applied sciences

MDPI

Article

Modeling and Profiling of Aggregated Industrial Network Traffic

Mehrzad Lavassani [1,2,*], Johan Åkerberg [1] and Mats Björkman [1]

[1] Division of Networked and Embedded Systems, Mälardalen University, 721 23 Västerås, Sweden; johan.akerberg@mdh.se (J.Å.); mats.bjorkman@mdh.se (M.B.)
[2] Division of Industrial Systems, RISE—Research Institutes of Sweden, 852 33 Sundsvall, Sweden
* Correspondence: mehrzad.lavassani@ri.se

Abstract: The industrial network infrastructures are transforming to a horizontal architecture to enable data availability for advanced applications and enhance flexibility for integrating new technologies. The uninterrupted operation of the legacy systems needs to be ensured by safeguarding their requirements in network configuration and resource management. Network traffic modeling is essential in understanding the ongoing communication for resource estimation and configuration management. The presented work proposes a two-step approach for modeling aggregated traffic classes of brownfield installation. It first detects the repeated work-cycles and then aims to identify the operational states to profile their characteristics. The performance and influence of the approach are evaluated and validated in two experimental setups with data collected from an industrial plant in operation. The comparative results show that the proposed method successfully captures the temporal and spatial dynamics of the network traffic for characterization of various communication states in the operational work-cycles.

Keywords: industrial network; aggregated traffic classes; traffic modeling

Citation: Lavassani, M.; Åkerberg, J.; Björkman, M. Modeling and Profiling of Aggregated Industrial Network Traffic. *Appl. Sci.* **2022**, *12*, 667. https://doi.org/10.3390/app12020667

Academic Editors: Wei-Chang Yeh, Xiao-Zhi Gao and Omprakash Kaiwartya

Received: 15 December 2021
Accepted: 6 January 2022
Published: 11 January 2022

1. Introduction

The concept of Industrial IoT encompasses the joint applicability of operation, internet, and information technologies to expand the efficiency expectation of automation to green and flexible processes with innovative products and services. A requirement for ensuing this integration is the transformation of industrial network infrastructures to enable the accommodation of new traffic from different technologies. This transformation is step-wise and needs many considerations to ensure the successful development of future industrial networks. One essential consideration is to ensure the continuous operation of the existing system, the machinery, and infrastructure also known as brownfield installation, to avoid risk of downtime.

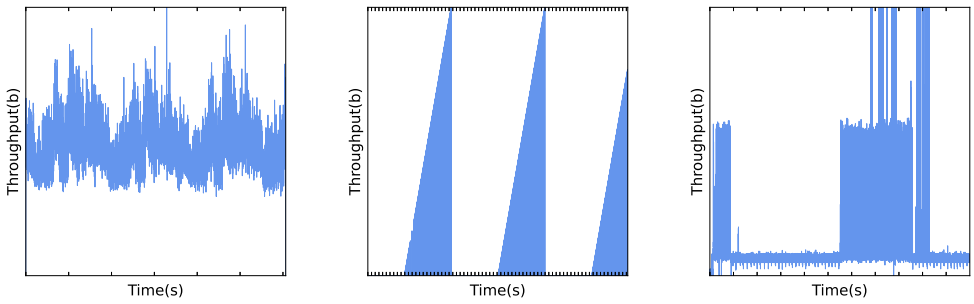
The importance and benefits of consolidated networks have been discussed from various aspects, and their challenges have been addressed from different technical perspectives [1–3]. One dominant challenge for industrial networks to overcome is satisfying the diverse and, in some cases, contradictory requirements of the Internet technology (IT) and operational technology (OT) systems, like real-time performance and high throughput. Time-Sensitive Networking (TSN) [4] provides a toolbox to provide mechanisms for any possible traffic type that are predicted to coexist in the future industrial networks.

The Orchestration of various applications in industrial ecosystems, with heterogeneous industrial communication protocols such as Open Platform Communication (OPC), MTConnect and message queue telemetry transport (MQTT), raises as a significant obstacle for system integration [5]. The works more concerned with the interoperability of different systems, by providing a communication middleware satisfying control systems requirements, are mostly presented in Open Platform Communications Unified Architecture (OPC-UA) [5,6].

The future industrial networks pose demanding requirements, and in some cases unpredictable challenges, on network infrastructure. One of the impacted domains by these challenges is network resource management, where integration of new technologies and applications into the existing networks translate to scaling-up communication and new configuration with no negative impacts on the performance of the ongoing processes. Recently, the academic and industrial communities started the conversation on the importance of studying brownfield installations and the need to support the legacy systems to guarantee the performance when transforming industrial networks [3,7]. In [3] the challenges in automation for future industrial networks are detailed. It is concluded that additional insights from the traffic of existing installation are prerequisites for the integration of new technologies and providing the intermediate steps required for the evolutionary transformation of the industrial automation networks. The additional insight from the state of ongoing processes can identify the current resource plan of the ongoing network communication in terms of bandwidth, dynamic throughput, delivery times, and scheduled traffics. This insight can contribute to the development of integration strategies that ensure adequate outcomes based on predefined performance metrics in the complex integrated networks for network management tasks such as provisioning and dimensioning [8–10].

In other words, resource management with respect to the new integration characteristic and required resources, while safeguarding the performance of legacy systems and ongoing processes. Hereof, a clear view of the operational state of the existing network is an undeniable prerequisite. However, this prerequisite is not a trivial goal to achieve due to the unavailability of data from industrial plants and the complexity of their traffic.

The unavailability of data from industrial plants is a major factor limiting the application of promising methods for developing solutions that address the existing and foreseeable future challenges for network management. Consequently, the research works that chose the next best solution of developing methods using the lab or simulated data do not necessarily obtain the expected results when applied to the traffic collected from brownfield. Figure 1 illustrates examples of traffic collected from (a) a paper mill operational network [3], (b) an SCADA system testbed for wind turbine [11], and (c) a SCADA lab with industrial equipment [12]. It is evident that the traffic patterns and their complexity differ vastly even in small time intervals between the three sources of data.



(**a**) Paper mill (**b**) Wind turbine testbed (**c**) SCADA lab

Figure 1. The complexity of the communication patterns from three different sources is visualized. Even in smaller time intervals, the aggregated traffic of the production line is more complex, while also showing the operational work-cycles.

The reason for the complexity of the traffic collected from brownfield is twofold: overly tailored networks to fulfill the performance guarantees of a specific application or use-case, and long life span of the industrial systems where integration of new applications resulting in significant differences between the current states from the initial, and theory-supported, configuration. The evident consequences of this complexity are the high cost of network reconfiguration with flow-based measurement and modeling, and divergence from the theoretical class-based assumptions when modeling the aggregated traffic classes [10].

1.1. Related Work

Several recent studies tried to tackle the challenges associated with brownfield traffic from various angles. Flow-based modeling is proposed in [13] to profile communication patterns in industrial IP networks for intrusion detection. However, the flow-based modeling proved unpractical because of the complex communication patterns. In [14] inter-arrival time and correlation models were defined as additional key parameters for traffic characterization and flow-based modeling in SCADA networks. The method showed promising results for anomaly detection on lab-generated data [15], but applying on real traffic, the performance was unsatisfactory [16]. Four on-off-based models were proposed in [8] to model the communication flows. The application of this approach was discussed in relation to different traffic classes in industrial networks.

