# Approximate Holistic Aggregation in Wireless Sensor Networks

Ji Li
Department of Computer Science
Georgia State University
Atlanta, Georgia 30302
Email: jli30@student.gsu.edu

Siyao Cheng
School of Computer Science
Harbin Institute of Technology
Harbin, China, 150001
Email: cheng@hit.edu.cn

Yingshu Li Zhipeng Cai
Department of Computer Science
Georgia State University
Atlanta, Georgia 30302
Email: {yili, zcai}@gsu.edu

***Abstract*—Holistic aggregation results are important for users to obtain summary information from Wireless Sensor Networks (WSNs). Holistic aggregation requires all the sensory data to be sent to the sink, which costs a huge amount of energy. Fortunately, in most applications, approximate results are acceptable. We study the approximated holistic aggregation algorithms based on uniform sampling. In this paper, four holistic aggregation operations are investigated. The mathematical methods to construct their estimators and determine the optional sample size are proposed, and the correctness of these methods is proved. Four corresponding distributed holistic algorithms are presented. The theoretical analysis and simulation results show that the algorithms have high performance.**

## I. Introduction

The amount of sensory data generated by a Wireless Sensor Network (WSN) is always huge and redundant. The summary information returned by aggregation operations is thus more meaningful. Aggregation operations can be classified as distributable, algebraic and holistic ons. The information returned by holistic aggregations is the most valuable but all sensory data needs to be transmitted for the exact results, which costs high energy consumption. Considering approximate aggregation results are acceptable [1], [2], [3], some approximate holistic aggregation techniques were proposed, such as [4], which save lots of energy but have a fixed error bound and cannot satisfy an arbitrary precision requirement.

We take *frequency*, *distinct-count*, *value-to-rank* and *rank-to-value* to investigate the sampling based holistic aggregation algorithms in this paper. Our aim is to return an $(\epsilon, \delta)$-approximate holistic aggregation result. Although some $(\epsilon, \delta)$-approximate aggregation algorithms were proposed in [5], they can only deal with distributable and algebraic aggregations. The contributions of this paper are as follows: 1) The mathematical estimators for approximate aggregation operations are provided; 2) The methods to determine the optional sample size are designed, and their correctness is guaranteed; and 3) The distributed algorithms are provided, and the energy costs are analyzed. The simulation results are also presented.

## II. Problem Definition

A WSN with $n$ nodes is divided into $k$ disjoint clusters $C_1, C_2, \ldots, C_k$. Let $s_{ti}$ be the sensory value of node $i$ at time $t$. $S_t = \{s_{t1}, s_{t2}, \ldots, s_{tn}\}$ is the set of the sensory data at time $t$. $Dis(S_t)$ is the distinct set of $S_t$. We study the *frequency*, *value-to-rank*, *distinct count* and *rank-to-value* aggregation operations. For any $x \in Dis(S_t)$, the exact frequency $F(S_t, x) = \frac{|\{s_{tj}|s_{tj}=x \bigwedge 1\le j\le n\}|}{n}$. The exact value-to-rank $VR(S_t, x) = \frac{|\{s_{tj}|s_{tj}\le x \bigwedge 1\le j\le n\}|}{n}$. The exact distinct-count $DC(S_t) = |Dis(S_t)|$. For any given $r \in [0, 1]$, the exact rank-to-value $RV(S_t, r) = argmin_{s_{ti}} \frac{|\{s_{tj}|s_{tj}\le s_{ti} \bigwedge 1\le j\le n\}|}{n} \ge r$. We study how to obtain an $(\epsilon, \delta)$-approximate result. The definition of the $(\epsilon, \delta)$-estimator proposed in [5] and the problem are given as follows.

*Definition 1 ($(\epsilon, \delta)$-estimator):* For any $\epsilon > 0$ and $\delta \in [0, 1]$, $\widehat{I}_t$ is the $(\epsilon, \delta)$-estimator of $I_t$ if $Pr(\frac{|I_t, \widehat{I}_t|}{I_t} \ge \epsilon) \le \delta$.

Input: the sensory data set $S_t$ of a WSN, aggregation operator $Agg \in \{Frequency, Rank-to-value, DistinctCount, Value-to-Rank\}$, $\epsilon > 0, \delta \in [0, 1]$, $x$ (only for *value-to-rank*), and $r$ (only for *rank-to-value*).

Output: The $(\epsilon, \delta)$-approximate aggregation result of $Agg$.

