Opportunistic Sampling in Wireless Sensor Networks

Muhammad Umer Egemen Tanin Lars Kulik

National ICT Australia
Department of Computer Science and Software Engineering
University of Melbourne, Victoria 3010, Australia
{mumer,egemen,lars}@csse.unimelb.edu.au

ABSTRACT

In an active WSN where user queries are regularly processed, a significant proportion of nodes relay and overhear data generated by other nodes in the network. In this paper, we propose to exploit this mode of data communication towards a gradual buildup of global knowledge. We show that by harnessing the multihop and multipath communication advantages, only a few user queries in a WSN can lead to an accumulation of accurate global knowledge at node level. This global knowledge can greatly improve numerous WSN applications when used in data validation, event detection, and query optimization.

1. INTRODUCTION

Localized information processing is a hallmark of wireless sensor network (WSN) technology. It attempts to minimize the frequency and size of data transmission among distant nodes by preprocessing and data sharing among neighboring nodes. One of the foremost challenges in localized information processing in WSNs is the need for global knowledge at node level. Consider, for instance, the requirement in TinyDB [5] to maintain `MIN` and `MAX` aggregates at each node for efficient query forwarding. Similarly, localization of essential WSN operations such as event discovery, range query optimization and data validation require global statistical knowledge at the node level [8].

In this paper, we propose *opportunistic sampling*; a novel sampling method for accumulating global knowledge at node level. Opportunistic sampling is based on the key insight that an operational WSN, actively being queried by its users, already possesses a certain degree of data distributed in the network. Since WSN communication is multihop and multipath (broadcast) in nature, data from a source to a sink node is relayed and overheard by several intermediate nodes. Due to this multihop, multipath (M2) communication advantage, each user query may leave a trace of collected data among the communicating nodes. We propose that nodes opportunistically sample this data or piggy-back augmented information as they respond to user queries. The data thus collected can then be used to answer future global aggregation requirements at node level without further communication. Our experimentation shows that if the M2 advantage is properly harnessed, nodes accumulate surprisingly accurate global knowledge after responding to only a small number of user queries.

ACM GIS '09 , November 4-6, 2009. Seattle, WA, USA

1.1 Motivating Example

Figure 1 presents a simple example to motivate the case for opportunistic sampling based on the M2 advantage. The figure shows a WSN deployed as a grid where each node can communicate with its all one hop neighbors. Assume that in each case the sink node issues the following query:

```
SELECT Humidity
FROM Sensors
SAMPLE PERIOD 30 min
FOR 4 hrs
```

Figure 1 shows that in response to the above query, data collection is performed using a multihop routing tree rooted on the sink node. We propose that in addition to propagation of incoming data towards higher tree levels, each node samples and saves this data for global statistics computation and future queries. The advantage of this opportunistic sampling is magnified by the M2 advantage of underlying routing structure. Due to the synchronization of listen periods among all nodes in the routing tree, internal nodes can receive and sample the incoming data from a region much larger than their individual transmission range.

Figure 1(a) highlights the region from which an arbitrary node (node B in the figure) in the network overhears or receives humidity data. We refer to this area as the M2 coverage of node B. It can be seen from Figures 1(b), 1(c) and, 1(d) that different locations of the sink node results in a different level of M2 coverage of node B. For node B, the presence of sinks at the four shown locations leads to the accumulation of humidity information from the entire network. Hence, in this example, node B can report an accurate global statistic, such as `MIN`, `MAX` or `AVG`, without issuing a single query itself.

The above example shows the potential strength of M2 coverage based opportunistic sampling in a simple setting. Computation of the extent of M2 coverage for a given node can be viewed as a spatial modeling problem involving WSN routing structure and locations of the given node and sinks. In this paper, we focus on the modeling of M2 coverage available in a typical WSN and use this model to show that

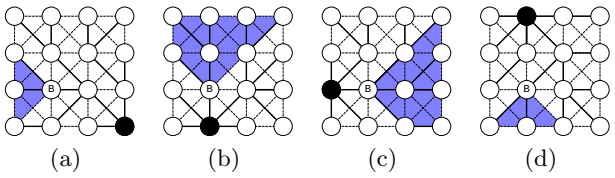


Figure 1: A motivating example for opportunistic sampling based on the M2 advantage. The dark circles represent the sinks. Solid edges represent the flow of data towards the sinks while dashed edges between two nodes show that the nodes can overhear each other.

only a small number of user queries lead to very high levels of M2 coverage in a network. We also use this coverage model to devise a localized M2 coverage estimation technique for individual nodes.

