Quality-aware Aggregation & Predictive Analytics at the Edge

Natascha Harth
School of Computing Science, University of Glasgow
n.harth.1@research.gla.ac.uk

Christos Anagnostopoulos
School of Computing Science, University of Glasgow
christos.anagnostopoulos@glasgow.ac.uk

***Abstract*—We investigate the quality of aggregation and predictive analytics in edge computing environments. Edge analytics require pushing processing and inference to the edge of a network of sensing & actuator nodes, which enables huge amount of contextual data to be processed in real time that would be prohibitively complex and costly to transfer on centralized locations. We propose a quality-aware, time-optimized edge analytics model that supports communication efficient predictive modeling within the edge network. Our idea rests on the capability of edge nodes to intelligently decide when and which data to deliver and process in light of minimizing the communication overhead and maximizing the quality of analytics results. We provide mathematical modeling, performance and comparative assessment over real datasets showing its benefits in edge computing environments.**

***Keywords*-Edge predictive analytics, quality of analytics, communication efficiency, optimal stopping theory.**

I. Introduction

Focusing on the increase of sensing & computing devices in Internet of Things (IoT) environments, delivering data continuously towards centralized locations e.g., Cloud, is constrained by bandwidth, energy, computational power and data storage. Aggregation & predictive analytics at the edge of an (IoT) network is an emerging area [1] trying to overcome these constrains by analyzing data *close* to the sources. Analytics at the edge over dynamic data is *different* from big data analytics over data at rest. It means carrying out the same kind of analysis, but moving more of it to the edge of the network, e.g., a car, an agricultural equipment in the field, or any other industrial device and exploiting only the local available resources. *Pushing* as much computing workload for analytics (e.g., regression/predictive models, outliers/concept drift detection), as close to the edge as possible brings serious benefits, particularly where communication costs are high or where instant action/decision is needed. But, today's edge capabilities are still relatively unsophisticated in light of quality of analytics, lacking anything like the computing power Cloud can provide.

We rest at the fact that pushing analytics towards the edge is feasible because of the increase of computational power on sensing & actuator devices. Their capabilities enables them to reduce network traffic and latency by supporting *in-network* real-time data analytics. However, analytics over contextual data at the edge should be *imperatively* provided with high quality of outcomes, e.g., low prediction errors, avoiding false alarms, taking into account efficient communication due to the above-mentioned constrains [1]. Within an edge network, it is deemed appropriate to introduce a methodology for providing quality aggregation and predictive analytics tasks departing from the traditional in-network data processing/delivery methods by exploiting the computing capability of the edge nodes.

Motivation: Let us consider the following motivating scenario that demands high quality analytics & efficient communication at the edge for car driver mirco-sleep identification [2]. In-car and driver mood-fatigue detection sensors locally sense the surrounding environment of the vehicle, road, and driver's physiological, facial and driving behavior [2] and transmit data via 5G towards the Cloud for processing/classification and alert the driver/emergency services if irregular patterns are detected. We identify three cases: (a) the driver fell into a micro-sleep and the system identifies that; (b) the driver is awake while the system identifies a micro-sleep pattern; (c) the driver is awake and the system identifies that. Case (a) demands low latency and real-time reaction to prevent accidents with high certainty on the analytics outcome; using unreliable broadband for data transmission cannot support real-time identification leading to horrible consequences. Case (b) encounters a false alarm, e.g., due to missing or obsolete transmitting values, resulting in bad consequences, e.g., the driver got shocked by the risen alarm causing the car go off the road. In case (c) the car sensors continuously transmit data without any action occurrence, thus, resulting in humongous redundant values and unnecessary bandwidth consumption increasing latency. From such cases, it is challenging to support sophisticated decisions on *when* and *which* data to process and deliver for supporting high quality of real-time analytics at the edge taking into account the induced communication overhead and latency. The research **challenges** this paper focuses on are: (1) deciding *which* data to communicate at the edge network *without* loosing quality of data and analytics outcomes at destination; (2) deciding *when* to deliver data in light of obtaining high quality of analytics; (3) reducing *unnecessary* communication between/among edge devices and/or Cloud for saving bandwidth and decreasing latency.

A. Related Work & Contribution

Many baseline approaches [3] (and the references therein) collect *all* data from IoT environments, e.g., Wireless Sensor Networks (WNS) to centralized locations for centrally performing analytics tasks requiring, thus, *all* devices to continuously sensing and communicating. However, due to bandwidth, latency and energy constrains alternative methodologies have been studied [4], [5], [6] especially for WSNs based on *selective forwarding*. In these approaches data are conditionally transmitted to central locations reducing communication overhead. However, such approaches focus only on communication efficiency and are unaware of the analytics tasks performed at the destination, thus, cannot be adopted to support high quality of analytics. Advanced selective forwarding methods [7], [8] deal with dynamic optimal decisions of finding the best time to deliver data in light of communication efficiency and reconstruction error minimization at the destination. Nonetheless, such optimal decision making is limited on communication overhead, not applied on the network edge and not taking into account its impact on the quality of advanced analytics like aggregation and predictive tasks. From the edge-analytics perspective, recent works [9], [10], [11] exploit the computational power of devices to launch (lightweight) algorithms directly at the data sources. However, such approaches are unaware of communication efficiency in the edge network as supported by the above-mentioned selective forwarding approaches. Our previous work [12] investigates the impact of a prediction-based selective forwarding decision, purely from the communication objective, on aggregation and predictive analytics. This signals the necessity of introducing a hybrid and sophisticated decision making model on *when* & *which* data to process and deliver for trading between quality of (advanced) analytics and communication efficiency at the network edge. Our proposed method in this paper advances on time-optimized data forwarding and data processing decisions based on the historical patterns of data forwarding decisions and the predictive capability of the edge nodes to determine the *best time* and the *most appropriate* data to deliver in light of maximizing the quality of aggregation and predictive analytics tasks at the destination being, in parallel, communication efficient. This is achieved based on a quality-aware, intelligent monitoring scheme over the cumulative reconstruction error at the destination under the principles of the Optimal Stopping Theory (OST) [13]. Our **contribution** is summarized as follows:

- an optimal, quality-aware decision making model determining when and which data to deliver in the network edge in light of maximizing the quality of analytics by being communication efficient;
- mathematical analysis based on the principles of the theory of optimal stopping [13] and incremental methods for evaluating the optimal decision in real-time;
- two real-time model variants exploiting the computational capabilities of the collaborating edge devices over real contextual data streams;
- comparative & performance assessment with aggregation and linear regression models using statistical & information theoretic metrics comparing our model with the methodologies [12], [4], [5], [6] following the selective forwarding scheme;

The paper is organized as follows: Section II discusses the rationale and provides fundamental definitions for the quality analytics metrics, while Section III presents the overall approach and problem formulation. Section IV elaborates on the solution fundamentals, while Section V reports on the performance and comparative assessment. Finally, Section VI concludes the paper with future research directions.

