon(\hat{v_j} = f(v_i))|$ is the error in predicting the value at sensor v_j given the data at v_i, and ε is the user's specified error tolerance and δ is the degree of the vertex with the minimum degree in the graph [10].

Variables: matrix $X_{\{}K \times N\}$ with elements x_{ik}, and a vector $U_{\{K\}}$ with elements u_k s.t. $u_k = 1$ is set S_k was selected, and 0 otherwise, and:

$$x_{ik} = \begin{cases} 1, & \text{If node } v_i \text{ is in set } S_k \\ 0, & \text{Otherwise} \end{cases} \quad (2)$$

Objective Function: The objective function is to maximize the number of disjoint dominating sets, i.e., $max \sum_k u_k$.

Constraints: The problem has five set of constraints. The first set of constraints (C_1) ensures that if a set S_k exist (i.e. $u_k = 1$), all nodes in G that are not in S_k have an incoming edge from a node in S_k. For $i = 1, ..., N$, $k = 1, ..., K$: $x_{ik} + \sum_j P_{ij} x_{jk} \geq u_k$.

The second set of constraints (C_2) is that if a node is selected in one group, it cannot be selected for any other group. For $i = 1, ..., N$, and $k = 1, ..., K$: $\sum_k x_{ik} \geq 1$.

The third set of constraints (C_3) ensures that for each vertex v_i is a domatic partition, there is less than $\rho\%$. This constraint corresponds to having at least one of the reliable cliques containing v_i within the domatic partition. To write this constraint, we define two auxiliary functions F_{OR} and F_{AND} on L variables as follows: $F_{OR}(D_1, ..., D_L) = D_1 \vee D_2 \cdots \vee D_L$ and $F_{AND}(D_1, ..., D_L) = D_1 \wedge D_2 \cdots \wedge D_L$.

The functions F_{OR} translates to the following linear constraints: (i) $F_{OR}(D_1, ..., D_L) \geq D_l$, for $l = 1, \ldots, L$, (ii) $F_{OR}(D_1, ..., D_L) \leq D_1 + D_2 \cdots + D_L$ and (iii) $0 \leq F_{OR}(D_1, ..., D_L) \leq 1$. The function F_{AND} translates to the following linear constraints: (i) $F_{AND}(D_1, ..., D_L) \leq D_l$, for $l = 1, \ldots, L$, (ii) $L-1+ F_{AND}(D_1, ..., D_L) \geq D_1+D_2 \cdots + D_L$, and (iii) $0 \leq F_{AND}(D_1, ..., D_L) \leq 1$.

Constraint $C3$ states that if a node v_i is in the group S_k, then there is at least one reliable clique A_r with $a_{ri} = 1$, such that $A_r \subset S_k$. If $A_r \subset S_k$, then the expression $C3_r$: $\sum_{i=1}^{N} a_{ri}x_{ik} = |A_r|$ would hold. Since at least one reliable clique should hold for each node in a set, we have the following constraints for each S_k, $k = 1, \ldots, K$.

$$\begin{aligned} x_{1k} = &\quad (a_{11} \wedge C3_1) \vee \cdots \vee (a_{R1} \wedge C3_R) \\ \vdots &\quad \vdots \\ x_{Nk} = &\quad (a_{1N} \wedge C3_1) \vee \cdots \vee (a_{RN} \wedge C3_R) \end{aligned}$$

The fourth set of constraints (C_4) ensures that for each v_i not in a domatic partition, there is a substitute group such that the combination of substitute nodes has less than $\rho\%$. This constraint corresponds to having at least one of the substitute cliques corresponding to $v_i \notin S_k$ within each domatic partition S_k. Constraint $C4$ states that if a node v_i is not in the group S_k, then there is at least one substitute clique B_s with $d_{si} = 0$, such that $B_s \subset S_k$. If $B_s \subset S_k$, then the expression $C4_s$: $\sum_{i=1}^{N} b_{si}x_{ik} = |B_s|$ would hold. For $k = 1, \ldots, K$:

$$\begin{aligned} x_{1k} = &\quad (d_{11} \wedge C4_1) \vee \cdots \vee (d_{S1} \wedge C4_S) \\ \vdots &\quad \vdots \\ x_{Nk} = &\quad (d_{1N} \wedge C4_1) \vee \cdots \vee (d_{SN} \wedge C4_S) \end{aligned}$$

The last set of constraints (C_5) ensures that the variables u_k and x_{ik} are within the [0,1] range. For $i = 1, ..., N$, $k = 1, ..., K$, $0 \leq x_{ik} \leq 1$, and $0 \leq u_k \leq 1$. Note that, we extract the P matrices and $K = (\delta + 1)$ from our modeling studies.

