# How does enterprise IoT traffic evolve? Real-world evidence from a Finnish operator[⋆]

B. Finley[a,*], J. Benseny[b], A. Vesselkov[c], J. Walia[b]

[a]*Department of Computer Science, University of Helsinki, Helsinki, Finland*
[b]*Department of Communications and Networking, Aalto University, Helsinki, Finland*
[c]*Department of Military Technology, National Defence University, Helsinki Finland*

**Abstract**

The adoption of Internet of Things (IoT) technologies in businesses is increasing and thus enterprise IoT (EIoT) is seemingly shifting from hype to reality. However, the actual use of EIoT over significant timescales has not been empirically analyzed. In other words, the reality remains unexplored. Furthermore, despite the variety of EIoT verticals, the use of IoT across vertical industries has not been compared. This paper uses a two-year EIoT dataset from a major Finnish mobile network operator to investigate device use across industries, cellular traffic patterns, and mobility patterns. We present a variety of novel findings: EIoT traffic volume per device has increased three-fold over the last two years, the share of LTE-enabled devices has remained low at around 2% and that 30% of EIoT devices are still 2G only, and there are order of magnitude differences between different industries' EIoT traffic and mobility. We also show that daily traffic can be clustered into only three patterns, differing mainly in the presence and timing of a peak hour. Beyond these descriptive results, modeling and forecasting is conducted for both traffic and mobility. We forecast the total daily EIoT traffic through a temporal regression model and achieve an error of about 15% over medium-term (30 to 180 day) horizons. We also model device mobility through a Markov mixture model and quantify the upper bound of predictability for device mobility.

*Keywords:* IoT, M2M, Empirical Measurements, Cellular Network
*2010 MSC:* 00-01, 99-00

## 1. Introduction

According to a recent survey [2], 29% of companies globally utilize Internet of Things (IoT) devices, suggesting that IoT use by companies (also referred to as enterprise IoT or EIOT) is shifting from hype to reality. Furthermore, EIoT is gaining momentum across different industries that use IoT devices in unique ways for solving diverse problems. Despite the variety of EIoT verticals and applications, the actual usage of IoT across industries has never been empirically explored. Additionally, the few studies [3, 4, 5] that have analyzed EIoT device usage from commercial cellular networks are relatively old and have only analyzed short timescales (i.e., typically less than a few weeks). Thus the evolution of EIoT usage over longer timescales remains uninvestigated.

To address these gaps, this work aims to study the evolution of traffic, mobility, and population characteristics of EIoT devices on several different timescales and on both the general and industry-level. Additionally, the work aims to illustrate the feasibility of modeling and predicting several of these diverse characteristics. Impact-wise, traffic, mobility, and population characteristics all are relevant factors for network operators in terms of both technical network operations, network planning, and longer term network investment strategies.

Towards these aims, we analyze a two-year EIoT dataset from a major Finnish mobile network operator (MNO) that includes data traffic volumes, customer industry class, and device features. We first perform a descriptive analysis of the traffic, mobility, and population characteristics while highlighting insights along the way. We then, in terms of modeling and prediction, cluster daily temporal traffic patterns, forecast longer term traffic volumes, model mobility through a Markov mixture model, and calculate the upper bound of mobility predictability (for an ideal model).

Overall, the work gives a holistic view of the evolution and current state of EIoT usage in a major MNO, thus illustrating the reality instead of the hype. We note that Finland is an early EIoT adopter with the 6th most M2M modules per capita of OECD countries (23 per 100 inhabitants) [6]. We also note that we do not claim the results can be greatly generalized to other operators or countries. Instead we hope that the results can represent a case in a larger process of cross-study comparison between different operators and countries, thus building up general patterns and theory.

The results of this study are relevant to both researchers and practitioners. In particular, researchers studying the impact of EIoT on cellular networks can use the identified EIoT traffic and mobility patterns to improve modeling. Furthermore, providers of EIoT connectivity and other services can get a better understanding of the requirements and challenges of IoT devices in different verticals, which will allow them to address customer needs. Also we note that the broad and diverse characteristics and methods of the study were chosen to give, as mentioned, a holistic guide such that future work can focus in on more specific details and with more specialized methods.

We briefly describe the structure of the remainder of the paper. Section 2 summarizes related work on empirical IoT traffic analysis, Section 3 describes

the dataset, and Section 4 details the basic descriptive results including temporal, spatial (i.e., mobility), and EIoT device population aspects. Section 5 presents further analysis and modeling results covering temporal clustering, temporal forecasting, and mobility modeling. Finally Section 6 discusses the limitations and Section 7 the conclusions and implications.