II. Rationale & Fundamentals

A. Rationale

We abstract an edge network architecture through Edge Nodes (ENs) forming a layer between Sensing & Actuator Nodes (SANs) and the Cloud. Several SAN are connected to each EN, e.g., cloudlet, sink node in a WSN. Since ENs are located close to the SANs, contextual data should be efficiently transferred to them in real-time. The fundamental desiderata to materialize analytics at the edge are: (D1) the autonomous nature of SANs to locally perform sensing and determine whether and which data to transfer to ENs or not in light of minimizing the required communication (overhead) at the expense of accurate and quality analytics tasks performed on ENs; (D2) the capability of ENs to locally reconstruct undelivered data and perform aggregation/predictive analytics tasks. We assume a tree-like topology in which a SAN i is connected with its EN j. We notate the neighborhood of EN j as the set of SANs $\mathcal{N}_j = \{1, \ldots, n_j\}$, i.e., $i \in \mathcal{N}_j$. We assume a discrete time domain $t \in \mathbb{T} = \{1, 2, \ldots\}$ such that SAN i, at every time instance $t \in \mathbb{T}$, senses a d-dimensional row vector $\mathbf{x}_t = [x_{1t}, \ldots, x_{dt}] \in \mathbb{R}^d$ of contextual parameters, e.g., temperature, humidity, air pollutant chemical compounds. Hereinafter, we refer to $\mathbf{x}_t \in \mathbb{R}^d$ as *context vector* at time t which forms the communication between SAN i and EN j.

A sliding window $\mathcal{W}$ is specified by a fixed-size temporal extent $N > 0$ by appending new context vectors and discarding older ones on the basis of their appearance. At time $t \in \mathbb{T}$, a sliding window $\mathcal{W}$ is a sequence of all context vectors observed from $t - N$ to $t - 1$, i.e., $\mathcal{W} = (\mathbf{x}_{t-N}, \mathbf{x}_{t-N+1}, \ldots, \mathbf{x}_{t-1})$ and is most widely used in continuous analytics [14], [15]. Aggregation analytics are evaluated over $\mathcal{W}$, which change over time as the window slides. There are three categories of aggregation functions: distributive, algebraic and holistic [16]; notably `MAX` and `MIN` are distributive, `AVG` is algebraic computed from `SUM` and `COUNT`, and `QUANTILE`, `MEDIAN` are holistic. For instance, `AVG` is defined over $\mathcal{W}$ as: $h(\mathcal{W}) = \frac{1}{N} \sum_{k=t-N}^{t} \mathbf{x}_k$.

One the most used predictive analytics models is the multivariate linear regression [17]. Given a $\mathcal{W}$ with vectors $\mathbf{x}_t = [\mathbf{x}_t^{in}, y_t^{out}] \in \mathbb{R}^d$ representing input-output pairs within the last N measurements, the linear regression model estimates the *current* coefficient $\mathbf{w} \in \mathbb{R}^d$, which interprets the current dependency of $\mathbf{x}^{in}$ with y^{out}:

$$\mathbf{w} = \arg\min_{\mathbf{w}' \in \mathbb{R}^d} \frac{1}{N} \sum_{t=1}^{N} \left(y_t^{out} - (\mathbf{x}_t^{in})^\top \mathbf{w}' \right)^2 \quad (1)$$

The predicted output $\hat{y}^{out}$ provided by the *actual* linear model $\mathbf{w}$ over $\mathcal{W}$ is $\hat{y}^{out} = (\mathbf{x}^{in})^\top \mathbf{w}$ and the Root Mean Square Error (RMSE) over n predictions is defined as:

$$\epsilon = \left(\frac{1}{n} \sum_{k=1}^{n} (y_k^{out} - \hat{y}_k^{out})^2 \right)^{1/2}. \quad (2)$$

The methodologies in [12], [4], [5], and [6] adopt a *selective forwarding* rule to decide whether to deliver a context vector in the edge network or not in light of minimizing the communication overhead defying the quality of analytics tasks. Such methodologies are based on an *Instantaneous Decision Making* (IDM) using only (i) the current vector $\mathbf{x}_t$ and (ii) the expected (predicted) vector $\hat{\mathbf{x}}_t$. To apply this methodology in our context, SAN i is equipped with a vector prediction algorithm $f_i(\mathbf{x}_{t-1}, \ldots, \mathbf{x}_{t-N})$, which uses the recent past $N \geq 1$ sensed vectors stored in window $\mathcal{W}$ of size N to predict the future vector $\hat{\mathbf{x}}_t$ at time t:

$$\hat{\mathbf{x}}_t = f_i(\mathbf{x}_{t-1}, \ldots, \mathbf{x}_{t-N}) = f_i(\mathcal{W}). \quad (3)$$

SAN i after sensing $\mathbf{x}_t$ predicts $\hat{\mathbf{x}}_t$ with prediction error:

$$e_t = d^{-\frac{1}{2}} \|\mathbf{x}_t - \hat{\mathbf{x}}_t\|, \quad (4)$$

where $\|\mathbf{x}\| = (\sum_{k=1}^{d} x_k^2)^{1/2}$ is the Euclidean norm of $\mathbf{x}$ and $d^{-1/2}$ is a normalization factor to ensure that $e_t \in [0, 1]$, given that $\mathbf{x} \in [0, 1]^d$ is scaled in the d-dimensional unit cube; each dimension $x_k, k = 1, \ldots, d$ ranges in $[0, 1]$. Such prediction capability yields SAN able to decide whether to send $\mathbf{x}$ to its EN or not for processing based on a θ-based IDM rule:

- **Case I** If predicted $\hat{\mathbf{x}}_t$ differs from actual $\mathbf{x}_t$ w.r.t. *decision threshold* $\theta \in (0, 1)$, i.e., $e_t > \theta$, SAN i sends $\mathbf{x}_t$ to EN j.
- **Case II** Otherwise, i.e., $e_t \leq \theta$, SAN i does not send $\mathbf{x}_t$ to EN j and EN j is responsible for reconstructing the undelivered vector locally.

In Case I, EN j receives the actual $\mathbf{x}_t$ from SAN i. In Case II, EN j should adopt a reconstruction function

$$\tilde{\mathbf{x}}_t = g_j(\mathbf{u}_{t-1}, \ldots, \mathbf{u}_{t-M}) = g_j(\mathcal{W}), \quad (5)$$

of the recent $M \geq 1$ vectors $\mathbf{u}$ from its window $\mathcal{W} = (\mathbf{u}_{t-M}, \ldots, \mathbf{u}_{t-1})$ to reconstruct the undelivered $\mathbf{x}_t$, notated by $\tilde{\mathbf{x}}_t$ using only historical vectors. The vectors $\mathbf{u}$ in the EN's $\mathcal{W}$ correspond to either the actual $\mathbf{x}$ from SAN i (Case I) or the past locally re-constructed vectors $\tilde{\mathbf{x}}$ from g_j (Case II): $\mathbf{u}_t = \mathbf{x}_t$ if $e_t > \theta$ (Case I); otherwise $\mathbf{u}_t = \tilde{\mathbf{x}}_t = g_j(\mathcal{W})$ (Case II). The reconstruction error at EN j is then:

$$a_t = \begin{cases} 0 & \text{Case I,} \\ \|\mathbf{x}_t - \tilde{\mathbf{x}}_t\| & \text{Case II.} \end{cases} \quad (6)$$

The aggregation & regression analytics functions are running on EN j *for each* sliding window $\mathcal{W}$ containing M received and/or re-constructed vectors from the SANs $i \in \mathcal{N}_j$ depending on cases I and II. We now introduce the *discrepancies of the analytics* on EN due to the fact that EN does not always receive the actual vectors from SAN.

B. Definitions

Definition 1 (Aggregation Analytics Discrepancy)**.** *Given a pair (SAN i, EN j), the aggregation discrepancy γ between the analytics output on EN j derived from aggregation function h over its window $\mathcal{W}$ and the actual analytics output on SAN i over window $\mathcal{W}^*$, which contains the actual context vectors (ground truth) is: $\gamma = \|h(\mathcal{W}) - h(\mathcal{W}^*)\|$.*

The discrepancy γ denotes how much the aggregation results over $\mathcal{W}$ with vectors $\mathbf{u}$ on EN j *differ* from the aggregation results over $\mathcal{W}^*$ with actual vectors $\mathbf{x}$, should SAN i have sent them all to EN j. In Case I we obtain $\gamma = 0$, while in Case II, $\gamma \geq 0$ since EN j needs to reconstruct undelivered vectors. We require to tolerate a low γ in light of communication efficiency.