In [9] a methodology is proposed for evaluating the availability of resources for new traffic integration for network configuration management. Inter-arrival time and packet size were the measured parameters for flow-based resource estimation.

A case study [10] highlighted the disparity between the common characteristics and modeling assumptions based on traffic classes and those seen in the data collected from brownfield with aggregated traffic classes. Ref. [17] further emphasized this finding by experimental evidence from three case studies for 5G systems.

Addressing the complexity of provisioning and configuration of the scaled-up IT-OT networks, and inaccurate results from the commonly applied traffic measurement for characterization and class-based model assumptions, a new measurement with the scope on network total traffic is introduced in [10]. The measurement applied on an aggregated traffic showed the mirroring bandwidth consumption patterns related to the operational work-cycles.

1.2. Proposed Approach

Despite the valuable contributions, there is still a palpable gap in the literature on topics relevant to studying the aggregated traffic classes of brownfields to enable the next step of integration with new technologies. While network performance metrics are considered for developing methodologies for new traffic types integration, there are still no key parameters for characterization or measuring the performance of the consolidated networks. The existing related work, study industrial network communication for means of network traffic modeling. However, there are still unaddressed issues in studying the dynamic behavior of networks that call for new and continuous probes and assessments. There is no proper model that considers network traffic as a whole; the focus so far has been on the characterization and modeling of each traffic flows in the network. Indeed this approach poses difficulties and complexities to network resource management considering the IIoT application demands in terms of scalability and flexibility. While the existing work-cycles in the traffic patterns, due to the operational work-flows, are acknowledged, it is not incorporated into solution development for modeling or characterization of network traffic.

This paper addresses the support and securing of the performance of brownfield and avoiding performance bottlenecks in new integration, considering the challenge of future industrial networks for network resource management. The main goal is to model the network traffic dynamics to identify various communication states and profile their characteristics. The identified states and the profiled characteristics then can create the bases of network monitoring, and provide the required insight in decision making for provisioning or scaling up by safeguarding the requirements of the ongoing traffic in the network. The contributions of this work are as follows:

- Two network parameter indicators (PIs), transmission volatility and transmitter volatility, are introduced to capture the temporal and spatial dynamics in bandwidth utilization to formulate the communication intensity and identify the work-cycles.
- Work-cycles are modeled, and their communication states are profiled based on their statistical summary and dynamic characteristics, including the introduced parameters.

- A two-step approach is proposed to model the aggregated traffic classes collected from brownfield, with respect to transmission intensity and throughput, utilizing the introduced parameters. The proposed approach is validated through comparative analysis of its performance in terms of accuracy of prediction and consistency of generated labels, with a set of unsupervised learning algorithms.

2. Network Traffic Modelling

Integrating IT and OT traffic introduces new challenges for network management in complexity, scalability, and analysis accuracy, either for replacement or integration of new technologies. These challenges have a direct impact on network performance and configuration management.

2.1. Network Traffic Measurement

The essential step in network management for performance and configuration is understanding the dynamics of the existing traffic in the network, i.e., modeling the network traffic. The modelling usually builds on the characterization of network flows and profiling traffic types for analyzing the requirements as well as guaranteeing intended performance criteria in the provisioning and dimensioning of the network.

Flow-based measurement for network configuration management lacks the required scalability due to the complexity of communication patterns between devices. Moreover, the common assumption on single-type traffic for modeling traffic types does not hold for traffic collected from brownfield since various types of traffic can be generated from the same sources [10,17]. To overcome the challenges of flow-based characterization and remove the assumption of single-type traffic, an approach is proposed in [10] that sets the measurement scope on the network level, i.e., accumulated traffic in a specific time interval. We adopt the same measurement for outlining the network traffic modeling. While the network level measurement can reduce the analysis complexity and reveal the repetition of communication patterns correspondence to the system work-cycle, it also affirms enough variances to exclude the deterministic traffic modeling of the communication system.

In general, the generated traffic in the network consists of periodic, sporadic, and burst traffic types [18], which means in any time interval we have a mixture of the traffic of some or all of these types. Since not all the involved parameters are known, e.g., sporadic, irregular IT traffics, various size payloads, and a different number of transmissions, the accumulated traffic in any specific time intervals cannot be considered deterministic.

To characterize the traffic dynamics using throughput and bandwidth consumption, the uncertainties need to be parametrized. In flow-based characterization, part of the uncertainties can be parameterized through packet generation intensity parameter. Each data stream and generated packets from one source at different points in time is studied. In the network-level measurement, the parameterization of the intensity parameter needs to consider the accumulated transmissions in the network.

2.2. Network-Level Transmission Intensity

The variances of the throughput in any specific time frame are the direct result of the total number of transmitted packets and the sum of the transmitted data. Since the number of devices in the network and communication between network entities are predefined, it is valid to assume the throughput turbulence is mainly a variety of changes in the number of transmitting devices and the number of transmissions by these devices, in any time interval. In other words, for profiling different communication states of the network in each work cycle, the number of active transmitters, the number of transmissions by devices, and the amount of transmitted data need to be taken into account. We define the following parameters to capture and quantify the network-level transmission intensity (network-level dynamic).

2.2.1. Transmitter Volatility, *tsv*

It is defined to capture the dynamic behavior of the devices or transmitters. The number of devices in the network is constant, but not the number of active transmitters in each time frame. Previous studies also showed that the number of transmissions by each device in various time frames during each work cycle differs [3,10].

Therefore, the throughput contribution of each transmitter based on the number of transmissions, or generated packets, in each time frame, is one element for quantifying the network dynamically. Since the scope is on the network level, *tsv* is defined as a relative change of transmitter's behavior over time in terms of the number of transmissions. Consequently, the dynamic of throughput within each time frame can be studied relative to *tsv*. That is, studying the relation between throughput and transmitter changes where the varying amount of data transmitted by the same device, due to the number and size of the transmitted packets, can be accounted for.

2.2.2. Transmission Volatility, *txv*

On the network-level *txv* is defined to capture the throughput dynamic over time by accounting for the varying number of transmissions. In each time frame the *txv* is the relative impact of the number of transmitted packets on the total throughput in the network.

Utilizing the two defined parameters, the dynamic behavior of the active elements in the network can be quantified by considering all the alternating elements, i.e., number of the active transmitter, number of transmissions, and the size of the transmitted packets, without requirements of per-device specification. The network-level transmission intensity, *ti* can then be formulated to capture the dynamic behavior of the devices as:

$$ti = tv_{s,x} = tsv \times txv \tag{1}$$

where *txv* represents the number of transmissions, *tx*, at each time instance, and *tsv* captures the number of active transmitters, *uts*.

2.3. Proposed Traffic Modelling Approach

The main goal of the presented work is to model the network traffic dynamics to identify various communication states and profile their characteristics. For this purpose, we propose a two-step approach, Figure 2, and detail the methodology in the following.