The accuracy of global statistics collected by opportunistic sampling may be affected by rapid temporal variation in a sensed phenomenon. If the query arrival rate is low compared to the temporal variation in the phenomenon, nodes will have fewer opportunities to sample information. The sampled information may thus become stale by the time a global statistic is required. However, the temporal variation of a phenomenon can be determined at an application level using past data. Nodes can thus measure the suitability of sampled information in terms of its freshness and may either proceed with using the information for computing a global statistic or may issue a new query.

The potential of opportunistic sensing may be affected by the nature of user queries. If queries involve aggregates, for instance seeking average humidity instead of all humidity values as in the above example, raw sensor readings are suppressed by in-network aggregation during transmission. As a result, the opportunistic information sampled at each node may only suit a certain type of global statistic. In such cases, we propose that each node piggy-backs some augmented information along with the partial query result it transmits. For example, if the global statistic `MAX` is required, each node can piggy-back augmented information in the form of partial `MAX` aggregate based on incoming data and its own humidity reading. Regardless of the type of query being processed, if all nodes add to and sample augmented information, a large proportion of nodes may accumulate enough local information to accurately compute the required statistic.

An alternative of the above piggy-backing strategy could be to exploit the partial query results computed by each node. A number of techniques for sharing of partial query results in future queries have been proposed in the literature (e.g., [3]) that can be adopted in our opportunistic sampling system. We leave the investigation of effectiveness of partial query results in global statistics computation as a future work.

Motivated by the need of global statistics for localized and distributed information processing, a number of methods for collecting and distributing such statistics have been proposed in the WSN literature. DIMENSIONS [2] and other distributed data indexing systems [7] aim to provide fast in-network search of pre-computed aggregates. Such systems propose an in-network storage structure such that WSN nodes possess global aggregates of multiple granularity levels. As an orthogonal approach, the *spatial gossiping* method aims to build and distribute a uniform level aggregate throughout the entire network. In [8], Sarkar et al. propose the hierarchical spatial gossip (HSG) algorithm that combines in-network storage and gossiping approaches to accumulate multi-resolution aggregates at each node in the network. A common thread among current global statistics collection techniques is their specialized nature, i.e., these approaches are purpose-built for a certain type of global statistic and have to be triggered regularly, which incurs significant costs.

The multipath nature of WSN communication has been exploited before for a variety of goals including fault tolerant in-network data aggregation [6] and spatial suppression in data collection [4]. To the best of our knowledge, our work is the first to exploit the coverage properties of multihop routing for data collection.

2. COVERAGE MODEL

We assume a network of N nodes uniformly deployed inside a $d \times d$ square. Two nodes can communicate with each other if the distance between their locations is less than R, the transmission radius. Query forwarding and data collection is performed using a TinyDB-style random data collection tree [4].

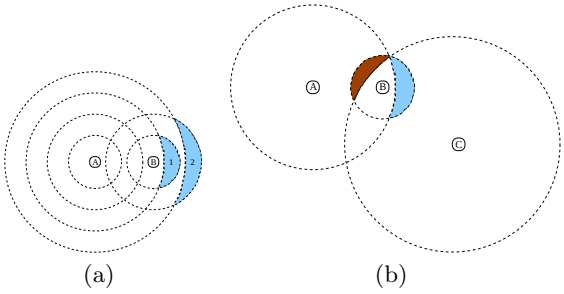


Figure 2: Estimating the M2 coverage.

M2 Coverage.