Definition 2 (Regression Performance Discrepancy)**.** *Given a pair (SAN i, EN j), the regression discrepancy δ is defined as the difference of the RMSE ϵ derived from the linear model $\mathbf{w}$ estimated over EN j's window $\mathcal{W}$ and the RMSE ϵ^* derived from the ground truth linear model $\mathbf{w}^*$ estimated over the actual SAN i's vectors in $\mathcal{W}^*$: $\delta = |\epsilon - \epsilon^*|$.*

Definition 3 (Model Fitting Discrepancy)**.** *Given a pair (SAN i, EN j), the model fitting discrepancy δ' is defined as the distance $\delta' = \|\mathbf{w} - \mathbf{w}^*\|$ from the model $\mathbf{w}$ estimated over EN j's window $\mathcal{W}$ and the ground truth model $\mathbf{w}^*$ estimated over the actual SAN i's vectors in $\mathcal{W}^*$.*

The δ discrepancy measures the difference in the quality of the *predictive performance* of the predictive analytics (linear regression analytics) performed at EN j. The RMSE ϵ^* refers to the prediction w.r.t. $\mathbf{w}^*$ over the actual pairs $(\mathbf{x}^{in}, y^{out})$. Since EN j may not receive all the actual pairs due to Case II, the derived model $\mathbf{w}$ results to a RMSE $\epsilon \neq \epsilon^*$.

The δ' discrepancy measures the *distance* of the derived linear model at EN j from the ground truth model at SAN i. Due to Case II, the model fitting achieved in EN j might be different (coefficients-wise) from the ground truth linear model fitting. We require to tolerate a low δ in terms of prediction performance and a low δ' in terms of linear model fitting by being communication-efficient.

C. Problem Fundamentals

IDM attempts to increase the communication efficiency by saving significant network bandwidth but at the expense at analytics discrepancies. Fundamentally, IDM disregards the history of analytics discrepancies that ENs are experiencing. The vector forwarding decision is purely based on the current prediction error on SANs and does not take into consideration the past predictions. Obviously, the analytics discrepancies are unknown to EN since not all the actual vectors are sent for communication efficiency. On the other hand, such discrepancies are unknown to the SANs because, even if the actual vectors are sensed locally, SANs are not equipped with reconstruction and analytics functions. The only information a SAN has is a series of its prediction errors $\{e_t\}$. We will show that based on this series our approach provides high quality of analytics while being communication efficient.

There are two major concerns in IDM:(**C1**) If θ is relatively high, SAN i scarcely updates EN j with actual vectors. Hence, EN j loses significant information, which is expected degrading analytics results. We encounter the same situation if the prediction error e_t is relatively small; if for a low θ we encounter $e_t \ll \theta$, EN j does not follow the data stream. This is happening e.g., when the predictor f_i of SAN i is very accurate. This is counterintuitive, since we desire to have an accurate predictor f_i, but its instantaneous outcome for decision making leads to the situation of information loss on EN j. Figure 1 (upper) shows the case where f_i predictor (here, exponential smoothing) at SAN i produces predictions close to the actual data, thus, resulting in no communication and thus information loss at EN j, which re-constructs the data with g_j (adopting exponential smoothing). (**C2**) If there are certain outliers or novel cases/significant events in SAN i, SAN i delivers the associated vectors to EN j and then transits back to the state of non delivering vectors to EN j. In this situation, EN j accumulates most of the time outliers and novel vectors and, again, the re-constructed data do not follow the in-between actual data; see Figure 1 (lower).

Departing from IDM, given a decision threshold $\theta \in (0, 1)$ at SAN i, we will derive sufficient conditions for a novel, time-optimized, analytics discrepancy-aware decision making, where the vector forwarding decision is a function of both the desired error bound and correlation among data. When θ is very tight or the correlation is not significant, SAN i always has to forward its vectors to EN j. Due to the characteristics and inherent dynamics of SANs' data, e.g., underlying data distribution evolves over time, prediction techniques may not work efficiently for a set of less predictable data. Moreover, there might be dependencies among data from neighboring SANs (data locality in $\mathcal{N}_j$), thus, EN j is capable of learning those dependencies in a communication-efficient way, as will be shown later.

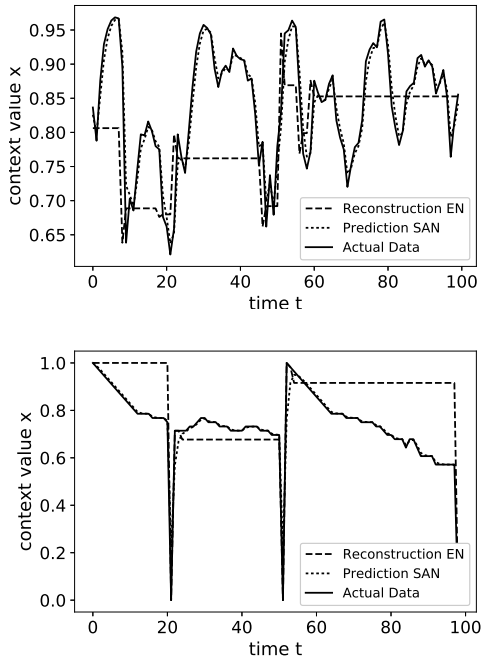


Figure 1: The actual data sensed at SAN, the predicted data at SAN and the reconstructed data at EN vs. time demonstrating (upper) concern C1 and (lower) concern C2.

III. Quality Analytics-Aware Decision Making

A. Overall Approach

As IDM is not capturing the variability of the data stream inside EN, it results in information loss of the actual SAN's data. Our approach departs from IDM, at the first instance, by taking into consideration past IDM decisions, i.e., decisions purely based on either $e_t > \theta$ or $e_t \leq \theta$. Our approach elaborates on historical prediction error-aware decision making, which takes into consideration IDM decisions made at past time instances $\tau < t$ and the current decision at time t to decide on vector forwarding. Our method quantifies this historical context by accumulating the prediction errors $e_\tau : \tau \leq t$ with $e_\tau \leq \theta$ from which non-forward decisions were taken. We encounter the past non-forward decisions as useful information in our method since this cumulative error on SAN i relates to the cumulative reconstruction error on EN j, thus, influencing the quality of analytics, as proved later. The non-forward decision indicates that the error is tolerable w.r.t. θ, however, its cumulation before a forward decision results to information loss, thus, cumulation of reconstruction error on the EN. Our idea is to exploit even those relatively small discrepancies for decision making and to tolerate up to a certain extent this cumulation. Nonetheless, in IDM, it is not instantly obvious the impact of the error cumulation on the quality of analytics at EN. We enforce SAN i not only to track the

current prediction error at t but also the cumulative error up to t. Obviously, SAN i cannot monitor the expected reconstruction error at EN j to take a decision at t; recall that only the expected prediction error (up to t) is available to SAN i. Based only on this information, the challenge is to monitor the behavior of the cumulative prediction error at SAN investigating which is its relation with the expected reconstruction error at EN, and up to which discrepancy tolerance this cumulative error is allowed to be in order to forward vectors from SAN i to EN j. We show that by monitoring the expected prediction error at SAN i suffices to take more sophisticated and certain decisions on vector forwarding/non-forwarding.