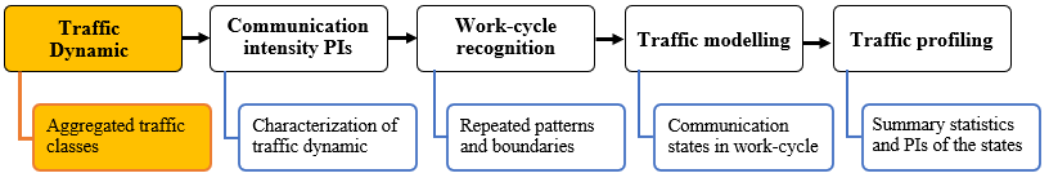


Figure 2. The proposed two-step approach overview; it builds on the network-level modeling method proposed in [10].

Previous studies and literature indicate the projection of the operational work-cycles in the repeated transmission and throughput patterns of the aggregated traffic [10,14]. It is reasonable to deduce that the repeated patterns provide similar information. That is a large dimension without any significant information gain, which can result in inaccurate modeling due to overfitting. Therefore, the proposed approach aims to exclude similar information and then to model the network traffic and profile the communication states. Accordingly, the goal of modeling the network traffic dynamics can be achieved by defining the following objectives:

1. Recognizing repeated communication patterns to identify work-cycles with respect to bandwidth consumption and communication intensity.
2. Profiling distinguishable transmission states for identifying the possible state-space in each work-cycle.

The two objectives are closely linked as the first objective serves as a prerequisite of the second objective. The state-space of the model can be reduced by identifying the work-cycles boundaries and consequently eliminating the dynamic imposed by insignificant and inconsequential variances in different work-cycles.

The following section details the proposed approach and presents the results for the realization of the two objectives on the data collected from brownfield with aggregated traffic classes.

3. Modelling Aggregated Traffic Break-Down

In this work, the aggregated traffic is modeled by realizing two interlinked objectives: work-cycle recognition and transmission states identification and profiling. The s taken to achieve the.

We first start with formulating the problem. Let $Y_{1:T} = \{y_1, y_2, \ldots, y_T\}$ be an observed stream of data generated in the network at time $t = 1, 2, \ldots T$, where each y_t is the joint reading of all flows, $y_t \subset R^n$. The recorded data from various streams in a specific time interval can be considered as a $(x \times t)$ matrix with $x = 1, 2, \ldots X$ be the data streams contributing to the y_t. The case of bandwidth utilization, with l_{tx} indicating the payload of each stream at each point, can be formulated as:

$$y_t = \sum_{t=1}^{T} \sum_{x=1}^{X} l_{tx} \tag{2}$$

3.1. Objective 1: Work-Cycle Recognition

Our first objective is to identify the work-cycles to find boundaries of the repeated communication pattern within a time interval. Algorithm 1 describes the steps for achieving the first objective. As mentioned in the previous section, the throughput turbulence results from variances in the number of active devices, the number of transmissions, and the size of transmitted packets. Considering network-level throughput (2), we quantify the throughput, tp_{tot} rate of change, roc_{tp}, dependent on transmission intensity, as

$$roc_{tp} = \frac{tp_{tot}}{tsv \times txv} = \frac{\sum_{t=1}^{T} \sum_{x=1}^{X} l_{tx}}{\sum_{t=1}^{T} uts_t \times tx_t} \tag{3}$$

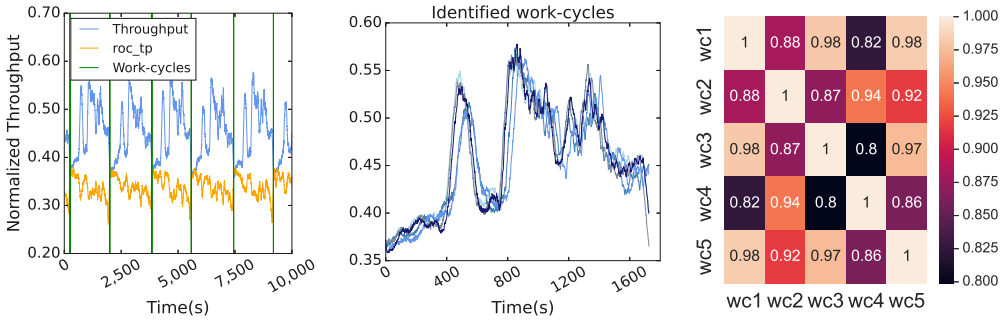
where uts_t is the number of active transmitters or devices, and tx_t is the total transmitted packets in the specific time interval. The largest changes happen when the accumulated impact of the combined parameters are the highest. The work-cycle identification is then defined as a Change Point Detection (CPD) problem where the most significant abrupt change indicates the boundaries of the repeated pattern in the dynamic turbulence sequence. The beginning and the end of each work-cycle can be identified by the abrupt decrease of the throughput, i.e., the time interval between two consecutive highest rank change points.

Applying Equation (3) on the data stream generates a number of non-overlapping time windows that segments the data into work-cycles. Figure 3a illustrates the outcome of work-cycle identification utilizing the throughput dynamics dependent on the proposed transmission and transmitter volatility parameters. Figure 3b shows the segmented data with new indexing for mapping the patterns, prerequisites for identifying the various operational mods/states of the aggregated traffic modeling. The accuracy of the method is checked using Spearman correlation and represented in Figure 3c. The similarity between the patterns varies between 0.88 to 0.98. The high similarity of the data sequences in each segment shows the success of the work-cycle recognition process utilizing the defined parameter indicators to capture transmission intensity.

Algorithm 1 Work-cycle recognition.

Data: Network traffic data, $Y_{1:T}$.
Results: Approximated work-cycle boundaries.

Set:
counter $n, i = 0$, $list_{roc}$, temp_threshold $= \Delta t (= 60)$, spatial_threshold $= \Delta s (= 0.1)$.
for $t \leftarrow t : 1 : T$ **do**
    Calculate total throughput: $tp_{tot} = \sum_1^X l_x$.
    if $x \neq 0$ **then**
        Calculate transmitted packets: $txv \leftarrow \sum_1^X tx_x$.
        Calculate active devices: $tsv \leftarrow n++$.
    end if
    Calculate transmission intensity: $ti = tsv \times txv$.
    Calculate throughput rate of change: $roc_t = tp_{tot}/tsv \times txv$.
    Append $list_{roc}[t] = roc_t$.
end for
Rank roc_t in $list_{roc}$ where $roc_t - ti_t < \Delta s$.
Find the change points: $List\_CP[i], i \leftarrow 1 : m < |Y|$.
$x_{init} = list_{roc}[1]$, $interval = 0$.
for x in $list_{roc}[t], t \leftarrow 1 : T - 1$ **do**
    $interval++$.
    if $(\overrightarrow{x_{init}x_{t+1}} \neq \overrightarrow{x_{t+1}x_{t+2}})$ & $(interval > \Delta t)$ & $(roc_t < \Delta s)$ **then**
        Append $list\_CP \leftarrow list_{roc}[t]$.
        Update $x_{init} = list_{roc}[t+1]$
    end if
end for



(**a**) (**b**) (**c**)

Figure 3. The work-cycles recognition is realised through: deploying communication intensity parameters to find work-cycles boundaries (**a**); the non-overlapping segments resemble similar repeated patterns (**b**); the Spearman's correlation shows high degree of similarity between the identified work-cycles (**c**).