## 2. Related Work

Shafiq et al. [4] were the first to analyze EIoT [1] data from a commercial cellular network in the US. They examined the traffic generated by more than a million EIoT devices over one week in August 2010 and found that such devices are less mobile than smartphones, generate more uplink than downlink traffic, and often have synchronized activity. Ref. [5] confirmed these last two observations by analyzing EIoT device data collected over several weeks in 2013 by a European mobile operator. Both studies concluded that the traffic generated by EIoT devices significantly differs from smartphones, indicating the need for MNOs to reassess network planning traditionally optimized for smartphone users.

In a more recent study, Andrade et al. [7] analyzed the traffic and mobility patterns of one million connected cars on a cellular network in the US. The authors concluded that the data traffic that cars generate differ both from smartphones and other IoT devices, and warned about the potential adverse impact that massive over-the-air firmware updates may have on network performance.

Several studies [8, 9, 10] similarly analyzed IoT data from a cellular network but with different objectives. The studies proposed methods for online and offline classification of IoT traffic that would give MNOs a more efficient way of identifying IoT devices compared to the traditional TAC-based (Type Allocation Code) approach.

From a mobility modeling perspective, smartphone mobility has been extensively modeled using empirical data. For example, based on a 13-month dataset, [11] modeled movement between highly visited locations in which transition speeds and pause times follow log-normal distributions. In another example, [12] modeled device location transitions via a semi-Markov process using a transition matrix and transition time distribution. However, as far as we know no study has modeled mobility for general EIoT devices. Though, empirical mobility models for specific IoT devices types (such as vehicles [13]) have been developed.

Several studies have also modeled longer term (>1 week) internet traffic volume trends (though only at the aggregate level without differentiating IoT devices). For example, [14] accurately fit a hyperbolic function to the CAGRs[2] of 20 years of fixed and mobile internet traffic volumes. Ref. [15] provides further

---

[1] They denoted such traffic as machine-to-machine (M2M).
[2] Compound annual growth rates

3

background into such long term traffic modeling including varying methods, timescales, and datasets.

## 3. Dataset

Before describing the dataset, we first note that an EIoT device (e.g., smart-meter, asset tracker, etc.) typically contains a generic communication module (CM) to transmit the data the device collects. Since these modules are often integrated, a device naturally inherits properties of the CM such as network capabilities. Therefore, in this work differentiating between properties of the EIoT device and CM is only required in a few cases. Hereafter, we note specifically when this is the case.

The main dataset of the analysis is a collection of data detail records (DDRs) of devices that use an IoT-specific enterprise subscription provided by a major nation-wide Finnish MNO. In other words, in this work, an EIoT device is defined as a device that uses a IoT-specific enterprise subscription (subscription based definition).

The dataset covers a period of two years from September 2016 to August 2018. Each record covers a single hour and contains the following fields: anonymized IMSI[3], anonymized cell ID, anonymized customer ID (hereafter company ID), device TAC, uplink traffic volume (in bytes), and downlink traffic volume (in bytes). If the device had traffic in more than one cell in a given hour then additional records for that hour for each cell were included. In other words, each record is uniquely identified by a triplet of (device, cell, hour). The dataset was extracted from the operator's network accounting system which receives aggregate statistics from base stations. We also note that the hourly time granularity of the dataset is a result of collecting the dataset from this network accounting system.

Additionally, the DDR dataset was joined with two other MNO provided datasets: a dataset of device features (from the GSMA device database) for all TACs found in the DDR dataset and a dataset of company industries for each company ID in the DDR dataset. The device feature dataset includes the following fields: device CM model name, device CM release year, and device network capabilities (i.e., EDGE, HSPA, LTE, etc.), while the company industry dataset is based on the standard Finnish TOL2008[4] industry classification. For industry-level analyses, we only include industries with at least 10 distinct companies and where the largest company accounts for a maximum of 80% of traffic or devices in the industry. For reference, we list these industries, their acronyms (used in figures), and brief descriptions in Table 1.

---

[3]We only refer to devices in this work and we assume a one-to-one relationship between IMSI and device since SIM cards are rarely swapped to different devices. Empirically we find that only 1.6% of IMSIs were used with multiple devices over the entire period.