Consider the case SAN i decides not to forward $\mathbf{x}_t$ to EN j and let us define $\hat{\mathbf{x}}_t = \tilde{\mathbf{x}}_t + \boldsymbol{\rho}_t$, where $\boldsymbol{\rho}_t$ is the *vector discrepancy* of the predicted vector at SAN i and the reconstructed vector at EN j, given that $e_t \leq \theta$ and $\mathbb{E}[\|\boldsymbol{\rho}\|] < \infty$. Our target is to relate the conditional expectation of the prediction error $\mathbb{E}[e|e \leq \theta]$ with the expected reconstruction error $\mathbb{E}[a]$ given that SAN i does not forward context vectors to EN j. We obtain that:

$$\begin{aligned}\mathbb{E}[a] &= \mathbb{E}[\|\mathbf{x}-\tilde{\mathbf{x}}\||e \leq \theta]P(e \leq \theta) + 0 \cdot P(e > \theta)\\ &= \mathbb{E}[\|(\mathbf{x}-\hat{\mathbf{x}}) + (\hat{\mathbf{x}}-\tilde{\mathbf{x}})\||e \leq \theta]P(e \leq \theta)\\ &\leq \mathbb{E}[\|\mathbf{x}-\hat{\mathbf{x}}\||e \leq \theta]P(e \leq \theta) + \mathbb{E}[\|\boldsymbol{\rho}\|\big|e \leq \theta]P(e \leq \theta)\\ &\leq (\mathbb{E}[e|e \leq \theta] + \mathbb{E}[\|\boldsymbol{\rho}\|\big|e \leq \theta])P(e \leq \theta) \quad (7)\end{aligned}$$

We obtain from (7) that the expected reconstruction error is bounded at least by the conditional expectation of the prediction error given that SAN i does not forward vectors to EN, which is *known* at SAN i and the conditional expectation of the discrepancy $\mathbb{E}[\|\boldsymbol{\rho}\|\big|e \leq \theta]$ derived by the intrinsic difference of the reconstructed and predicted vectors. Interestingly, if SAN and EN adopt the same algorithm for prediction and reconstruction, e.g., exponential smoothing, then $\boldsymbol{\rho}$ can be directly known to SAN. Based on this outcome, our idea is that SAN tracks the cumulative sum of prediction errors for decision making as this reflects the cumulative reconstruction error shown in (7).

Consider now the event $\{e_t > \theta\}$ where SAN forwards $\mathbf{x}_t$ to EN thus there is on reconstruction error. Our method takes also into account this decision to deal with a more sophisticated decision making since $e_t > \theta$ might not only reflect the capability of the prediction algorithm but also the fact that the sensed data on SAN is rather unpredictable with significant peaks or outliers or even events that are of high importance. This knowledge cannot be derived instantly adopting IDM. Instead, a continuous (not necessarily strictly sequential) observations of events $\{e_t > \theta\}$ is deemed appropriate to be taken into consideration for decision making. Our method proceeds with the quantification of these significant events by cumulating the scaled lower bound of the error excess w.r.t. θ, i.e., through a cumulative sum of quantities $\lambda\theta$ for each event $\{e_t > \theta\}$ with $\lambda > 0$. As it will be shown, the value of λ and the priority of vector forwarding upon the event $\{e_t > \theta\}$ leads to two variants.

Our method attempts to smooth the re-constructed data stream on the EN by taking into consideration (i) the cumulative prediction error avoiding concern C1 and (ii) the cumulation of events avoiding concern C2. This is achieved by optimally deciding on vector forwarding combining the error cumulation in cases $\{e_t \leq \theta\}$ and the significance of events in cases $\{e_t > \theta\}$. This leads to a model which drastically departs from IDM methods attempting to deal with the concerns C1 & C2.

B. Real-Time Decision Making Model

Our model optimally postpones vector forwarding in light of reducing communication and on the other hand increasing the quality of analytics. The problem is to identify *when* SAN should take a forward decision at t based on the current e_t, the cumulation of prediction errors and events occurrences up to t. Given a fixed θ, which is application specific, if SAN delays to forward vectors to EN, we gain communication efficiency but EN cannot re-construct the data, thus, degrading the quality of analytics. If SAN forwards data with a high rate, we achieve high quality analytics at the expense of communication overhead. We seek a stochastic decision making model to deal with this trade-off by maximizing the delay tolerance (thus saving communication) at the expense of quality of analytics. Based on the conditional expectation of the prediction error at SAN, which is an upper bound of the expected reconstruction error on EN (see (7)), we define the stochastic indicator Z_t whose value depends on $e_t = \|\mathbf{x}_t - \hat{\mathbf{x}}_t\|$:

$$Z_t = \begin{cases} \lambda\theta & \text{if } e_t > \theta, \\ e_t & \text{if } e_t \leq \theta. \end{cases} \quad (8)$$

A value $Z_t = \lambda\theta$ indicates an event which cannot be accurately predicted by f_i at SAN w.r.t. θ or signals a significant peak or outlier/novelty at t. A value $Z_t = e_t \leq \theta$ indicates the acceptable error in term of quality tolerance, which is accumulated also at EN. In both cases, the cumulation of Z_t values at $t = 0, 1, \ldots$ enforces SAN to decide on vector forwarding based on the history of $\{e_t \leq \theta\}$ and $\{e_t > \theta\}$. We abstract this cumulative enforcement through the cumulative sum of either prediction errors or tolerances up to t, i.e., $S_t = \sum_{\tau=0}^{t} Z_\tau$. Since the quantity S_t up to t provides information to SAN whether to further postpone vector forwarding or not, we define the *reward tolerance function* at time t as:

$$Y_t = \beta^t S_t = \beta^t \sum_{\tau=0}^{t} Z_\tau, \quad (9)$$

with tolerance discount factor $\beta \in (0, 1)$ acting as an adviser on delaying vector forwarding or not. Y_t represents the stochastic tolerance for non-forward decisions up to t. The idea is to postpone vector forwarding as much as

possible, thus saving communication, but not to degrade the analytics quality at EN. $\beta \to 1$ suggests SAN to further postpone vector forwarding in light of minimizing the communication, while $\beta \to 0$ suggests SAN to proceed with vector forwarding at an earlier stage in light of minimizing the reconstruction error at EN. Based on the randomness of the events (randomness of Z_t and S_t), SAN tries to find the *optimal vector forwarding time* t^* to maximize the expectation of Y_t, $\mathbb{E}[Y_t]$, given fixed β and θ. Formally:

Problem 1. *Find the optimal vector forwarding time t^* at which the supremum of the expectation of Y_t is attained:*

$$\sup_{t \geq 0} \mathbb{E}[Y_t]. \tag{10}$$

SAN tracks the Z_t values at $t = 0, 1 \ldots$, and decides to forward $\mathbf{x}_t$ at time t^*, which maximizes $\mathbb{E}[Y_t]$. Based on the value of $\lambda \in \{0, 1\}$ we contribute with two variants of our method to cope with both C1 and C2 and provide the trade-off between quality analytics and communication.

IV. Quality-Aware Optimal Vector Forwarding

A. Solution Fundamentals

We first prove that the optimal forwarding time t^* for Problem 1 exists provided in Theorem 1 based on the principles of OST [13] and provide an optimal forwarding rule for evaluating it at Theorem 2. Then, we report on the two proposed variants.