3.2. Objective 2: State Modelling and Profiling

The second objective is to identify the state space the communication in the network can obtain, i.e., identifying different states in which the multi-modal communication is functioning. That is, we want to partition the time into k consecutive and non-overlapping segments $\{t_{1:k}, s_i\}$ where t_k represents k^{th} segment of time with state s_i, $i = 1, \ldots, m$, that ends at time t_k. The data sequences in each segment should show similar dynamics and characteristics while dissimilar enough from the other segments to distinguish the states.

The only available ground truth and previous knowledge about the network are the evident work-cycles in traffic patterns, obtained from the first objective and supported in the literature [9,10,14]. Hence, the number, duration, or characteristics of each of the states are unknown, i.e., there exists no knowledge about the true model. The parameters that can describe the underlying communication pattern need to be learned from the limited available data for modeling the network traffic.

Furthermore, the available network data is unlabelled and lacks any additional information about the possible or existing operational states for deterministic network modeling.

In the absence of the ground truth for the operational modes evident in the network traffic patterns, the modes need to be approximated from the data. In other words, the data sequences with similar characteristics that can describe the dynamic characteristics of each state need to be identified.

The hidden Markov model (HMM) is an effective unsupervised method with strong support in Bayesian inference that has been proven effective applied on sequential data where the correct model, the order of the HMM, is unknown. HMM can represent the probability distributions over a sequence of observations and model the observations as a probabilistic function of the latent states. In compact notation an HMM can be defined as $\lambda = (A, B, \pi)$, where $\pi = \{\pi_j\}$ is the initial state distribution, $A = \{a_{ji}\}$ is the state transition probability, and $B = \{b_i\}$, $1 \leq j, i \leq m$ is the probability of the observation in the current hidden state.

To deploy HMM for modeling the network traffic, one important parameter to estimate is the order of HMM, i.e., the number of states that best describes the data.

3.2.1. Model Selection

Model selection has been one of the main concerns in deploying learning algorithms in real scenarios. Dynamic systems are hard to be deterministically modeled as there are many factors involved in the problem, from system parameters to the effect of the surrounding environment on system behavior. Furthermore, the process of identifying the complex relation and correlation between all parameters, as well as the element of noise, can be costly and time-consuming, if not impossible. In data-driven system identification, a learning algorithm is trained to uncover system model over recorded historical data, and then deployed for various configuration and provisioning management, diagnostic and prognostic purposes [19–23]. The question here is how to select a learning algorithm in the absence of an identified true model? To answer this question, we first need to find models with various orders that give the best approximation, or fit, for the targeted data.

There exist many methods in the literature for comparing models accuracy for various datasets and data types with different characteristics [24,25]. The majority of the methods are based on likelihood model selection where the model parameters, such as the number of states and samples, are not considered. Thus, increasing the number of states leads to a higher likelihood adds to system complexity without providing additional information [24]. Therefore, methods that consider the number of model parameters are desirable. Commonly applied methods for order estimation of sequential data with parameter consideration are Akaike's information criterion (AIC) [26] and Bayesian information criterion (BIC) [27] and efficient determination criterion (EDC) [28].

Akaike information criterion is one mathematically supported evaluation criterion of models. It is an estimator of expected relative (K-L) information based on the maximized log-likelihood function:

$$AIC = -2\log(\hat{L}) + 2k \tag{4}$$

where k is the number of estimated parameters in the approximated model, and $\hat{L}$ is the maximum likelihood of the model with the true order k. For small sample size where $\frac{n}{k} \lesssim 40$, AIC becomes:

$$AIC_c = -2\log(\hat{L}) + 2k + \frac{2k(k+1)}{(n-k-1)} \tag{5}$$

where n is the cardinality of the set, that is the number of elements in the sample set.

Baysian information criterion is closely related to AIC model selection methods, but introduces the sample size in the penalty term that provides

$$BIC = -2\log(\hat{L}) + k\log n \tag{6}$$

Efficient determination criterion encompasses *AIC* and *BIC* and introduces a strictly increasing function that results in a strongly consistent order estimation.

$$EDC = -2\log(\hat{L}) + k\log\log n \tag{7}$$

As all assumptions are excluded from the model selection process, in this work, the order of the model is set where the difference between the three methods are minimized. The identified work-cycles from the first objectives are the data used for state identification. The throughput in each work-cycle is influenced by the same conditions with minor differences, i.e., varying transmission intensity posed by unknown parameters.

The results of applying the three methods for model selection are presented in Figure 4a. There are two possible choices of orders: 2 and 4, with the second-order shared between all the three methods. However, the results of the fourth-order are less consistent between the methods; the model order by EDC maps those by AIC and BIC for the order of 2, but it shows signs of overfitting with higher orders, i.e., 4. Therefore, the model selection process suggests two different states in the communication dynamics within each work-cycle. A second-order HMM model identified the various states and transition probabilities between each state. Figure 4b shows the mapping of the identified states on the work-cycle data. The colour-coded identified states indicate consideration of both temporal and spatial features. The data sequences identified for the two states show high similarities, while those belonging to different states resemble dissimilarities. The Pearson correlation of the two states' data on average for different work-cycles is 0.58. Considering the value range of 0 to 1, with the highest positive correlation at 1 and the highest negative correlation at 0, the acquired correlation value indicates a very small correlated behavior of the states. The distribution of the data points also supports dissimilarities of the identified states, Figure 4c.

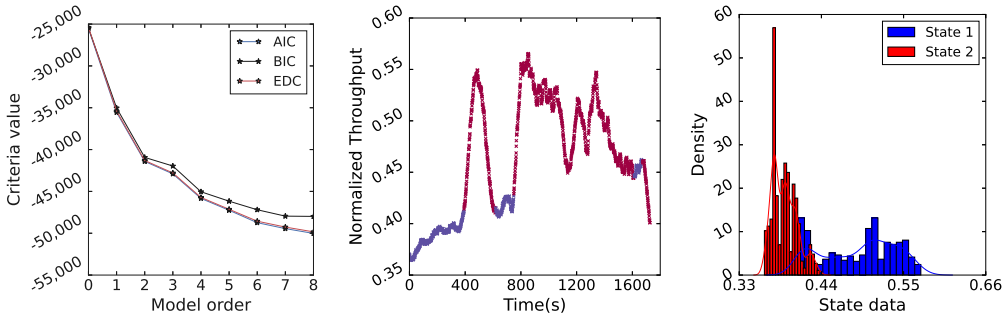


Figure 4. The criterion values, y-axis, suggest a second order HMM for the dataset consisting of 5 work-cycles (**a**); the identified states by the second-order model are color-coded in (**b**); (**c**) shows distribution of the data in each state.

3.2.2. State Profiling

After identifying the states, the second objective is achieved by profiling distinguishable transmission states in a work-cycle. A set of labels was produced in the modeling process to identify the data points belonging to each state. The summary statistics of the identified states are presented in Table 1. Since the transition between the states is time-varying, each state is profiled by the summary statistics and the communication intensity parameters, i.e., transition volatility and transmission volatility. Each varying length state can be profiled as

$$\{(t_{1:k}, s_i)|s_i = \{\mu_i, \sigma_i, M_i, max_i, uts_i, tx_i\}, i = 1, 2\}.$$

Table 1. The summary statistic of the identified states.