The M2 advantage virtually extends the sensing coverage of each node much beyond its communication radius. Figure 2(a) shows an example where a query, $Q1$, is issued by a node A. All nodes in the network synchronize their transmission periods depending on their level in the query tree formed by successive rebroadcasts of $Q1$. An arbitrary node, represented as node B in this example, receives $Q1$ during the third broadcast round and sets its listen and transmit periods accordingly. Region 1 in Figure 2(a) represents node B's first level M2 coverage, i.e., the area within direct transmission range of node B, from where it expects to receive (or overhear) data once the network starts responding to $Q1$. This coverage region is formed by the part of node B's transmission range that does not intersect with the query wavefront that delivers $Q1$ to node B. All nodes falling inside the *overlapping* part of node B's transmission range and the query wavefront must have received $Q1$ before node B and must have set their transmission time periods accordingly. Thus, during its listen time-period, node B cannot receive or overhear any data from these nodes.

Node B's M2 coverage increases as nodes in its first level

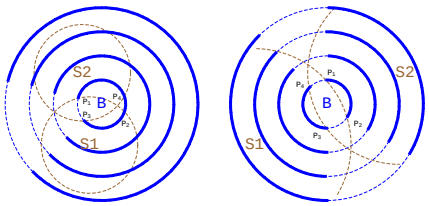
M2 coverage rebroadcast $Q1$ further into the network. Analogous to the case above, the second level M2 coverage for node B will include the part of node B's second level broadcast that does not overlap with the corresponding wavefront for $Q1$. Depicted as region 2 in Figure 2(a), node B's second level M2 coverage is greater in area than its first level M2 coverage. Further re-broadcasts of $Q1$ increases node B's M2 coverage in a similar manner.

Multi-sink M2 Coverage.

Figure 2(b) extends the above example by including a second query, $Q2$, issued by node C. Node B can now overhear or receive data from all nodes in the non-overlapping part of its transmission range and the query wavefront that delivers it query $Q2$. Consequently, the M2 coverage for node B is further increased. The overall M2 coverage for node B can now be computed by estimating the complement of joint intersection of node B's transmission range with wavefronts $Q1$ and $Q2$. In general for K sinks, the exact M2 coverage computation requires the area of joint intersection of $K+1$ circles. Unfortunately, a model cannot be established based on this computation as the close form representation of the area defined by the intersection of $K+1$ unequal circles in general positions is not known [1]. We identify the joint intersection of unequal circles as an interesting future problem, however, to make the M2 coverage computation tractable we propose an approximation technique in the section below.

3. DISCRETE COVERAGE APPROXIMATION

Figure 3(a) presents an example to explain the main idea behind discrete M2 coverage approximation. In this example, a node B receives queries from sink nodes $S1$ and $S2$. The two query wavefronts intersect node B's transmission range at points P_1, P_2 and P_3, P_4, respectively (Figure 3(a)). Considering sink $S1$ alone, node B's first level M2 coverage can be estimated by the number of border nodes present on the arc represented by the tuple (P_1, P_2). Accounting for the M2 coverage extended through the border nodes automatically includes that offered by any inner nodes.



(a) A node with one uncovered arc. (b) A node with two uncovered arcs.

Figure 3: Discrete approximation for the M2 coverage.

Assuming a uniform network density, the number of nodes can be approximated by the length of the arc (P_1, P_2). Introduction of a second sink, $S2$, increases the first level M2 coverage for node B by the addition of border nodes located on the arc (P_4, P_3). As a result, the only uncovered part on node B's boundary remains to be the arc (P_3, P_1), referred to here as node B's *uncovered arc*.

Uncovered arcs occur as segments on the transmission border of a node i that fall inside the area of overlap of a set of wavefronts intersecting node i's transmission range. Given M_i such arcs, the first level M2 coverage for node i, $\widehat{\lambda_{i1}}$, can then be approximated as follows:

$$\widehat{\lambda_{i1}} = 2\pi R - \sum_{m=1}^{M_i} R\theta_m - \epsilon_{i1} \quad (1)$$

where R is the fixed radio transmission range, θ_m is the angle (in radians) of arc m and, ϵ_{i1} is an error constant. Since the M2 coverage approximation relies on a circular representation of query broadcast centered at a given node, certain parts of the broadcasted region may lie outside the deployment area. The error constant, ϵ_{i1}, is used to subtract the area of such parts from the M2 coverage.