Theorem 1. *The optimal vector forwarding time for Problem 1 exists.*

Proof: Based on the principles of OST [13], to prove the existence of t^* we need to prove that the conditions A1 and A2 for Y_t in (9) are satisfied: (A1) $\limsup_t Y_t \leq Y_\infty = 0$ almost surely and (A2) $\mathbb{E}[\sup_t Y_t] < \infty$. A1 implies that with the elapse of time ($t \to \infty$) the reward should go to zero, i.e., $Y_\infty = 0$, since no vector delivery with indefinite horizon is useless due to extremely high reconstruction error at EN; $Y_\infty = 0$ represents the reward of an endless non delivery phase. A2 implies that the expected reward under any policy is finite. We first focus on the supremum limit of Y_t notated by $\limsup_t Y_t$, i.e., the limit of $\sup_t Y_t$ as $t \to \infty$ or $\lim_{t\to\infty}(\sup\{Y_k : k \geq t\})$. Note, Z_t are non-negative and from the strong law of numbers $(\frac{1}{t}\sum_{k=1}^{t} Z_k) \to \mathbb{E}[Z]$:

$$Y_t = t\beta^t(S_t/t) = t\beta^t(1/t)\sum_{k=1}^{t} Z_k \sim t\beta^t\mathbb{E}[Z] \overset{\text{a.s.}}{\to} 0,$$

that is $\lim_{t\to\infty} \sup_t Y_t = 0$. Moreover, we have $Y_\infty = 0$ by definition, thus, A1 is satisfied. For A2, we have

$$\sup_t Y_t = \sup_t \beta^t \sum_{k=1}^{t} Z_k \leq \sup_t \sum_{k=1}^{t} \beta^k Z_k \leq \sum_{k=1}^{\infty} \beta^k Z_k.$$

Hence,

$$\mathbb{E}[\sup_t Y_t] \leq \sum_{k=1}^{\infty} \beta^k \mathbb{E}[Z] = \mathbb{E}[Z]\frac{\beta}{1-\beta} < \infty.$$

Therefore, it is shown that the optimal forwarding time in (10) exists. ∎

Theorem 2. *SAN decides to forward vector $\mathbf{x}_{t^*}$ at time t^*:*

$$t^* \quad = \quad \inf\{t \geq 1 | \sum_{k=1}^{t} Z_k \geq \frac{\beta}{1-\beta}\mathbb{E}[Z]\}. \tag{11}$$

Proof: Since Y_t are non-negative, Problem 1 is *monotone* [18] thus the optimal time t^* is obtained by the *one-stage look-ahead optimal rule* (1-sla):

$$t^* = \inf\{t \geq 1 | Y_t \geq \mathbb{E}[Y_{t+1}]\}.$$

The adoption of 1-sla is optimal since $\sup_t Y_t$ has finite expectation (equal to $\mathbb{E}[Z]\frac{\beta}{1-\beta}$) and $\limsup_t Y_t = 0$ almost surely as proved in Theorem 1. Hence, t^* is estimated through the principle of optimality; suppose that $S_t = s$ and SAN decides that it is optimal to forward a vector. Then, the current reward of $\beta^t s$ is at least as large as any expected $\mathbb{E}[(\frac{\beta}{1-\beta})^{t+\tau}(s+S_\tau)]$, which means that: $s(1-\mathbb{E}[(\frac{\beta}{1-\beta})^\tau]) \geq \mathbb{E}[(\frac{\beta}{1-\beta})^\tau S_\tau]$ for all times τ. This must hold true for all $s' \geq s$, thus, the optimal time t^* for some s_0 must be of the form $t^* = \inf\{t \geq 1 | S_t \geq s_0\}$. That is, SAN forwards at the first t for which $S_t \geq s_0$. Now, the tolerance for forwarding s_0, must be the same as the tolerance for continuing using the 1-sla that forwards the first time the sum of tolerances is positive. That is, s_0 must satisfy the equation

$$s_0 = \mathbb{E}[(\frac{\beta}{1-\beta})^\tau(s_0 + S_\tau)],$$

with $\tau = \inf\{t \geq 1 | S_\tau > 0\}$. Since Y is non-negative, we obtain $\tau \equiv 1$ and $S_\tau \equiv Y$ [18] and then replacing with $s_0 = \frac{\beta}{1-\beta}\mathbb{E}[Y]$ we obtain: $t^* = \inf\{t \geq 1 | \sum_{k=1}^{t} Z_k \geq \frac{\beta}{1-\beta}\mathbb{E}[Z]\}$. ∎

B. Evaluation of the Optimal Vector Forwarding Time

Theorem 2 provides us with the optimal forwarding time t^* for SAN. At each time t SAN observes the events $\{e_t \lesseqgtr \theta\}$, evaluates Z_t and S_t. If the criterion (11) holds true then SAN forwards $\mathbf{x}$ to EN and resets the sum to zero starting a new 'era' of optimal vector forwarding. The triggering of (11) requires the knowledge of $\mathbb{E}[Z]$ at SAN, which is now associated with the conditional expectation of the prediction error $\mathbb{E}[e|e \leq \theta]$ as discussed in Section III-A, which is known to SAN. We contribute with an incremental mechanism to estimate $\mathbb{E}[Z]$ based on the expected prediction error on SAN. Specifically we obtain that

$$\begin{aligned} \mathbb{E}[Z] \quad &= \mathbb{E}[Z|e > \theta]P(e > \theta) + \mathbb{E}[Z|e \leq \theta]P(e \leq \theta) \quad (12) \\ &= \quad \lambda\theta - \int_0^\theta (\lambda\theta - e)p(e)de = \lambda\theta - \mathcal{I}(\theta), \end{aligned}$$

where $\mathcal{I}(\theta) = \int_0^\theta (\lambda\theta - e)p(e)de$ and $p(e)$ is the Probability Density Function (PDF) of the prediction error in SAN. Notably, the criterion (11) is based on the estimation of $\mathcal{I}(\theta)$, which involves estimation of $p(e)$. The approximation of $p(e)$ at t, notated by $\hat{p}^{(t)}(e)$, is based on incremental Kernel Density Estimation (KDE) from the sequence $e_1, \ldots, e_t$:

$$\hat{p}^{(t)}(e) \quad = \quad \frac{1}{t}\sum_{k=1}^{t} K_h(e - e_k), \tag{13}$$

where $K_h(u)$ is a kernel function, unimodal, symmetric, non-negative that centers at zero and integrates to unity while the window h controls the degree of smoothing of the estimation. The PDF of e is then estimated incrementally as:

$$\hat{p}^{(t)}(e) \quad = \quad \frac{t-1}{t}\hat{p}^{(t-1)}(e) + \frac{1}{t}K_h(e - e_t) \tag{14}$$

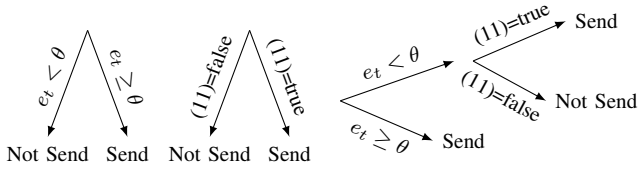
with $\hat{p}^{(1)}(e) = K_h(e - e_1)$. The integral $\mathcal{I}(\theta)$ can be then incrementally estimated based on $\hat{p}^{(t-1)}(e)$ and e_t at $t > 1$ based on the recursion:

$$\mathcal{I}^{(t)}(\theta) = \frac{t-1}{t}\mathcal{I}^{(t-1)}(\theta) + \frac{1}{t}\int_0^\theta (\theta - u)K_h(u - e_t)du \tag{15}$$

When e_t is obtained by SAN *only* the evaluation of $q^{(t)}(e) = \frac{1}{t}\int_0^\theta (\theta - u)K_h(u - e)du$ is needed for checking the criterion (11) with initial $\mathcal{I}^{(1)}(\theta) = q^{(1)}(e_1)$ at time $t = 1$. There are certain kernels K_h that can be adopted here, e.g., Epanechnikov and Gaussian kernel. The Gaussian kernel is mostly used due to its convenient mathematical properties and, especially, when dealing with PDF estimation. We adopt the Gaussian $K_h(u) = \frac{1}{\sqrt{2\pi}h}e^{-\frac{1}{2}(\frac{u}{h})^2}$ where the optimal value of h is $h^* = 1.06 \min\left(\hat{\sigma}, \frac{\hat{R}}{1.34}\right) T^{-\frac{1}{5}}$ [19], where $\hat{\sigma}$ is the standard deviation of e, $\hat{R}$ is the interquartile range, and T is the number of training error values. Based on K_{h^*}, SAN easily calculates $q^{(t)}(e_t)$[1] and evaluates the criterion (11) for vector forwarding in $O(1)$ through $\mathcal{I}(\theta)$ in (13).