## III. Preliminaries

$U(m) = \{X_1, X_2, ..., X_m\}$ denotes a uniform sample of $S_t$ with sample size $m$. $X_i$ $(1 \le i \le m)$ is a random variable satisfying $\Pr(X_i = s_{tj}) = \frac{1}{n}$ for any $s_{tj} \in S_t$. $f_j = F(S_t, s_{tj}^{(d)})$, $\widehat{f}_j = \widehat{F(S_t, s_{tj}^{(d)})}$. The estimator of $F(S_t, x)$, $R(S_t, x)$, $DC(S_t)$ and $RV(S_t, r)$ are defined as follows, $\widehat{F(S_t, x)} = \frac{|\{X_i|X_i=x \bigwedge 1\le i\le m\}|}{m}$, $\widehat{R(S_t, x)} = \frac{|\{X_i|X_i\le x \bigwedge 1\le i\le m\}|}{m}$, $\widehat{DC(S_t)} = \sum_{s_{tv}^{(d)} \in U(m)} \frac{1}{Pr(s_{tv}^{(d)} \in U(m))}$ $\widehat{RV(S_t, r)} = argmin_{s_{ti}^{(d)}}\{\sum_j \widehat{f}_j \ge r \bigwedge s_{tj}^{(d)} \le s_{ti}^{(d)}\}$. Let $RV(S_t, r)$ and $\widehat{RV(S_t, r)}$ be the $k$-th and $\widehat{k}$-th smallest values in $Dis(S_t)$. The error between $RV(S_t, r)$ and $\widehat{RV(S_t, r)}$ is defined as $\frac{|\sum_{i=1}^{k-1} f_i - \sum_{i=1}^{\widehat{k}-1} f_i|}{\sum_{i=1}^{k-1} f_i}$, then we have

*Theorem 1:* $\widehat{F(S_t, x)}$ is an $(\epsilon, \delta)$-estimator of $F(S_t, x)$ if $m \ge \frac{\phi^2_{\delta/2}}{\epsilon^2}(\frac{n}{n_{min}} - 1)$. $\widehat{R(S_t, x)}$ is the $(\epsilon, \delta)$-estimator of $R(S_t, x)$ if $m \ge \frac{\phi^2_{\delta/2}}{\epsilon^2}(\frac{1}{\mu} - 1)$, where $\mu$ denotes the lower bound of $R(S_t, x)$. $\widehat{DC(S_t)}$ is an $(\epsilon, \delta)$-estimator of $DC(S_t)$ if $m \ge \frac{ln(n\epsilon^2) - ln(n\epsilon^2 + 4n_{max} ln(2/\delta))}{ln(1 - n_{min}/n)}$. $\Pr(\frac{|\sum_{i=1}^{k-1} f_i - \sum_{i=1}^{\widehat{k}-1} f_i|}{\sum_{i=1}^{k-1} f_i} \ge \epsilon) \le \delta$ if $m \ge (\frac{\phi_{\delta/2} n_{max} n}{2\epsilon n_{min}(nr - n_{max}) - 2n_{min} n_{max}})^2 (\frac{nr}{n_{min}} + 1)$.□

1063-6927/15 $31.00 © 2015 IEEE
DOI 10.1109/ICDCS.2015.86

## IV. $(\epsilon, \delta)$-Approximate Aggregation Algorithms

The following theorem provides a method to determine the sample size for distinct-count based on approximate frequency.

*Theorem 2:* $\widehat{DC(S_t)}'$ is an $(\epsilon, \delta)$-estimator of $DC(S_t)$ if $m \geq max(m_1, m_2)$. $m_1 = (\ln(n(\sqrt{\epsilon+1}-1)^2) - ln(n(\sqrt{\epsilon+1}-1)^2 + 4n_{max}ln(2/(1-\sqrt{1-\delta}))))/(ln(1-n_{min}/n))$, $m_2 = \frac{\phi_{(1-\sqrt{1-\delta})/2}}{\epsilon'^2}(\frac{n}{n_{min}}-1)$, where $\epsilon'$ satisfies $\frac{\epsilon'(1-(1-\epsilon')n_{min}/n)^{\lfloor \frac{m_1}{2} \rfloor}}{1-\epsilon'} = \sqrt{1+\epsilon}-1$.□

The algorithms are listed as follows. The algorithm for value-to-rank is similar with that for frequency and thus omitted. The communication and energy costs for frequency, value-to-rank, distinct-count and rank-to-value are $O(\frac{1}{\epsilon^2}ln\frac{1}{\delta})$, $O(\frac{1}{\epsilon^2}ln\frac{1}{\delta})$, $O(\frac{1}{\epsilon_2^2}ln\frac{1}{\delta})$ and $O(\frac{1}{\epsilon^2}ln\frac{1}{\delta})$.

**Algorithm 1** $(\epsilon, \delta)$-Approximate Frequency

**Input:** $\epsilon, \delta$
**Output:** $(\epsilon, \delta)$-approximate frequency

1: $m = min(\lceil \frac{\phi^2_{\delta/2}}{\epsilon^2}(\frac{n}{n_{min}}-1)\rceil, n)$;
2: generate $Y_i$ following $\Pr(Y_i = l) = \frac{n_l}{n}$, where $n_l$ is the number of the nodes in cluster $C_l$ $(1 \leq i \leq m, 1 \leq l \leq k)$;
3: $m_l = |\{Y_i \mid Y_i = l\}|$ $(1 \leq i \leq m, 1 \leq l \leq k)$, the sink sends $m_l$ to each cluster head;
4: sample and calculate the partial frequency in each cluster;
5: transfer and merge partial frequency in the spanning tree;
6: get frequency $Fre$ from the sink node;
7: **return** $Fre$;