Computation of the overall M2 coverage for a given node is greatly simplified by the above boundary nodes' based approximation. As shown in Figure 3, all first level uncovered arcs for a node i grow linearly with a factor R (the transmission radius) during node i's subsequent level broadcasts. The second level M2 coverage of node B is a sum of first level M2 coverage of all nodes located on the covered parts of its transmission border. In general, the node coverage for node i for level j, can be computed as:

$$\widehat{\lambda_{ij}} = 2\pi Rj - \sum_{m=1}^{M_i} Rj\theta_m - \epsilon_{ij} \quad (2)$$

The overall M2 coverage for node i can then be computed as:

$$\widehat{\Lambda_i} = \sum_{j=1}^{D} \widehat{\lambda_{ij}} \quad (3)$$

where D is a constant large enough to guarantee that broadcast reaches the entire deployment area.

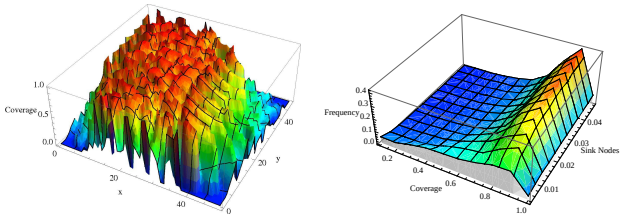
Given the locations of a set of sink nodes, the above approximation model allows a WSN node to locally determine the extent of its M2 coverage. This computation forms the basis of the opportunistic sampling system. Based on this knowledge nodes can decide to respond to requests for global statistics using local sampled information or inject new queries in the network.

4. SIMULATIONS

This section presents a simulation based analysis of opportunistic sampling technique. We simulate a network of 2500 nodes, set in a grid of 1 meter resolution and study the coverage, cost and accuracy of our proposed technique.

4.1 M2 Coverage

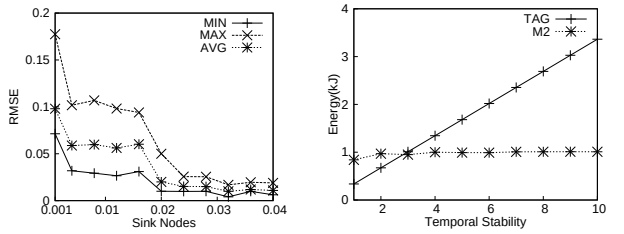
We show that only a small number of user queries (or sink nodes) lead to a high overall M2 coverage in a network. Figure 4(a) shows the approximate M2 coverage of each node (as a fraction of network area) due to 10 randomly placed sinks. This experiment shows that due to only a small number of sinks (0.4% of entire network), a large proportion of nodes (42%) achieve near exhaustive (> 95%) M2 coverage. To analyze this result thoroughly, in Figure 4(b) we report the M2 coverage histogram as the number of sink nodes is increased up to 5% of the network. The frequency axis shows the number of nodes in a histogram bin while all values are normalized to the total number of nodes in the network. This experiment shows that for a number of sink nodes as low as



(a) Approximate M2 coverage with 10 randomly placed sinks

(b) M2 coverage histogram

Figure 4: Approximate M2 coverage



(a) Effect of the number of sink nodes on estimation accuracy. Number of sink nodes normalized to the total number of nodes in the network.

(b) Effect of temporal stability on communication cost. Temporal stability is presented as the number of estimations required per unit of time.

Figure 5: Opportunistic sampling for global statistics collection.

2% of the entire network, 64% of the nodes cover more than 80% of the deployment area, while 87% of the nodes cover more than half of the deployment area.

4.2 Global Statistics Collection

This section reports experimentation with opportunistic sampling targeted at global statistics computation at node level. We assume that users issue aggregate queries which are processed using in-network aggregation. During query processing nodes continuously collect, sample and transmit augmented information in the form of `MIN`, `MAX` and `AVG` aggregates of incoming data. In response to any request for a global statistic by a user or an application (e.g., an event detection application), a node first attempts to serve this request locally, using its sampled information. It broadcasts the request in the network only if its sampled information is not sufficient to compute the answer.