C. Quality-Aware Model Variants

We propose two variants depending on the value of λ, which plays a significant role on decision making. With $\lambda = 1$, we obtain the *pure* Optimal Vector Forwarding (OVF) variant, which uses θ as the least tolerance value if e_t exceeds θ in (8). OVF *always* increases the cumulative sum S_t even if the predictor f in SAN produces accurate forecast w.r.t. θ or not. This presents a *strict* variant which takes into consideration even the relatively small prediction errors for deciding on vector forwarding. Figure 2 (b) shows the OVF decision tree, which is purely based on S_t and a forwarding decision is triggered w.r.t. (11) compared to IDM (Figure 2 (a)). With $\lambda = 0$, we obtain a variant which imposes a *penalty* only when the predictor f in SAN forecasts correctly the expected context and acts immediately when



(a) IDM (b) OVF (c) Hybrid OVF

Figure 2: The decision trees for IDM and variants OVF and HOVF. OVF is triggered based on (11); HOFV combines both decision trees of IDM and OVF.

the prediction error exceeds θ. In this case, S_t does not always monotonically increase thus making a forwarding decision based on the inherent prediction capability of SAN. This variant, called as Hybrid OVF (HOVF), combines both the pure OVF in the case where $\{e_t < \theta\}$, thus, accumulating only the tolerances due to the prediction capability of the SAN (coping with C1), and the IDM in the case where the current prediction error exceeds θ, thus, capturing *immediately* any significant event/outlier/novelty (coping with C2). Figure 2(c) shows the HOVF decision tree fusing decisions of OVF for C1 and IDM for C2.

Both variants require a vector prediction algorithm f_i at SAN i following the evolving nature of the data streams and a reconstruction algorithm g_j at EN j that supports the analytics tasks for the pair (i, j). We are seeking to reduce the computational power for prediction and reconstruction at SAN and EN thus using a small fraction of their computing power we adopt the multivariate exponential smoothing [20], used for time series forecast, as an ideal predictor with computational complexity $O(d)$ in a d-dimensional space[2]. Exponential smoothing weighs the current vector with the historic vectors and is adopted as the function f_i for predicting $\hat{\mathbf{x}}$ and as the function g_j for re-constructing $\tilde{\mathbf{x}}$. At time t, a smoothed vector $\mathbf{s}_t$ is calculated by using the current vector $\mathbf{x}_t$ and the previous smoothed vector $\mathbf{s}_{t-1}$:

$$\mathbf{s}_t = \alpha\mathbf{x}_t + (1 - \alpha)\mathbf{s}_{t-1}, \tag{16}$$

initializing with $\mathbf{s}_0 = \mathbf{x}_0$ and $\alpha \in [0, 1]$. A higher α denotes more importance to the current vector and less importance to the historic vectors; normally, $\alpha = 0.7$ [20]. The calculated smoothed vector $\mathbf{s}_{t-1} = [s_{1,t-1}, \ldots, s_{d,t-1}]$ refers to the predicted vector $\hat{\mathbf{x}}_t$: $\hat{\mathbf{x}}_t = f_i(\mathcal{W}) = \mathbf{s}_{t-1}$ with the window $\mathcal{W} = (\mathbf{s}_{t-1})$ at SAN i containing only the recent smoothed vector. EN j, at time t either receives $\mathbf{x}_t$ or nothing. In the former case EN j inserts the delivered $\mathbf{x}_t$ into its window $\mathcal{W}$ (which is associated with SAN $i \in \mathcal{N}_j$) discarding the oldest vector, i.e., $\mathbf{u}_t = \mathbf{x}_t$. In the latter case EN j re-constructs the undelivered vector with the available vectors $\mathbf{u}$ reside currently in its $\mathcal{W}$ using exponential smoothing $g_j(\mathcal{W})$.

[1]Due to space limitations, the formulate of $q^{(t)}(e_t)$ with h^* is omitted.

[2]Holt-Winters smoothing can be adopted with the same complexity.

V. Performance & Comparative Assessment

A. Experimental Setup & Analytics Quality Metrics

We compare the performance of the OVF and HOVF variants with the models in [12], [4], [5], and [6], which implement the IDM methodology, over real contextual data described in [21]. The dataset contains $T = 10^4$ context vectors in a 12-dimensional $(d = 12)$ real data space corresponding to sensing air quality parameters reflecting 12 SANs of an edge network connecting with EN. For examining the reconstruction and aggregation analytics (experimenting with the AVG aggregation function, i.e., $h(\mathcal{W}) \equiv$ AVG), all context vectors are normalised and scaled, i.e., each parameter $x \in \mathbb{R}$ is mapped to $\frac{x-\mu}{\sigma}$ with mean value μ and variance σ^2 and scaled in [0,1], thus vector $\mathbf{x} \in [0,1]^d$. For examining the discrepancy in the regression analytics (in terms of performance and model fitting), after vector noramlisation, we divided the 12 air quality sensors to four SANs, where in each SAN two sensors serve as the $\mathbf{x}^{in}$ while the remaining sensor serves as the response y^{out}; we added the constant 1 to $\mathbf{x}^{in}$ to allow intercept in the linear regression model. The regression task at EN is therefore to predict the value of the 3rd sensor in each SAN using the first two. The discrepancy in the regression RMSE is achieved by 10-fold cross validation. The threshold $\theta \in \{10^{-5}, \ldots, 0.3\}$ ranging from sensitive to less sensitive quality of data capturing a range of context-aware applications, while the OVF (HOVF) factor β ranges in $\{0.1, \ldots, 0.999\}$ for investigating the impact of the forwarding tolerance on the quality of analytics. For all SANs and the EN, the exponential smoother (predictor and re-constructor) adopts $\alpha = 0.7$ as suggested in [20], while the window at each SAN is $N = 1$ (due to exponential smoothing) and at EN, $M = 10$, for each SAN.

