























**Figure 11: Number of messages in network per second (log-scale), while varying the number of sensors.**

ments, which is better than the performance of the algorithms on the synthetic datasets. This is due to the *smooth* nature of the data set, except for the measurements observed from October 28th to November 1st, where a major failure was detected in the systems and they reported deviating values. The results for the environmental 2-d dataset were comparable to those obtained with the synthetic 2-d dataset.

### 10.3 Memory and Communication Costs

In order to verify the efficiency of our technique, we ran experiments to measure the maximum amount of memory required by the *D3* algorithm per node. There are two components of our algorithm that affect the memory consumption: sample maintenance and variance estimation. The memory requirement of the former is upper-bounded by  $O(d|R|)$ , and of the latter by  $O(\frac{d}{\epsilon^2} \log|W|)$ .

We ran experiments using the real datasets, and assuming a 16-bit architecture, i.e., 2 bytes per number. We varied the size of the sliding window  $|W|$  (10000-20000), as well as the sample fraction  $f$  (0.25-5). The experiments showed that in all cases the actual values of the maximum memory consumption of the variance estimation procedure is around 55%-65% less than the theoretic upper bound.

We also ran experiments in order to quantify the number of messages that are generated, by scaling up the number of the nodes in our testbed. We compare our algorithms, *D3* and *MGDD* against the centralized approach. For our approach we take into account the number of messages generated due to the incremental sample propagation. We do not account for the messages sent when a local outlier is identified, since these are infrequent. We assume that each sensor generates one reading every 1 second. The size of the window  $|W|$  was set to 10240, the sample size  $|R|$  was set to 1024, and the sample fraction  $f$  was equal to 0.25. Figure 11 shows the number of messages generated per second (in log scale), while scaling up the number of nodes. As expected, the *D3* approach gives better savings compared to both *MGDD* and centralized. We observe that the *D3* algorithm requires approximately two orders of magnitude fewer messages, and hence the best method with respect to optimizing communication cost.

## 11. RELATED WORK

Madden and Franklin [31] present a framework for the efficient execution of queries in a sensor network. The problem of evaluating aggregate operators in a sensor network is addressed by Madden et al. [32]. Yao and Gehrke [45] investigate the problem of query processing in sensor networks. In a complementary study, Bonfils and Bonnet [9], describe an algorithm for mapping a tree of query operators on the sensor

network.

A recent study [16] proposes a sensor data acquisition technique, based on models that approximate the data with probabilistic confidences. This general technique results in reduced communication costs, without sacrificing much of the accuracy [15]. However, any special characteristics of the data distribution, such as periodic drifts, have to be explicitly encoded in the space of models considered. In our work, we describe a more general technique, which can efficiently overcome this limitation. Moreover, we observe all the data values, and can therefore reason about outliers, whereas the above technique aims at minimizing the cost of making *some* observations that will ensure the user-defined probabilistic confidence thresholds are met.

A framework for modeling sensor network data is also proposed by Guestrin et al. [20]. The goal in this approach is for the nodes in the network to collaborate in order to fit a global function to each of their local measurements. This is a parametric approximation technique, and as such, requires the user to make an assumption about the number of estimators required to fit the data. This model has more parameters to fit than the approach that we propose, where we only have to estimate a single parameter, thus reducing the requirements of in-network computation. Cormode and Garofalakis [13] describe a technique for approximate query tracking based on *sketches*. Their technique can efficiently operate in a distributed, online setting. Even though it can be generalized, it is mainly geared toward discrete domains and the unrestricted window model. In order to work for sliding windows, it would require to store all the values of the window, which is something we avoid doing in the framework we propose.

Greenwald and Khanna [19] study the problem of computing order statistics in a sensor network. Another recent study [41] addresses the problem of approximating the data distribution for computing order statistics, as well as range queries. There has also been work on predicting and caching the values generated by the sensors [35, 26], which can result in significant communication savings. Nevertheless, it is not obvious how to use this approach in our setting, since distance-based outliers require the computation of the number of neighboring values. It may be the case that values within the change detection threshold defined by the above approaches are outliers, and values outside this threshold are not. In addition, our model is designed to efficiently compute the distribution of a region, and therefore, identify outliers by combining the data from multiple sensors.

A similar approach for outlier detection in streaming data is described by Yamanishi et al. [44]. In contrast to our work, their method does not operate on sliding windows, but rather on the entire history of the data values, using an exponential forgetting factor for discounting the effect of the older values. Furthermore, the above approach is not geared towards a distributed environment, such as a sensor network.

There is extensive literature in the statistics community regarding outlier detection [6], as well as in the database community [3, 28, 38, 10]. However, none of these approaches is directly applicable to a sensor environment, either because they assume knowledge of the input data distribution, or because they are not tailored to operate online. There has been work on the special case of identifying outliers in streaming *time-series* data [37, 34]. Nevertheless, the significance of the temporal ordering is a major difference from the semantics of the problem we are considering in this study.