	Mean-μ	**Std-σ**	**Max**	**Median (50%)-M**
State 1	0.412	0.030	0.470	0.402
State 2	0.501	0.035	0.565	0.507

At this point, the profiles can be used as the basis for identifying the states, given a sequence of observations in a work cycle. The steps taken to achieve the second objective are described in Algorithm 2.

Algorithm 2 State modeling and profiling.

Data: Work-cycles dataset, wcd.
Results: Profiled states.

Initialize model parameters:
Model order: $model\_k, k \leftarrow 1:10$,
Sample size: $n \leftarrow |wcd|$,
Model evaluation criterion: $model\_ec = [AIC, BIC, EDC]$.
for $model\_k$ & n **do**
    Fit and predict probability: $model\_k(wcd, n)$.
    Calculate likelihood: $\hat{L}(wcd)$.
    Calculate model accuracy score: $model\_score_k(wcd)$.
    for $model\_ec$ **do**
        Calculate evaluation criterion: $model\_ec(n, k, model\_score_k(wcd))$.
        Append to EC_List: $EC\_List[k, AIC, BIC, EDC] \leftarrow model\_ec()$.
    end for
end for
Select $model\_k$ where $EC\_List[k, AIC_{min}, BIC_{min}, EDC_{min}]$.
$model = HMM(k, wcd)$.
$state_{1:k} = model.predict(wcd)$.
for $state_i[items], i \leftarrow 1:k$ & $items = wcd[state_i]$ **do**.
    Calculate $\mu_i, \sigma_i, M_i, max_i, uts_i, tx_i$.
    Profile $s_i = \{\mu_i, \sigma_i, M_i, max_i, uts_i, tx_i\}$.
end for

To provide support for the effectiveness of the proposed approach, including the introduced parameters and deployed learning model, a set of experiments were carried out where the results are presented in the following section.

4. Results and Discussion

This section presents the comparative results for validating the proposed network modeling approach in capturing the temporal and spatial communication dynamic. We first introduce two scenarios in which we carry out the comparative study, including the experiment setup and evaluation metrics for performance analysis. Further, we discuss the results and their importance in network management with respect to the continuous and uninterrupted operation of the network while enabling the evolution of OT systems.

4.1. Data Collection and Dataset

The data was captured from the Iggesund paperboard factory. It is a typical process automation factory, and the network can be considered as an example of production networks in manufacturing. The communication between different systems is provided by the configuration of several virtual LANs (VLANs). The experiments of this study cover a part of the operational network consisting of 5 control systems, with 43 stations connected to the server network and 32 process controllers on various VLANs; 337 devices in total [3,10].

The network traffic was captured by enabling mirroring of the traffic recorder port connected to one of the switches in the production network. The initial captured traffic

flows are huge files containing packet dumps from the network, in *.pcap* format, and consist of both IT and OT traffic. The resolution of the collected data is in microseconds with varying communication intervals from milliseconds to seconds. To illustrate the results of this experiment, 6,554,498 consecutive recorded transmissions were selected to cover several operational cycles.

4.2. Validation Scenarios

In the proposed approach, the introduced parameters were used to identify the work-cycles and reduce the state-space for higher accuracy of state recognition. An HMM with an order of 2 was deployed to capture the spatial dynamics within each work-cycle. The impact of the proposed two-step phase identification approach is validated by comparing the results with (1) a one-step work-cycle and state identification, and (2) a two-step identification with a set of learning algorithms.

4.2.1. One-Step Approach and Work-Cycles

This scenario aims to validate the proposed two-step approach and the impact of the work-cycle recognition as a prerequisite for network modeling. For this purpose, this scenario investigates the accuracy of the learning algorithms in capturing the spatial dynamics in the system for work-cycle and mode/state identification. The proposed parameters are excluded from this experiment, and a set of unsupervised learning algorithms are applied directly on the original dataset, containing data of 5 work-cycle duration.

The choice of unsupervised clustering algorithms is due to the lack of labels for the data and removed assumption of previous knowledge about work-cycles profiles and dynamic characteristics. The specific algorithms are selected to cover various possible clustering approaches, i.e., distance-based KM, hierarchical AC, and graph-based SC methods. The set of algorithms includes HMM, K-means (KM), Agglomerative clustering (AC), and Spectral clustering (SC). The model with an order of 5 and 2 were applied for (1) identifying the 5 work-cycles and (2) identifying the 2 states in each of the work-cycles.

4.2.2. Two-Step Approach and Communication Mode

This scenario aims to show the efficiency of HMM in state-mode identification and capturing the temporal and spatial dynamics, compared to a set of unsupervised learning algorithms. In this setup, the introduced parameters were first applied for identifying the work cycles. Then the new dataset was fitted to the same set of learning algorithms of the first scenario to identify the states/modes in each work-cycle. The results were labels for the data, which then were fitted to a logistic regression model to evaluate the prediction consistency of the algorithm set. The accuracy was evaluated by classification reports in terms of precision, recall, and f1-score of the predicted labels.

4.3. Comparative Results

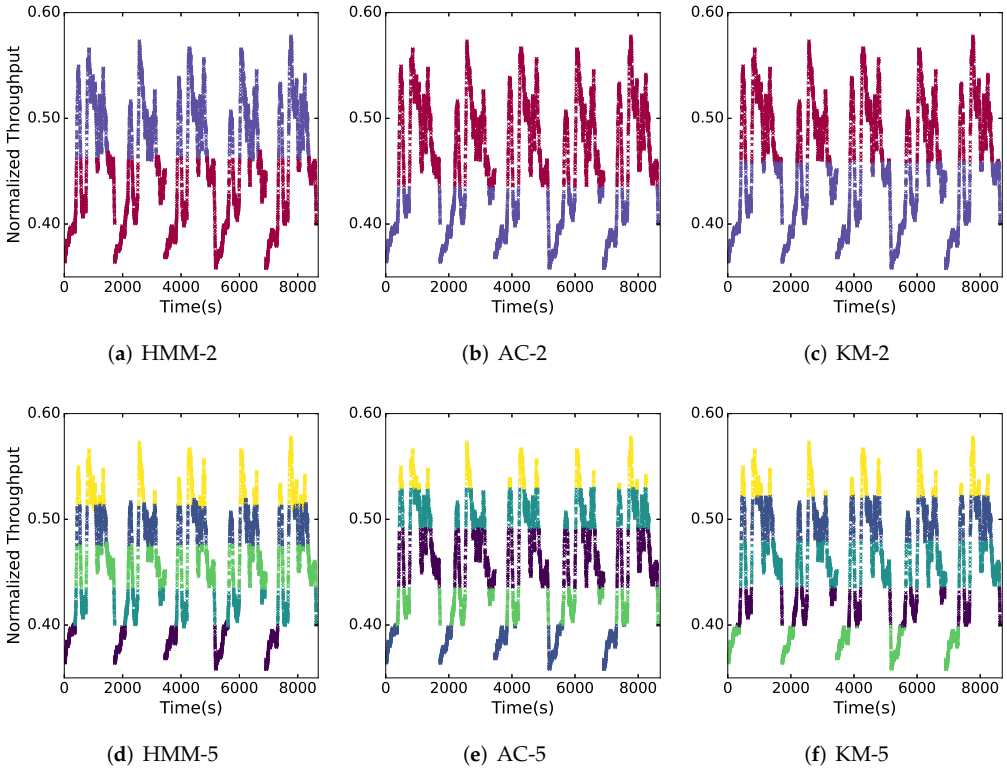
In what follows, we present and discuss the results of the introduced scenarios to validate the proposed two-step approach and the selected modeling tool for state/mode identification in communication work-cycles.