Our target is to compare OVF (HOVF) with IDM variants in terms of communication overhead, reconstruction error, quality of aggregation tasks, and quality of predictive analytics (regression performance and model fitting). We measure the *percentage of communication*, i.e., context vectors transmitted by each model for each pair (SAN i, EN j) against the baseline solution, which forwards all actual vectors from SANs to EN. In terms of analytics quality, we examine the reconstruction error a due to undelivered vectors and the aggregation analytics outcome γ adopting the Symmetric Mean Absolute Percentage Error (SMAPE) per SAN. We use SMAPE as a quality metric due to its unbiased properties [22] representing a percentage value in $[0, 100]$ defined as: $\text{SMAPE} = \frac{100}{T}\sum_{t=1}^{T}\frac{a_t}{\|\mathbf{x}_t\|+\|\tilde{\mathbf{x}}_t\|}$ and $\text{SMAPE} = \frac{100}{T}\sum_{t=1}^{T}\frac{\gamma_t}{\|h(\mathcal{W})\|+\|h(\mathcal{W}^*)\|}$. Moreover, we adopt Kullback-Leibler (KL) divergence as a quality metric to measure the the *information loss* EN experienced due to the applied models over each reconstructed vector dimension $\tilde{x}$ from the actual dimension x after estimating their PDFs $p(\tilde{x})$ and $p(x)$, respectively, defined by: $KL(p(\tilde{x})\|p(x)) = \int_0^1 p(\tilde{x}) \log \frac{p(\tilde{x})}{p(x)} dx$. KL indicates the amount of information lost when EN j approximates the actual vectors at SANs i due to undelivered vectors. For assessing the quality of predictive analytics, we measure the discrepancy δ in the linear prediction performance w.r.t. RMSE and the model fitting discrepancy δ' in the actual and approximated linear models as defined in Section II-B.

B. Experimental Evaluation & Results

We identified during evaluation that the tolerance discount factor β is of high influence for the two optimal vector forwarding variants. In contrast, IDM does not depend on β and only variates with changing the threshold θ. Considering a fixed θ value, which is application-specific, for all models OVF, HOVF and IDM, we examine the quality of analytics at EN in terms of reconstruction error a, aggregation analytics γ, regression performance δ and model fitting δ' discrepancies as the tolerance factor β increases. As shown in Figure 3 increasing in β results in increasing a, γ, δ and δ' values for OVF and HOVF; IDM remains constant during all variations of β. A high β value refers to more tolerant OVF variants since they decide to further postpone vector forwarding in light of communication efficiency; however at the expense of quality of analytics. Specifically, OVF adopting $\beta > 0.8$ produces for all performance metrics a higher discrepancy than IDM. However, by adopting $\beta \leq 0.8$, OVF benefits of lower analytics discrepancies compared to IDM reflecting its flexibility of being less tolerant in terms of analytics quality while being communication efficient w.r.t. IDM. The HOVF variant achieves a better trade-off between tolerance and communication efficiency than the OVF and IDM models considering both concerns C1 and C2 simultaneously. Specifically, HOVF obtains an asymptotic behavior towards IDM with increase of tolerance in light of being communication efficient in all metrics. Interestingly, the analytics discrepancies for HOVF are with all values of β below the IDM, indicates that HOVF is deemed an appropriate method to be adopted for high quality analytics tasks w.r.t. IDM and OVF given a fixed θ; similar results are obtained for other θ values, which are not presented here due to space limitations. Even OVF can be preferred over IDM having $\beta \leq 0.8$ achieving higher quality analytics outcomes.

Besides the consideration of the analytics discrepancy, we also have to evaluate the information loss at EN with increasing the tolerance factor β (to achieve less communication overhead) with respect to how the reconstructed data PDFs at EN diverge from the actual data PDFs at SANs due to optimal vector forwarding decisions. Figure 4 (upper-left) illustrates that with the raise of β, the information loss, measured by KL metric, increases, which means that EN is less able to approximate the undelivered actual vectors of SANs. By adopting the IDM model as an upper bound of KL divergence value to compare the OVF and HOVF variants against, we observe that HOVF does not exceed this value of the IDM model even if the tolerance factor

is high ($\beta \rightarrow 1$). This denotes the capability of HOVF to optimally decide *not only when* to forward vectors to SAN but also *which* vectors helping SAN to accurately capture the statistical characteristics of the actual vectors at SANs. Similar behavior is achieved by OVF having $\beta \geq 0.8$; this differentiates HOVF from OVF in determining not only when but also which vector to deliver as reflected by the treatment of the concerns C1 and C2. Both optimal variants provide edge applications with the flexibility of achieving high quality of analytics, satisfiable capture of the statistical characteristics, and communication efficiency by tuning the tolerance factor β given a pre-determined application-specific accuracy threshold θ.

Figure 4 (upper-right) examines the induced (%) of communication for $\theta \in \{0.01, 0.06\}$ for all models against the tolerance factor β. Obviously, as $\beta \rightarrow 1$ both optimal forwarding models reduce the communication between SANs and EN, where IDM is not flexible to tune this percentage. It is interesting to mention that the HOVF variant exhibits an asymptotic behavior towards the IDM with an increase of β (for both θ values), thus, liaising this with its quality of analytics performance in Figures 3 and 4 (upper-left), demonstrates the successful trade-off between quality of analytics and communication efficiency. The HOFV variant provides us with the flexibility of obtaining high analytics quality while being communication efficient, which are the fundamental desiderata in edge analytics as discussed in Section II-A, while the OVF variant can support both desiderata having $\beta \in (0.5, 0.8)$. Figures 4 (lower-left/right) and 5 (upper-left/right) show the trade-off (%) of communication against reconstruction error a, aggregation discrepancy γ, KL and regression discrepancy δ for OVF and HOVF varying β from the lowest to highest with $\theta = 0.06$; for IDM varying θ from the lowest to highest. Both variants outperform the efficiency of IDM. Therefore, by adopting HOVF not only decrease the communication overhead and provides less information loss on EN, but also guarantees better quality of analytics at EN. Finally, Figure 5 (lower-right) shows the expected intermediate time between two consecutive forwarding decisions, i.e., the expected *delay* for vector forwarding, for all models against β having $\theta = 0.06$ for IDM. HOVF assumes the lowest delay in vector forwarding, since it attempts to forward the most appropriate vectors for achieving high quality of analytics, as discussed above. Interestingly, as $\beta \rightarrow 1$, the expected delay of HOVF approaches that of IDM indicating that even both models assume quite similar forwarding rates, HOVF intelligently choses to forward the best vectors for guaranteeing high analytics quality compared to the quality-unaware IDM. OVF appears more communication efficient especially for high β values at the expense of quality of analytics, while IDM remains inflexible in adapting to communication overhead constraints and accuracy of analytics results.

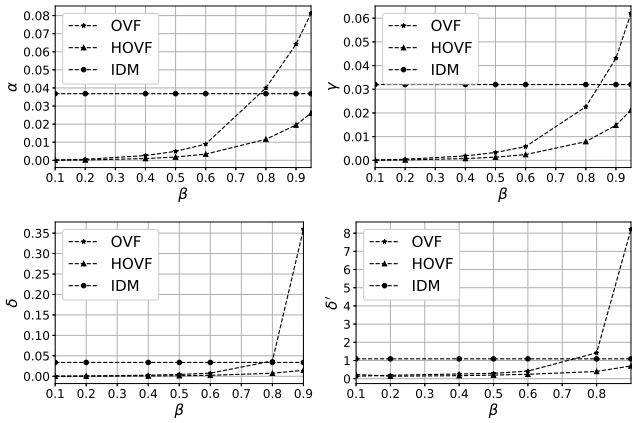


Figure 3: (Upper-left) reconstruction error a vs. β; (upper-right) aggregation discrepancy for AVG γ vs. β; (lower-Left) regression discrepancy δ vs. β; (lower-right) model fitting discrepancy δ' vs. β. Fixed threshold $\theta = 0.06$.