Recent work [22] gives an online technique to compute the JS divergence. This approach can be applied for some of the applications we consider, such as identifying faulty sensors but does not impact our algorithms for finding outliers.

## 12. CONCLUSIONS

In this paper, we study the problem of outlier detection in sensor networks. Outlier detection is very important in this context, since it enables the analyst to focus on the interesting events in the network. We propose a framework based on the approximation of the distribution of the sensor measurements. The techniques we describe operate efficiently in an online fashion. Moreover, they distribute the computation effort among the nodes in the network, thus better exploiting the available resources and cutting back on the communication and processing costs. We evaluated our approaches with a set of experiments with real and synthetic datasets. The experimental evaluation shows that our algorithm can achieve very high precision and recall rates for identifying outliers, and demonstrate the effectiveness of the proposed approach. As future work, we plan to evaluate our techniques in a real sensor network.

**Acknowledgments:** We would like to thank Samuel Madden for providing us the source code of the TAG simulator. The research of Vana Kalogeraki and Dimitrios Gunopulos is supported by NSF Grant 0330481.

## References

- [1] Crossbow Technology Inc. <http://www.xbow.com/>.
- [2] Earth Climate and Weather, University of Washington. <http://www-k12.atmos.washington.edu/k12/grayskies/>.
- [3] Andreas Arning, Rakesh Agrawal, and Prabhakar Raghavan. A Linear Method for Deviation Detection in Large Databases. In *KDD*, 1996.
- [4] Brian Babcock, Mayur Datar, and Rajeev Motwani. Sampling From a Moving Window Over Streaming Data. In *SODA*, 2002.
- [5] Brian Babcock, Mayur Datar, Rajeev Motwani, and Lidan O’Callaghan. Maintaining Variance And k-medians Over Data Stream Windows. In *PODS*, pages 234–243, USA, 2003.
- [6] V. Barnett and T. Lewis. *Outliers in Statistical Data*. John Wiley and Sons, Inc., 1994.
- [7] Shai Ben-David, Johannes Gehrke, and Daniel Kifer. Identifying Distribution Change in Data Streams. In *VLDB*, Toronto, ON, Canada, 2004.
- [8] Bjorn Blohsfeld, Dieter Korus, and Bernhard Seeger. A Comparison of Selectivity Estimators for Range Queries on Metric Attributes. In *SIGMOD*, 1999.
- [9] B. Bonfils and P. Bonnet. Adaptive and decentralized operator placement for in-network query processing. In *IPSN*, 2003.
- [10] M.M. Breunig, H.-P. Kriegel, R.T. Ng, and Jörg Sander. LOF: Identifying Density-Based Local Outliers. In *SIGMOD*, 2000.
- [11] Paul G. Brown and Peter J. Haas. Techniques for warehousing of sample data. In *ICDE*, 2006.
- [12] Kaushik Chakrabarti, Minos N. Garofalakis, Rajeev Rastogi, and Kyuseok Shim. Approximate Query Processing Using Wavelets. In *VLDB*, 2000.
- [13] Graham Cormode and Minos N. Garofalakis. Sketching streams through the net: Distributed approximate query tracking. In *VLDB*, pages 13–24, 2005.
- [14] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. John Wiley & sons, 1991.
- [15] Amol Deshpande, Carlos Guestrin, and Samuel R. Madden. Using Probabilistic Models for Data Management in Acquisitional Environments. In *Proc. CIDR*, 2005.
- [16] Amol Deshpande, Carlos Guestrin, Samuel R. Madden, Joseph M. Hellerstein, and Wei Hong. Model-Driven Data Acquisition in Sensor Networks. In *VLDB*, Toronto, ON, Canada, 2004.
- [17] D. Ganesan, B. Greenstein, D. Estrin, J. Heidemann, and R. Govindan. Multiresolution storage and search in sensor networks. *ACM TOS*, 1(3):27–315, 2005.
- [18] Anna C. Gilbert, Yannis Kotidis, S. Muthukrishnan, and Martin Strauss. Surfing Wavelets on Streams: One-Pass Summaries for Approximate Aggregate Queries. In *VLDB*, Rome, Italy, 2001.
- [19] M.B. Greenwald and S. Khanna. Power-Conserving Computation of Order-Statistics over Sensor Networks. In *PODS*, 2004.
- [20] Carlos Guestrin, Peter Bodik, Romain Thibaux, Mark Paskin, and Samuel Madden. Distributed Regression: an Efficient Framework for Modeling Sensor Network Data. In *IPSN*, Berkeley, CA, 2004.