4.3.1. One-Step Approach and Work-Cycles

This phase of the experiment is carried out to demonstrate the importance of work-cycle identification as a prerequisite for communication state/mode modeling. Among the algorithms in the comparison set, K-Means and Agglomerative clustering can be deployed on the dataset without any assumptions on the expected number of clusters, i.e., the number of work-cycles. Preliminary, the data were fitted to this subset of unsupervised learning algorithms with the addition of MeanShift clustering. The identified number of clusters by all the algorithms on the original dataset was 2.

The results of fitting the data, without the work-cycle identification step, to the set of algorithms with the order of 2 and 5 are illustrated in Figure 5. It is clear that none of the

learning algorithms succeeds in capturing the work-cycles or the operational mods. With a higher-order model, all the algorithms tend to set a higher number of spatial thresholds and miss the temporal dynamics of the system altogether, Figure 5d–f. The performance differences are limited to the threshold values, with a variance of 0.05 on the normalized filtered throughput values for the second-order model, and 0.04 for the fifth-order model.



(**a**) HMM-2 (**b**) AC-2 (**c**) KM-2

(**d**) HMM-5 (**e**) AC-5 (**f**) KM-5

Figure 5. The result of one-step approach for state-mode identification on the dataset with no labels for work-cycles. Each state is distinguished by a different color. The first row shows the second-order model identification. The second row shows the results of the fifth-order model. The algorithms fail to capture the temporal dynamics of the signal where work-cycles are not identified and increasing the order only adds to the number of spatial thresholds.

The algorithms do not detect the repeated temporal pattern since the increased amount of data provides more support for the distance and/or similarity between the values. While the data is clustered and the learning process shows acceptable results for algorithms without any parameter tuning, the prediction accuracy of the algorithms on the unseen data is very low; even though the unseen data bears high similarity with the training dataset, as illustrated in Figure 3b. One reason for this poor performance is that the algorithms do not take the order of the observed data into account. The commonly used unsupervised and clustering algorithms work based on maximizing the intracluster while minimizing the intercluster similarities. Therefore, two points with similar values close to a cluster center are more likely to be assigned to the same cluster despite their time of occurrence or order.

As discussed previously, throughput varies in different time intervals due to the dynamic behavior of the devices. Hence, it is accessible that sole spatial profiling does not provide sufficient information for estimating the required resource and planning for an ongoing process that expands in time.

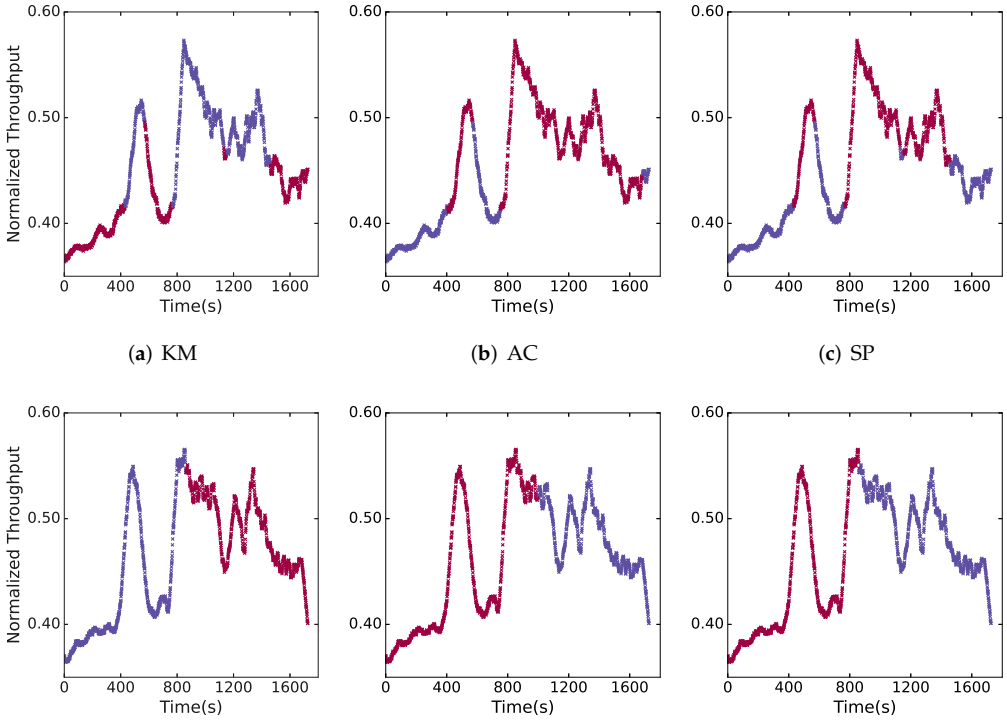
4.3.2. Two-Step Approach and Communication Mode

Table 2 details the results of the classification of the data into the two states utilizing labels generated by the algorithms set. Prediction of labels on the data is shown in Figure 6a–c. The classification results show high accuracy for all the algorithms, but visu-

alizing the prediction outcomes does not inspire identified states that can be utilized for profiling the communication intensity in each work-cycle. The summary statistics of the clusters provide negligible differences for being used as a parameter with high certainty for the classification of the states. In the specific case of SP, Figure 6c, the identified states show more of the temporal thresholds. Applying HMM for state identification, Figure 4b, is superior compared to the other algorithms in the algorithms set in terms of capturing the dynamic of communication intensity. HMM identifies the states based on the temporal and spatial features of the data, which provides distinguishable statistical summaries for profiling and modeling various dynamic behavior in each work-cycle.

Table 2. The prediction accuracy of the data is labeled by different algorithms.

Algorithm	Precision	Recall	F1-Score	Accuracy
K-Means	0.88	0.81	0.81	0.81
Agglomerative	0.95	0.95	0.95	0.95
Spectral	0.90	0.81	0.83	0.85



(**a**) KM (**b**) AC (**c**) SP

(**d**) KM (**e**) AC (**f**) SP

Figure 6. The result of state-mode identification by the algorithms in the algorithms set on the dataset with labeled cycles. The algorithms fail to capture the spatial dynamic to provide a baseline summary statistic for the identified states, top row; and with ordered data, only the temporal thresholds are identified, bottom row.

To account for the temporal dependencies the data was ordered and then fit the algorithm. Figure 6d–f illustrates the results. In this setup, the algorithms provide temporal thresholds for the signal and neglect spatial features The temporal thresholds can be of interest, but the identified states do not show significant differences that can be the basis of comparison. In other words, the summary statistic of the features set such as mean, minimum, maximum, and variance values are not distinguishable with the certainty required for state classification. Adding the state duration as a feature does not solve this issue either, since the turbulence in each state calls for another round of learning for any meaningful traffic dynamic profiling.

