Poster: Data-Aware Edge Sampling for Aggregate Query Approximation

Joel Wolfrath
University of Minnesota
wolfr046@umn.edu

Abhishek Chandra
University of Minnesota
chandra@umn.edu

I. Motivation and Background

Data stream processing is an increasingly important topic due to the prevalence of smart devices and the demand for real-time analytics. One estimate suggests that we should expect nine smart-devices per person by the year 2025 [1]. These devices generate data which might include sensor readings from a smart home, event or system logs on a device, or video feeds from surveillance cameras. As the number of devices increases, the cost of streaming the device data to the cloud over the wide-area network (WAN) will also increase substantially. Transferring and querying this data efficiently has become the focus of much academic research [2]–[5]. Edge computation affords us the opportunity to address this problem by utilizing resources close to the devices. Edge resources have many different use cases, including minimizing end-to-end latency or maximizing throughput [6], [7]. We restrict our focus to minimizing the required WAN bandwidth, which is an effort to address the increase in data volume.

One strategy for reducing bandwidth consumption involves performing sampling at the edge [8]. This technique seeks to transfer a subset of the incoming streaming data without substantially impacting the results of user queries. Figure 1 shows an example of a system which performs sampling at the edge prior to forwarding data to the cloud. If these samples are an accurate representation of the original streams, then simply transferring the sample rather than the original data can save on the transfer cost. More sophisticated sampling techniques consider data-aware approaches to sampling. For example, Trihinas, Pallis, and Dikaiakos [9] propose a method for sampling a stream that considers the evolution of the stream over time. This involves continually estimating the variability in the stream and increasing the sampling rate if the stream becomes more variable. However, this approach assumes the user can exercise control over the sampling rate and does not consider dependencies that may exist between the devices themselves.

Another strategy for handling bandwidth constraints uses the cloud to simulate values from a stream. In this case, the user builds a machine learning model to simulate values from a stream when directly sampling the device is too expensive. Memon and Maheswaran [10] built a recurrent neural network model in the cloud which simulates values from a data stream based on historical data. Their approach is able to produce simulated data under a variety of conditions, but it does not attempt to detect and exploit dependencies between streams.

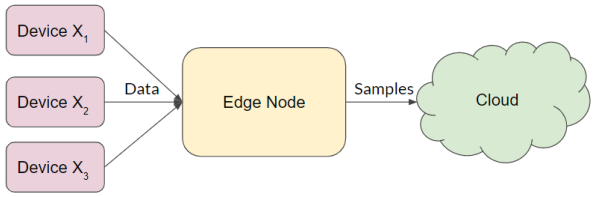

Fig. 1. **System Topology.** This example has 3 devices generating data with one edge node performing sampling prior to forwarding data to the cloud.

Both the data-aware sampling approach and the cloud simulation approach address WAN bandwidth constraints in different ways. This line of research considers a hybrid approach which seeks to systematically trade-off between edge sampling and simulation in the cloud. We restrict our focus to aggregate queries, which estimate quantities like counts, averages, standard deviations, and order statistics (e.g. minimum, maximum, or median).

II. Problem Statement and Proposed Solution

We assume there exists a set of k data-generating devices which produce key-value pairs over time. We also assume there exists an edge node which receives data from these devices or has the ability to query the devices directly. The edge node aggregates this data and is tasked with selecting a subset of the data to forward to the next node in the system, subject to some constraints. There are two main goals associated with this task:

1) Selecting a subset of data points that maximizes the amount of information gained about quantities of interest to the user.
2) Minimizing the amount of bandwidth (cost) required to convey this information to the next node in the system.

We formulate a more complex data-aware approach for sampling which attempts to exploit dependencies in the data. Recent research suggests that devices that are located in the same geographical regions may exhibit some kind of dependency [11], [12]. We seek to leverage these dependencies

to produce more accurate simulations from a stream when required. We examine two main sources of dependency:

1) **Auto-correlation**: Indicates the present values of a stream are dependent on its own past values.
2) **Cross-correlation**: Indicates the present values of a stream, Y are dependent on the current or past values of another stream X.