To evaluate the effectiveness of the new approach we compared the new sleeping coordination technique with the base case of sleeping strategy that uses the same intersensor models, but does not consider missing data models. The comparison was done by enforcing that the lifetimes of the networks for both the base case and new approach are identical. For each case, we calculate the percentage of missing data. Table 1 shows the results. The first two columns show the number of nodes in the experiment and the maximal allowed error. The next two columns show the percentage of missing data for temperature sensors when base case and new approach are used respectively. The last two columns show the same data for percentage of humidity missing. All experiments with 54 or less nodes are conducted on actual data traces. The large instances use the interacting particle model. While the base case coordination was never able to recover more than two third of data, the new approach consistently recovered more than 92%.

# of nodes	err rate (%)	Temp B Rec(%)	Temp N Rec(%)	Hum B Rec(%)	Hum N Rec(%)
27	2	40.6	7.1	35.6	4.9
	3	41.3	7.7	35.3	4.9
40	2	39.8	6.8	35.0	5.1
	3	39.5	6.6	32.7	6.7
54	2	41.3	6.6	35.2	4.9
	3	43.4	6.1	33.5	5.8
100	2	40.1	8.0	35.9	6.3
	3	39.7	6.9	36.4	6.5
200	2	41.0	5.9	40.8	3.1
	3	41.3	5.5	43.9	7.3

Table 1. Percentage of missing data for the sleeping coordination approach without the missing data recovery (B) and for the sleeping coordination with the missing data recovery (N). The results are shown for temperature (Temp) and humidity sensors (Hum).

4. CONCLUSION

We have developed an approach for energy efficient energy management using sleeping in sensor networks in presence of missing data. We introduced interacting particle-based model and a simulator for missing data. Using combination of non-parametric statistical modeling and ILP formulation, we optimally addressed the problem and demonstrated significant improvements in ensuring completeness of collected data.

5. REFERENCES

[1] A. Cerpa, J.L Wong, L. Kuang, M. Potkonjak, and D. Estrin, "Statistical model of lossy links in wireless sensor networks," in *IPSN*, 2005, pp. 81–88.

[2] V. Rajendran, K. Obraczka, and J. J. Garcia-Luna-Aceves, "Energy-efficient collision-free medium access control for wireless sensor networks," in *Sensys*, 2003, pp. 181–192.

[3] C. Han, R. Kumar, R. Shea, E. Kohler, and M. Srivastava, "A dynamic operating system for sensor nodes," in *MobiSys*, 2005, pp. 163–176.

[4] F. Koushanfar, N. Taft, and M. Potkonjak, "Sleeping coordination for comprehensive sensing: Isotonic regression and domatic partitions," Tech. Rep., Intel Research, 2005.

[5] T. M. Liggett, *Interacting particle systems*, Springer-Verlag, 1985.

[6] R. Durrett, *Lecture notes on particle systems and percolation*, Wordworth & BrooksCole, 1988.

[7] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements Of Statistical Learning: Data Mining, Inference, And Prediction*, Springer, New York, 2001.

[8] V. Raghunathan, C. Schurgers, S. Park, and M.B. Srivastava, "Energy-aware wireless microsensor networks," in *IEEE Signal Processing Magazine*, 2002, vol. 19, pp. 40–50.

[9] M. R. Garey and D. S. Johnson, *Computers and intractability. A Guide to the theory of NP-completeness*, W. H. Freeman and Company, 1979.

[10] U. Feige, M. M. Halldorsson, G. Kortsarz, and A. Srinivasan, "Approximating the domatic number," *SIAM Journal of Computing*, vol. 32, no. 1, pp. 172–195, 2002.