4.4. Discussion on Results and Related Works

The overview of the proposed approach is illustrated in Figure 7. As presented through detailing the two-step approach in Section 3, it succeeds in identifying and profiling communication states in each work-cycle relevant to the operational mods. The two objectives of work-cycle recognition and state modeling serve the main goal of network traffic profiling. The defined network-level parameters, with respect to traffic dynamics, provided the basis for quantifying communication intensity. From the first objective, the continuous data stream was segmented into non-overlapping windows that approximated the boundaries of the work cycles. This process reduced the dimensions by excluding those with similar information, i.e., the repeated patterns. The result of the first objective was input for modeling the network traffic in each work-cycle. Since the data is sequentially observed, HMM was selected as an appropriate method for modeling the network dynamically. The order of the model was estimated using information gained from various methods, and a second-order HMM was used for the state identification step. The results showed dynamic similarities in the segments belonging to each state with both temporal and spatial features. The identified states then labeled the dataset. The summary statistics of each state, along with the defined parameter indicators, provided the basis for profiling network traffic dynamics based on the states of transmission intensity and throughput.

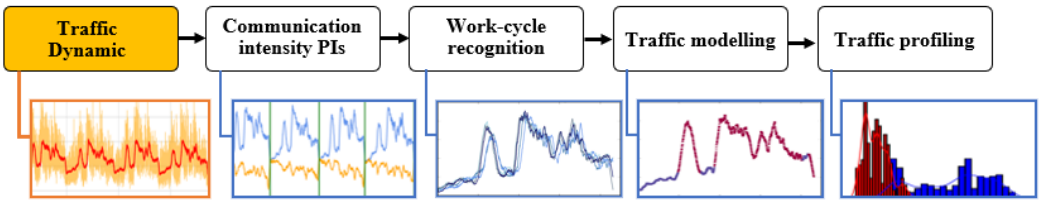


Figure 7. Overview of the proposed approach with the obtained results when applied on aggregated traffic classes data.

The effectiveness of this approach was demonstrated through the validation scenarios and comparative analysis of the results with a set of learning algorithms. The experimental results supported the importance of work-cycle identification as a prerequisite for traffic modeling when aggregated traffic is under study. Furthermore, the modeling process utilizing methods that take the order of the observations into account shows better performance than the tree or graph-based methods.

The industrial networks are under constant transformation either by integrating new technologies or adding advanced IT services to enhance flexibility, efficiency, and innovation. The former demands precise resource planning to avoid interruption of the ongoing operations, and the latter emphasizes data availability for additional insights into the system. In either case, imposed network changes require appropriate continuous adaptation, which is not a trivial task for network resource management. It calls for the full perception of the demands from the legacy systems and the new technology or the service, as well as a prediction of the consequences of their interactions in the ongoing network condition.

In the absence of sufficient research works addressing the lack of insight into brownfield installations and securing the performance of ongoing processes upon integration of new technologies and applications, the presented work undertakes the orderly steps to address the aforementioned requirements. The proposed approach extends the network-level characterization to aggregated traffic modeling for network resource management. The influence of this approach can be associated with the related works and the state of practice from various aspects. The identified work-cycle boundaries reduce the model's state-space while increasing the sample space for the learning process. The reduced dimensions by temporal segmentation remove the unnecessary complexity of stream modeling imposed by data points with insignificant information gain. It can potentially be combined with the modeling methods proposed in [8] to derive alternative solutions for communication state recognition, and improve the modeling results presented in [13,15] for more accurate

network traffic profiling. Furthermore, resource estimation for new technology or use-case integration can be carried out with more certainty since the accounted temporal dynamic gives more insight for provisioning periodic resource requirements, such as the presented method in [9].

A critical task in network resource management is to identify the possible communication bottlenecks [29–31]. The proposed method can be further deployed for network-level bottleneck estimation. The network-level bottleneck uncertainty parameter can be defined as the probability of bottlenecks with respect to the communication intensity in each state and the available resources. This parameter can be a baseline for provisioning based on resource estimation; states with lower communication intensity will have a lower bottleneck uncertainty level.

The future industrial networks are expected to accommodate both IT and OT traffic. In the proposed approach, the repeated patterns were utilized to model the traffic of OT systems. The self-similarity characteristic of IT traffic was detected and further discussed in the literature as one foundation for network modeling [32,33]. Considering both traffic bearing self-similar characteristics in their dynamics, a more efficient and realistic approximation of resources can be carried out by contemplating the consequence of their simultaneous accordance and possible interactions.

In this work, the first step is taken to address the identified gap within the active research and state of practice in network resource management, namely modeling and monitoring the traffic of industrial brownfield installation. It is indisputable that in the cross-section of computer science and operational technology more sophisticated methods can be developed with acceptable trade-offs between complexity and flexibility, as well as innovative solutions to address the existing gaps and challenges.

5. Conclusions

In the work presented, a two-step approach was proposed for modeling aggregated traffic classes of brownfield, and identifying various communication states to profile their characteristics with a network-level perspective. Two parameter indicators were defined to capture the transmission intensity of the traffic dynamic, and were deployed to identify the work-cycles in the streaming data. The operational states in each work-cycle were modeled with HMM as an effective method for sequential data. The importance of the first step and the method's performance for state identification was evaluated and validated in two experimental setups. The comparative results showed that the proposed method successfully captures the temporal and spatial dynamics of the network traffic for profiling various communication states in each work-cycle.

The impact of the approach was discussed in relation to network management challenges in future industrial networks and the related research in this field. Future work will expand the current modeling approach to online monitoring of communication dynamic and bottleneck prediction to optimize resource estimation and network resource management.