If we can accurately estimate the dependence in the data, we can allocate our samples more efficiently and simulate values from a stream more accurately. For example, if we know that two streams X and Y are sufficiently dependent, it may be optimal to only forward samples from stream X and let a model in the cloud simulate values from Y based on the observed values of X.

To that end, we formulate an information-theoretic optimization framework that runs at the edge and allocates samples given the dependence structure. In this case, we assume each stream follows a certain probability distribution, parameterized by a vector of parameters, θ_i. Our objective is to maximize the *Fisher Information*, which quantifies the amount of information about the distribution parameters present in the data [13]. Note that this information measure assumes a probability model for the stream; however, this assumption can be relaxed in practice. Let the number of real and simulated samples from a stream i be represented by $n_{i,r}$ and $n_{i,s}$ respectively. Then we denote the Fisher Information gained about the stream parameters by $\mathcal{I}(\theta_i,\ n_{i,r},\ n_{i,s})$. At a high level, this results in the following, simplified optimization problem:

$$\begin{aligned} \max_{n} \quad & \sum_{i=1}^{k} \mathcal{I}(\theta_i,\ n_{i,r},\ n_{i,s}) \\ \text{s.t.} \quad & \sum_{i=1}^{k} c_i(n_{i,r},\ n_{i,s}) \leq C \end{aligned} \tag{1}$$

where C is a bound on the overall cost and $c_i(n_{i,r}, n_{i,s})$ is a measure of the cost for transferring $n_{i,r}$ samples and simulating $n_{i,s}$ from stream i. This basic formulation is a convex optimization problem; however, the problem becomes non-convex if other constraints are enforced (depending on the application). In order to quantify the value of a simulated sample, we use the information theoretic concept of *mutual information*, which is able to capture arbitrary dependencies between two random variables. One drawback of this measure is that it requires knowledge of the pairwise joint distributions, which are not often realized in practice.

III. Empirical Results and Future Work

In order to evaluate the efficacy of our approach, we tested our framework against a real-world dataset with dependent streams. We used a UMASS trace containing temperature measurements from three homes located in western Massachusetts

This research was supported in part by the NSF under grant CNS-1717834.

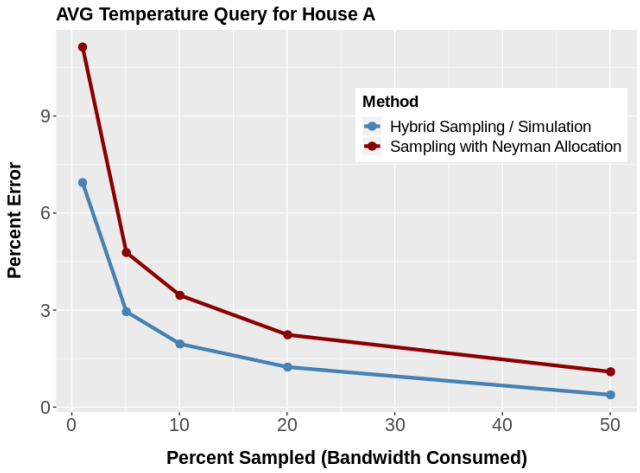

Fig. 2. **Bandwidth vs. Error Rate for the UMASS Temperature Trace captured at house A**.

[14] and performed our optimization to determine which samples should be sent and which should be simulated. We compared this approach against the *Neyman Allocation*, which simply allocates samples based on the estimated variability in each stream [15]. For each house, we measured the error associated with a query for the average temperature. Figure 2 compares the performance of the sampling methods across a variety of sample sizes. We used sample size as a proxy for bandwidth cost, since the increase in samples is proportional to the increase in WAN bandwidth required for the data transfer. Our hybrid method obtains a 1% error rate by sampling 30% percent of the observed values. The standard sampling approach required a sampling rate just over 50% to obtain the same average error. This suggests our combination of sampling and simulation has the potential to obtain comparable query accuracy with a 40% reduction in bandwidth requirements.