- [21] Sudipto Guha and Nick Koudas. Approximating a Data Stream for Querying and Estimation: Algorithms and Performance Evaluation. In *ICDE*, pages 567–576, San Jose, CA, USA, 2002.
- [22] Sudipto Guha, Andrew McGregor, and Suresh Venkatasubramanian. Streaming and sublinear approximation of entropy and information distances. In *In Proceedings of the seventeenth annual ACM-SIAM symposium on Discrete algorithm*, 2006.
- [23] Dimitrios Gunopulos, George Kollios, Vassilis J. Tsotras, and Carlotta Domeniconi. Approximating Multi-Dimensional Aggregate Range Queries over Real Attributes. In *SIGMOD*, 2000.
- [24] Chalermek Intanagonwiwat, Deborah Estrin, Ramesh Govindan, and John Heidemann. Impact of network density on data aggregation in wireless sensor networks. In *ICDCS*, 2002.
- [25] H. V. Jagadish, Nick Koudas, S. Muthukrishnan, Viswanath Poosala, Kenneth C. Sevcik, and Torsten Suel. Optimal Histograms with Quality Guarantees. In *VLDB*, New York, NY, USA, 1998.
- [26] A. Jain, E.Y. Chang, and Y.-F. Wang. Adaptive Stream Resource Management Using Kalman Filters. In *SIGMOD*, 2004.
- [27] Ralph M. Kling. Intel Mote: An Enhanced Sensor Network Node. In *Workshop on Advanced Sensors, Structural Health Monitoring, and Smart Structures*, Kanagawa, Japan, 2003.
- [28] E.M. Knorr and R.T. Ng. Algorithms for Mining Distance-Based Outliers in Large Datasets. In *VLDB*, NY, NY, 1998.
- [29] Lillian Lee. On the effectiveness of the skew divergence for statistical language analysis. In *Artificial Intelligence and Statistics 2001*, pages 65–72, 2001.
- [30] J. Lin. Divergence measures based on the shannon entropy. *IEEE Trans. Theory*, 37:145–151, 1991.
- [31] Samuel Madden and Michael J. Franklin. Fjording the Stream: An Architecture for Queries Over Streaming Sensor Data. In *ICDE*, 2002.
- [32] Samuel Madden, Michael J. Franklin, and Joseph M. Hellerstein. TAG: A Tiny Aggregation Service for Ad-Hoc Sensor Networks. In *OSDI*, 2002.
- [33] N. Malpani, J. Welch, and N. Vaidya. Leader Election Algorithms for Mobile Ad Hoc Networks. In *DIAL M Workshop*, 2000.
- [34] S. Muthukrishnan, Rahul Shah, and Jeffrey Scott Vitter. Mining Deviants in Time Series Data Streams. In *SSDBM*, 2004.
- [35] C. Olston, J. Jiang, and J. Widom. Adaptive Filters for Continuous Queries over Distributed Data Streams. In *SIGMOD*, 2003.
- [36] S. Papadimitriou, H. Kitagawa, P. Gibbons, and C. Faloutsos. Loci: Fast outlier detection using the local correlation integral, 2003.
- [37] Vasundhara Puttagunta and Konstantinos Kalpakis. Adaptive Methods for Activity Monitoring of Streaming Data. In *ICMLA*, 2002.
- [38] Sridhar Ramaswamy, Rajeev Rastogi, and Kyuseok Shim. Efficient Algorithms for Mining Outliers from Large Data Sets. In *SIGMOD*, 2000.
- [39] Dongmei Ren, Baoying Wang, and William Perrizo. Rdf: A density-based outlier detection method using vertical data representation. In *ICDM*, pages 503–506, 2004.
- [40] D. Scott. *Multivariate Density Estimation: Theory, Practice and Visualization*. Wiley & Sons, 1992.
- [41] Nisheeth Shrivastava, Chiranjeeb Buragohain, Divyakant Agrawal, and Subhash Suri. Medians and Beyond: New Aggregation Techniques for Sensor Networks. In *ACM SenSys*, Baltimore, MD, USA, 2004.
- [42] N. Thaper, S. Guha, P. Indyk, and N. Koudas. Dynamic multidimensional histograms. In *SIGMOD Conference*, 2002.
- [43] B. Warneke, M. Last, B. Liebowitz, and K. Pister. Smart dust: Communicating with a cubic-millimeter computer. *IEEE Computer Magazine*, pages 44–51, January 2001.
- [44] Kenji Yamanishi, Jun ichi Takeuchi, Graham J. Williams, and Peter Milne. On-Line Unsupervised Outlier Detection Using Finite Mixtures with Discounting Learning Algorithms. *Data Mining and Knowledge Discovery*, 8(3):275–300, 2004.
- [45] Yong Yao and Johannes Gehrke. Query Processing for Sensor Networks. In *CIDR*, Asilomar, CA, USA, 2003.
- [46] Fan Ye, Haiyun Luo, Jerry Cheng, Songwu Lu, and Lixia Zhang. A Two-Tier Data Dissemination Model for Large-Scale Wireless Sensor Networks. In *MOBICOM*, Atlanta, GA, USA, 2002.
- [47] S. Zhao, K. Tepe, I. Seskar, and D. Raychaudhuri. Routing protocols for self-organizing hierarchical ad hoc wireless networks. In *IEEE Sarnoff Symposium*, 2003.