Author Contributions: Conceptualization, J.Å., M.B. and M.L.; methodology, experiments and visualization, M.L.; writing—original draft preparation, review and editing, M.L.; review, supervision, J.Å.; project administration, funding acquisition, J.Å. and M.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research was partly funded by the Future Industrial Networks project, grant number 2018-02196, within the Strategic innovation program for process industrial IT and Automation, PiiA, a joint program by Vinnova, Formas and Energimyndigheten.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Tian, S.; Hu, Y. The Role of OPC UA TSN in IT and OT Convergence. In Proceedings of the 2019 Chinese Automation Congress (CAC), Hangzhou, China, 22–24 November 2019; pp. 2272–2276.
2. Garimella, P.K. IT-OT Integration Challenges in Utilities. In Proceedings of the 2018 IEEE 3rd International Conference on Computing, Communication and Security (ICCCS), Kathmandu, Nepal, 25–27 October 2018; pp. 199–204.
3. Åkerberg, J.; Furunäs Åkesson, J.; Gade, J.; Vahabi, M.; Björkman, M.; Lavassani, M.; Nandkumar Gore, R.; Lindh, T.; Jiang, X. Future industrial networks in process automation: Goals, challenges, and future directions. *Appl. Sci.* **2021**, *11*, 3345. [CrossRef]
4. Time-Sensitive Networking Task Group. Available online: https://www.ieee802.org/1/pages/tsn.html (accessed on 25 November 2021).
5. González, I.; Calderón, A.J.; Figueiredo, J.; Sousa, J. A literature survey on open platform communications (OPC) applied to advanced industrial environments. *Electronics* **2019**, *8*, 510. [CrossRef]
6. Givehchi, O.; Landsdorf, K.; Simoens, P.; Colombo, A.W. Interoperability for industrial cyber-physical systems: An approach for legacy systems. *IEEE Trans. Ind. Inform.* **2017**, *13*, 3370–3378. [CrossRef]
7. Gooneratne, C.P.; Magana-Mora, A.; Otalvora, W.C.; Affleck, M.; Singh, P.; Zhan, G.D.; Moellendick, T.E. Drilling in the fourth industrial revolution—Vision and challenges. *IEEE Eng. Manag. Rev.* **2020**, *48*, 144–159. [CrossRef]
8. Weissenberg, M.; Głąbowski, M.; Hanczewski, S.; Stasiak, M.; Zwierzykowski, P.; Bai, V. Traffic Modeling in Industrial Ethernet Networks. *Int. J. Electron. Telecommun.* **2020**, *66*, 145–153.
9. Soós, G.; Ficzere, D.; Varga, P. Investigating the network traffic of Industry 4.0 applications—Methodology and initial results. In Proceedings of the 2020 16th International Conference on Network and Service Management (CNSM), Izmir, Turkey, 2–6 November 2020; pp. 1–6.
10. Lavassani, M.; Åkerberg, J.; Björkman, M. From brown-field to future industrial networks, a case study. *Appl. Sci.* **2021**, *11*, 3231. [CrossRef]
11. Teixeira, M.A.; Salman, T.; Zolanvari, M.; Jain, R.; Meskin, N.; Samaka, M. SCADA system testbed for cybersecurity research using machine learning approach. *Future Internet* **2018**, *10*, 76. [CrossRef]
12. 4SICS Geek Lounge Dataset. Available online: https://www.netresec.com/?page=PCAP4SICS (accessed on 25 May 2020).
13. Faisal, M.A.; Cardenas, A.A.; Wool, A. Profiling Communications in Industrial IP Networks: Model Complexity and Anomaly Detection. In *Security and Privacy Trends in the Industrial Internet of Things*; Springer: Cham, Switzerland, 2019; pp. 139–160.
14. Lin, C.Y.; Nadjm-Tehrani, S. Understanding IEC-60870-5-104 traffic patterns in SCADA networks. In Proceedings of the 4th ACM Workshop on Cyber-Physical System Security, Incheon, Korea, 4–8 June 2018; pp. 51–60.
15. Lin, C.Y.; Nadjm-Tehrani, S. Timing Patterns and Correlations in Spontaneous {SCADA} Traffic for Anomaly Detection. In Proceedings of the 22nd International Symposium on Research in Attacks, Intrusions and Defenses ({RAID} 2019), Beijing, China, 23–25 September 2019; pp. 73–88.
16. Lin, C.Y.; Nadjm-Tehrani, S. A Comparative Analysis of Emulated and Real IEC-104 Spontaneous Traffic in Power System Networks. In *International Workshop on Cyber-Physical Security for Critical Infrastructures Protection*; Springer: Cham, Switzerland, 2020.
17. Mogensen, R.S.; Rodriguez, I.; Berardinelli, G.; Pocovi, G.; Kolding, T. Empirical IIoT Data Traffic Analysis and Comparison to 3GPP 5G Models. In Proceedings of the 2021 IEEE 94th Vehicular Technology Conference (VTC2021-Fall), Norman, OK, USA, 27–30 September 2021; pp. 27–30.
18. Cao, Y.; Li, Y.; Liu, X.; Rehtanz, C. Modeling and simulation of data flow for vlan-based substation communication system. In *Cyber-Physical Energy and Power Systems*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 75–101.
19. Ni, J.; Yin, W.; Jiang, Y.; Zhao, J.; Hu, Y. Periodic Mining of Traffic Information in Industrial Control Networks. In *International Conference on Advanced Information Networking and Applications*; Springer: Cham, Switzerland, 2020; pp. 176–183.
20. Gómez, S.E.; Hernández-Callejo, L.; Martínez, B.C.; Sánchez-Esguevillas, A.J. Exploratory study on class imbalance and solutions for network traffic classification. *Neurocomputing* **2019**, *343*, 100–119. [CrossRef]
21. Jiang, Y.; Wang, W.; Zhang, C. Research on Traffic Recognition Algorithms for Industrial Control Networks based on Deep Learning. In *3rd International Conference on Mechatronics Engineering and Information Technology (ICMEIT 2019)*; Atlantis Press: Paris, France, 2019.
22. Wang, Q.; Chen, H.; Li, Y.; Vucetic, B. Recent advances in machine learning-based anomaly detection for industrial control networks. In Proceedings of the 2019 1st International Conference on Industrial Artificial Intelligence (IAI), Shenyang, China, 23–27 July 2019; pp. 1–6.
23. Byrén, F. Machine Learning for Traffic Classification in Industrial Environments, Master's Thesis, KTH, Stockholm, Sweden, 2018.
24. Burnham, K.P.; Anderson, D.R. Multimodel inference: Understanding AIC and BIC in model selection. *Sociol. Methods Res.* **2004**, *33*, 261–304. [CrossRef]
25. Dorea, C.C.; Resende, P.A.; Gonçalves, C.R. Comparing the Markov Order Estimators AIC, BIC and EDC. In *Transactions on Engineering Technologies*; Springer: Dordrecht, The Netherlands, 2015; pp. 41–54.
26. Akaike, H. *Akaike's Information Criterion*; Springer: Berlin/Heidelberg, Germany, 2011; p. 25.
27. Konishi, S.; Kitagawa, G. Bayesian information criteria. In *Information Criteria and Statistical Modeling*; Springer: New York, NY , USA, 2008; pp. 211–237.
28. Zhao, L.C.; Dorea, C.C.; Gonçalves, C.R. On determination of the order of a markov chain. *Stat. Inference Stoch. Process.* **2001**, *4*, 273–282. [CrossRef]

29. Johannesson, A.; Shams, P. Data-Driven and Variant-Based Throughput and Bottleneck Prediction Using Ensembled Machine Learning Algorithms. Master's Thesis, Chalmers University, Gotenborg, Sweden, 2018.
30. Thürer, M.; Ma, L.; Stevenson, M.; Roser, C. Bottleneck detection in high-variety make-to-Order shops with complex routings: An assessment by simulation. *Prod. Plan. Control.* **2021**, 1–12. [CrossRef]
31. Subramaniyan, M.; Skoogh, A.; Muhammad, A.S.; Bokrantz, J.; Johansson, B.; Roser, C. A generic hierarchical clustering approach for detecting bottlenecks in manufacturing. *J. Manuf. Syst.* **2020**, *55*, 143–158. [CrossRef]
32. Willinger, W.; Taqqu, M.S.; Leland, W.E.; Wilson, D.V. Self-similarity in high-speed packet traffic: Analysis and modeling of Ethernet traffic measurements. *Stat. Sci.* **1995**, *10*, 67–85. [CrossRef]
33. Melo, E.F.; de Oliveira, H.D.M. An Overview of Self-Similar Traffic: Its Implications in the Network Design. *arXiv* **2020**, arXiv:2005.02858.