The chart in Figure 5(a) shows the effect of increasing number of sink nodes on estimation accuracy of `MIN`, `MAX` and `AVG` statistics. In this chart, the sink node values are shown as proportion of the network size, while accuracy is represented as normalized root mean squared error (RMSE). The chart shows that an increase in the number of active sink nodes increases the mean M2 coverage for all nodes leading to a better estimation. A relatively small number of active sink nodes (2%) is enough to guarantee a mean error less than 2% in `MIN` and `AVG` and less than 5% in `MAX` aggregates. This experiment assumes a balance between query arrival rate and temporal stability of the observed phenomenon. Therefore, the error in global statistics can be solely attributed to lack of coverage. We leave the investigation of errors introduced by stale information as a future work.

The chart in Figure 5(b) shows the cost of computing global statistics by opportunistic sampling (denoted as M2) in comparison to a proactive data aggregation technique using Tiny AGgregation(TAG) [4]. In our implementation of TAG, a designated base station collects aggregate data from the entire network employing in-network aggregation. It then broadcasts the final aggregates to all nodes in the network. A single run of this technique is quite efficient, however, its efficiency degrades quickly if the temporal stability of the observed phenomenon is low. Temporal stability refers to the ratio between the length of time for which a value is expected to remain valid to the total monitoring duration. For instance, in a WSN monitoring temperature in a building where we expect the average temperature to be changing one degree per hour, the temporal stability of MIN, MAX and AVG aggregates for a 12 hour monitoring cycle will be 1, resulting in re-computation of aggregates 12 times. If in the same network we expect 2% of nodes to be actively collecting data every hour, aggregate computation can be performed opportunistically with no extra cost. Figure 5(b) shows this trend for our simulated network.

5. CONCLUSIONS

We propose to exploit the multi-hop and multi-path (M2) advantage of the WSN communication paradigm. Based on the M2 advantage, we present an opportunistic sampling approach where WSN nodes gather global statistics by sampling incoming data during regular data collection. We model the M2 advantage for multi-sink WSNs and show that only a relatively small number of queries are enough to guarantee accurate global statistics. In future, we plan to extend this work for a larger range of global aggregates while building a realistic prototype for query processing. Moreover, we plan to study the impact of various query types and temporal variations of an observed phenomenon on the cost and accuracy of the opportunistic sampling method.

6. REFERENCES

[1] M. P. Fewell. Area of common overlap of three circles. Technical Note DSTO-TN-0722, Defence Science and Technology Organization, Australia, 2006.

[2] D. Ganesan, B. Greenstein, D. Estrin, J. Heidemann, and R. Govindan. Multiresolution storage and search in sensor networks. *Trans. Storage*, 1(3):277–315, 2005.

[3] R. Huebsch, M. Garofalakis, J. M. Hellerstein, and I. Stoica. Sharing aggregate computation for distributed queries. In *Proceedings of SIGMOD*, pages 485–496, 2007.

[4] S. Madden, M. J. Franklin, J. M. Hellerstein, and W. Hong. TAG: A Tiny AGgregation service for ad-hoc sensor networks. *SIGOPS Oper. Syst. Rev.*, 36(SI):131–146, 2002.

[5] S. Madden, M. J. Franklin, J. M. Hellerstein, and W. Hong. TinyDB: An acquisitional query processing system for sensor networks. *ACM Trans. Database Syst.*, 30(1):122–173, 2005.

[6] S. Nath, P. B. Gibbons, S. Seshan, and Z. R. Anderson. Synopsis diffusion for robust aggregation in sensor networks. In *In Proceedings of SenSys*, pages 250–262, 2004.

[7] S. Ratnasamy, B. Karp, S. Shenker, D. Estrin, R. Govindan, L. Yin, and F. Yu. Data-centric storage in sensornets with GHT, a geographic hash table. *Mob. Netw. Appl.*, 8(4):427–442, 2003.

[8] R. Sarkar, X. Zhu, and J. Gao. Hierarchical spatial gossip for multi-resolution representations in sensor networks. In *Proceedings of IPSN*, pages 420–429, 2007.