There are a few major challenges we are looking to address in our future work. First, our theoretical framework makes some strong assumptions, including a parametric model and knowledge of the pairwise joint distributions across the streams, which are difficult to estimate in practice. Another challenge is tolerating applications which perform outlier detection. If we simulate enough values with a model based on the expected value, we will be masking the variation present in the data. So we seek to establish criteria for bounding the number of simulated samples to prevent biasing these queries. Finally, we want to better understand the computational trade-off when performing sampling and optimization at the edge. Addressing these challenges will allow us to apply our framework to a wider range of applications.

References

[1] B. Safaei, A. M. H. Monazzah, M. B. Bafroei, and A. Ejlali, "Reliability side-effects in internet of things application layer protocols," in *2017 2nd International Conference on System Reliability and Safety (ICSRS)*, 2017, pp. 207–212.

[2] D. Kumar, J. Li, A. Chandra, and R. Sitaraman, "A ttl-based approach for data aggregation in geo-distributed streaming analytics," *Proc. ACM Meas. Anal. Comput. Syst.*, vol. 3, no. 2, Jun. 2019.

[3] B. Zhang, X. Jin, S. Ratnasamy, J. Wawrzynek, and E. A. Lee, "Awstream: Adaptive wide-area streaming analytics," in *Proceedings of the 2018 Conference of the ACM Special Interest Group on Data Communication*, ser. SIGCOMM '18, Budapest, Hungary, 2018, pp. 236–252.

[4] A. Jonathan, A. Chandra, and J. Weissman, "Multi-query optimization in wide-area streaming analytics," in *Proceedings of the ACM Symposium on Cloud Computing*, ser. SoCC '18, Carlsbad, CA, USA, 2018, pp. 412–425.

[5] A. Rabkin, M. Arye, S. Sen, V. Pai, and M. J. Freedman, "Making every bit count in wide-area analytics," ser. HotOS'13, Santa Ana Pueblo, New Mexcio, 2013.

[6] G. Amarasinghe, M. D. de Assunção, A. Harwood, and S. Karunasekera, "A data stream processing optimisation framework for edge computing applications," in *2018 IEEE 21st International Symposium on Real-Time Distributed Computing (ISORC)*, 2018, pp. 91–98.

[7] X. Fu, T. Ghaffar, J. C. Davis, and D. Lee, "Edgewise: A better stream processing engine for the edge," in *Proceedings of the 2019 USENIX Conference on Usenix Annual Technical Conference*, ser. USENIX ATC '19, Renton, WA, USA, 2019, pp. 929–945.

[8] Z. Wen, D. L. Quoc, P. Bhatotia, R. Chen, and M. Lee, "Approxiot: Approximate analytics for edge computing," in *2018 IEEE 38th International Conference on Distributed Computing Systems (ICDCS)*, 2018, pp. 411–421.

[9] D. Trihinas, G. Pallis, and M. D. Dikaiakos, "Adam: An adaptive monitoring framework for sampling and filtering on iot devices," in *2015 IEEE International Conference on Big Data (Big Data)*, 2015, pp. 717–726.

[10] S. Memon and M. Maheswaran, "Optimizing data transfers for bandwidth usage and end-to-end latency between fogs and cloud," in *2019 IEEE International Conference on Fog Computing (ICFC)*, 2019, pp. 107–114.

[11] J. Hribar and L. DaSilva, "Utilising correlated information to improve the sustainability of internet of things devices," in *2019 IEEE 5th World Forum on Internet of Things (WF-IoT)*, 2019, pp. 805–808.

[12] M. H. Mazhar and Z. Shafiq, *Characterizing smart home iot traffic in the wild*, 2020. arXiv: 2001.08288 [cs.NI].

[13] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. John Wiley and Sons, 2006.

[14] S. Barker, A. Mishra, D. Irwin, E. Cecchet, P. Shenoy, and J. Albrecht, "Smart*: An open data set and tools for enabling research in sustainable homes," ser. ACM Proc. SustKDD, 2012.

[15] J. Neyman, "On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection," *Journal of the Royal Statistical Society*, vol. 97, no. 4, pp. 558–625, 1934, ISSN: 09528385. [Online]. Available: http://www.jstor.org/stable/2342192.