**Algorithm 2** $(\epsilon, \delta)$-Approximate Distinct Count

**Input:** $\epsilon, \delta$
**Output:** $(\epsilon, \delta)$-approximate distinct count

1: $\epsilon_1 = \sqrt{1+\epsilon}-1$, $\delta_1 = 1-\sqrt{1-\delta}$;
2: $m_1 = \lceil \frac{ln(n\epsilon_1^2) - ln(n\epsilon_1^2 + 4n_{max}ln(2/\delta_1))}{ln(1-n_{min}/n)} \rceil$;
3: Solve equation $\frac{\epsilon_2(1-(1-\epsilon_2)\frac{n_{min}}{n})^{\lfloor m_1/2 \rfloor}}{1-\epsilon_2} = \epsilon_1$;
4: $m_2 = \lceil \frac{\phi^2_{\delta_1/2}}{\epsilon_2^2}(\frac{n}{n_{min}}-1)\rceil$;
5: $m = min(max(m_1, m_2), n)$;
6: get approximate frequency $Fre$ with sample size $m$;
7: **return** $\sum_{i=1}^{|Fre|} \frac{1}{1+(1-Fre.Count[i])^m}$;

## V. Simulation Results

The main simulation results are shown in Fig.1 and Fig.2.

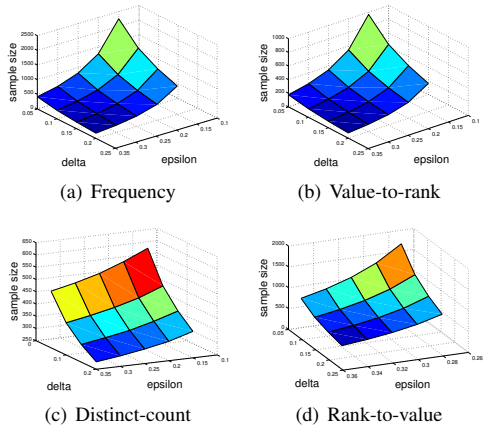
## VI. Conclusion

In this paper, the $(\epsilon, \delta)$-approximate algorithms for the frequency, value-to-rank, distinct-count and rank-to-value aggregation operations in WSNs are proposed. Furthermore, the sample size which can make the final result to satisfy the specified precision and failure probability requirements is derived. In addition, a cluster-based uniform sampling algorithm is provided. The simulation results show that the proposed algorithms have high performance.

**Algorithm 3** $(\epsilon, \delta)$-Approximate Rank-to-value

**Input:** $\epsilon$, $\delta$, rank $r$
**Output:** $(\epsilon, \delta)$-approximate value

1: $m_1 = \lceil (\frac{\phi_{\delta/2}n_{max}n}{2\epsilon n_{min}(nr-n_{max})-2n_{min}n_{max}})^2(\frac{nr}{n_{min}}+1)\rceil$;
2: $m_2 = \lceil (\frac{\phi_{\delta/2}n_{max}n}{2\epsilon n_{min}(n(1-r)-n_{max})-2n_{min}n_{max}})^2(\frac{n(1-r)}{n_{min}}+1)\rceil$;
3: $m = min(m_1, m_2, n)$;
4: get approximate frequency $Fre$ with sample size $m$;
5: **if** $m_1 < m_2$ **then**
6: find the min $p$ satisfying $\sum_{i=1}^{p} Fre.Count[i] \geq r$;
7: **else**
8: find the max $p$ satisfying $\sum_{i=p}^{|Fre|} Fre.Count[i] \geq 1-r$
9: **end if**
10: **return** $Fre.Value[p]$;




(a) Frequency (b) Value-to-rank

(c) Distinct-count (d) Rank-to-value

Figure 1. The relationship among $\epsilon$, $\delta$ and sample size.

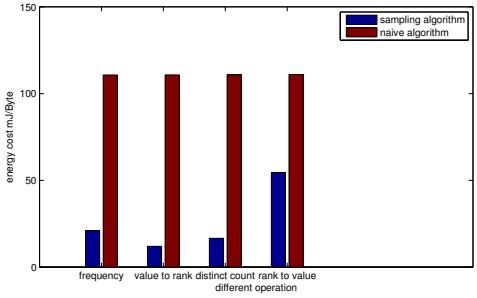



Figure 2. The energy cost comparison.

## References


[1] J. Considine, F. Li, G. Kollios, and J. Byers, "Approximate aggregation techniques for sensor databases," *ICDE*, pp. 449–460, 2004.
[2] Z. He, Z. Cai, S. Cheng, and X. Wang, "Approximate aggregation for tracking quantiles in wireless sensor networks," in *COCOA*, 2014.
[3] S. Cheng, J. Li, and Z. Cai, "$O(\epsilon)$-approximation to physical world by sensor networks," in *INFOCOM*, 2013.
[4] G. Cormode and M. Garofalakis, "Holistic aggregates in a networked world: Distributed tracking of approximate quantiles," in *In SIGMOD*, 2005, pp. 25–36.
[5] S. Y. Cheng and J. Z. Li, "Sample based $(\epsilon, \delta)$-approximate aggregation in sensor networks," *ICDCS*, pp. 273–280, 2009.