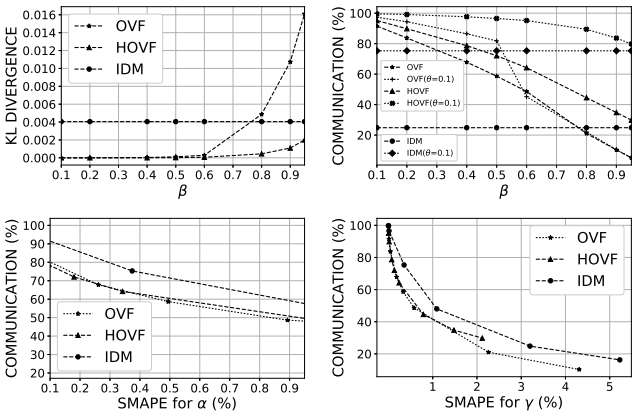


Figure 4: (Upper-left) KL divergence vs. β ($\theta = 0.06$); (upper-right) (%) communication vs. β for IDM, OVF and HOVF with $\theta = \{0.01, 0.06\}$; Trade-off for OVF, HOVF and IDM with $\theta = 0.06$ and all β between (%) communication and (lower-left) reconstruction error a, (lower-right) aggregation discrepancy γ.

VI. Conclusions & Future Work

We propose a novel, quality-aware and time-optimized decision making model for achieving high quality edge analytics while being communication efficient. We introduce the fundamental quality metrics and provide two variants exploiting the sensing & computational capabilities of nodes to perform on-line decision making. The edge nodes are enhanced to intelligently decide *when* and *which* data to deliver for guaranteeing high quality of data reconstruction, aggregation and linear regression analytics. We provide mathematical analyses based on the principles of optimal stopping theory, while evaluating and comparing the models performance with other methodologies following the

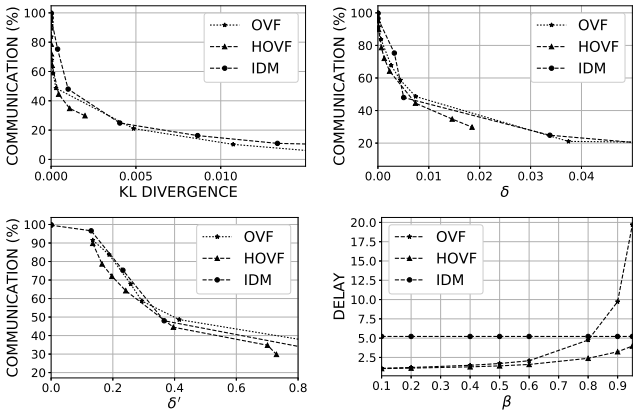


Figure 5: Trade-off for OVF, HOVF and IDM between (%) communication and (upper-left) KL divergence, (upper-right) regression discrepancy δ, (lower-left) model fitting δ'; (lower-right) expected delay vs. β ($\theta = 0.06$).

instantaneous decision making paradigm. Our approach is deemed appropriate to edge analytics being flexible to cope with the trade-off quality & communication overhead. Our future research agenda includes leveraging edge analytics by pushing predictive modeling & analytics to sensing/actuator devices expecting limited data transmission.

ACKNOWLEDGEMENT

This research is funded by the EU H2020 GNFUV Project/Action RAWFIE-OC2-EXP-SCI, under the EC Future Internet Research Experimentation (FIRE+) initiative.

REFERENCES

[1] M. Satyanarayanan, P. Simoens, Y. Xiao, P. Pillai, Z. Chen, K. Ha, W. Hu, and B. Amos, "Edge analytics in the internet of things," *IEEE Pervasive Computing*, vol. 14, no. 2, pp. 24–31, 2015.

[2] W. Tu, L. Wei, W. Hu, Z. Sheng, H. Nicanfar, X. Hu, E. C.-H. Ngai, and V. C. M. Leung.

[3] L. G. Rios *et al.*, "Big data infrastructure for analyzing data generated by wireless sensor networks," in *IEEE Big Data Congress*, 2014, pp. 816–823.

[4] A. Manjeshwar and D. P. Agrawal, "Teen: Arouting protocol for enhanced efficiency in wireless sensor networks," in *Proceedings of the 15th International Parallel & Distributed Processing Symposium*. IEEE, 2001, p. 189.

[5] ——, "Apteen: A hybrid protocol for efficient routing and comprehensive information retrieval in wireless sensor networks," in *IEEE IPDPS '02*, 2002, pp. 48–.

[6] H. Jiang, S. Jin, and C. Wang, "Prediction or not? an energy-efficient framework for clustering-based data collection in wireless sensor networks," *IEEE Trans. on Parallel and Distributed Systems*, vol. 22, no. 6, pp. 1064–1071, 2011.

[7] N. Cheng, N. Lu, N. Zhang, T. Yang, X. S. Shen, and J. W. Mark, "Vehicle-assisted device-to-device data delivery for smart grid," *IEEE Trans. on Vehicular Technology*, vol. 65, no. 4, pp. 2325–2340, 2016.

[8] C. Anagnostopoulos, "Time-optimized contextual information forwarding in mobile sensor networks," *J. Parallel and Distributed Computing*, vol. 74, no. 5, pp. 2317–2332, 2014.

[9] G. Kamath, P. Agnihotri, M. Valero, K. Sarker, and W.-Z. Song, "Pushing analytics to the edge," in *IEEE GLOBECOM*, 2016, pp. 1–6.

[10] C. Anagnostopoulos, "Quality-optimized predictive analytics," *Applied Intelligence*, vol. 45, no. 4, pp. 1034–1046, 2016.

[11] M. Gabel, D. Keren, and A. Schuster, "Monitoring least squares models of distributed streams," in *21th ACM SIGKDD KDD*, 2015, pp. 319–328.

[12] N. Harth, C. Anagnostopoulos, and D. Pezaros, "Predictive intelligence to the edge: Impact on edge analytics," *Evolving Systems*, 2017.

[13] G. Peskir and A. Sirjaev, *Optimal stopping and free-boundary problems*. Birkhuser Basel, 2006.

[14] M. Dallachiesa, G. Jacques-Silva, B. Gedik, K.-L. Wu, and T. Palpanas, "Sliding windows over uncertain data streams," *KAIS*, vol. 45, no. 1, pp. 159–190, 2015.

[15] K. Patroumpas and T. Sellis, "Maintaining consistent results of continuous queries under diverse window specifications," *Information Systems*, vol. 36, no. 1, pp. 42–61, 2011.

[16] J. Gray, S. Chaudhuri, A. Bosworth, A. Layman, D. Reichart, M. Venkatrao, F. Pellow, and H. Pirahesh, "Data cube: A relational aggregation operator generalizing group-by, cross-tab, and sub-totals," *Data Mining & Knowledge Discovery*, vol. 1, no. 1, pp. 29–53, 1997.

[17] M. Kuhn and K. Johnson, *Applied predictive modeling*. Springer, 2013, vol. 26.

[18] H. Robbins, D. Sigmund, and Y. Chow, "Great expectations: the theory of optimal stopping," *Houghton-Nifflin*, vol. 7, pp. 631–640, 1971.

[19] B. W. Silverman, *Density estimation for statistics and data analysis*. Chapman and Hall, London, 1986, vol. 26.

[20] J. Durbin and S. J. Koopman, *Time series analysis by state space methods*. OUP Oxford, 2012, vol. 38.

[21] S. De Vito, E. Massera, M. Piga, L. Martinotto, and G. Di Francia, "On field calibration of an electronic nose for benzene estimation in an urban pollution monitoring scenario," *Sensors and Actuators B: Chemical*, vol. 129, no. 2, pp. 750–757, 2008.

[22] C. Tofallis, "A better measure of relative prediction accuracy for model selection and model estimation," *J. Operational Research Society*, vol. 66, no. 8, pp. 1352–1362, 2015.