[4]TOL2008 is based on the EU's classification of economic activities, NACE Rev.2 [16], prescribed in EC Regulation No 1893/2006

As previously mentioned, in this work, an EIoT device is defined as a device that uses a enterprise IoT-specific subscription (i.e., subscription based definition). However, to further ensure that only EIoT devices were included in the analysis, we first manually checked all unique device CM models from the dataset and categorized them as IoT, maybe-IoT (typically PCI Express data cards that can also be used in laptops), and non-IoT (typically smartphones and feature phones) based on online research. We then filtered out all non-IoT devices and any device with an invalid TAC code (since in those cases we did not have any device information). This filtering removed 5.7% of devices.

We also note that since we only know the CM model name (and not the EIoT device model name), we do not know the EIoT device type. For example, we might know that a device has a Cinterion EU3-E module, but we do not know if the device is a smart meter, asset tracker, etc. Unfortunately inferring the device type from just the CM is not feasible because, as mentioned, these modules are generic and many manufacturers do not identify the CMs in their devices. We even attempted to scrape FCC and other regulatory approval reports to identify the CMs in devices but with limited success.

In any case, we know that the device population includes EIoT devices like payment terminals, smart-meters, location trackers, and surveillance cameras. In terms of requirements, some of these devices require very little bandwidth. Payment terminals, for example, typically use the ISO-8583 financial message standard which requires only several kilobytes per payment transaction or even less [17]. While other devices such as video surveillance cameras can require over one Mbps depending on the resolution (e.g., a 720p 30 frame per second H.264 camera requires about 1.9 Mbps). Overall, the analysis further illustrates this requirement diversity, especially across industries.

Finally, to give an idea of the full scale of the analysis, the DDR dataset covers hundreds of companies, hundreds of thousands of devices, and tens of millions of records. We also note that for business confidentiality and privacy reasons we normalize some of the numerical results, however, this normalization does not change the interpretations or conclusions. Finally, for illustration purposes we use moving averages[5] in several figures to help emphasize longer-term trends and smooth out short-term fluctuations.

## 4. Descriptive Results

### 4.1. Traffic statistics

First, we examine the traffic of cellular EIoT devices over time to understand its evolution. Figure 1 shows the four-week moving average of traffic per device. We find that total EIoT traffic per device increased three-fold, whereas downlink traffic increased six-fold. Most of the traffic growth occurred between September 2016 and 2017. Comparatively, [4] reported a total EIoT traffic increase of

---

[5]The moving average is essentially a low-pass filter in signal processing.

Table 1: Description of industries based on NACE Rev.2 [16]

| Industry (abbreviated) | Acronym | Description |
|---|---|---|
| Administrative and support | AS | Activities supporting general business operations, except professional activities; e.g., rental and leasing, recruitment, security and investigation |
| Electricity and gas | EG | Providing electric power, natural gas, steam, hot water and the like through a permanent infrastructure |
| Information and communication | IC | Publishing activities, including SW; broadcasting; telecommunications and IT activities |
| Manufacturing | MF | Physical or chemical transformation of materials or components into new products, e.g., food, textiles, computers and electronics |
| Professional activities | PA | Activities making specialized knowledge available to users, e.g., consultancy and engineering |
| Transportation | TR | Provision of passenger or freight transport, and associated activities such as terminal and parking facilities, cargo handling, and storage |
| Wholesale and retail trade | WR | Wholesale and retail sale of any goods, including associated operations, such as assembling and packing; repair of motor vehicles and motorcycles |

250% during 2011. Furthermore, despite some fluctuations, the traffic does not demonstrate any seasonal patterns.

The volume of traffic increased in all industries, as shown by the four-week moving average of traffic per device in Figure 2. We observe significant differences in traffic volumes between industries, with devices in *Manufacturing* and *Administrative and support* generating on average 10 MB and 2 GB per device, respectively. This difference might be explained by *Administrative and support* containing security companies that may generate video traffic via surveillance cameras. The most substantial increase occurred in *Electricity and gas* where traffic grew twelve-fold to 19 MB per device. This could be due to a 1-hour metering obligation deadline for smart meters in Finland. We also analyzed the traffic evolution for the subset of companies that had active IoT devices in both the first and last months of the observation period and found similar trends. This indicates that the increase in traffic over time includes both companies that already use IoT and new companies adopting IoT.

To explore industry-specific differences in traffic depending on the day of the week, we study the daily traffic per device for August 2018. Figure 3 illustrates this traffic. We find that most industries do not show significant variation depending on the day of the week. However, in *Professional activities* and *Manufacturing* industries, we observe weekday-weekend patterns, with traffic halving during weekends.

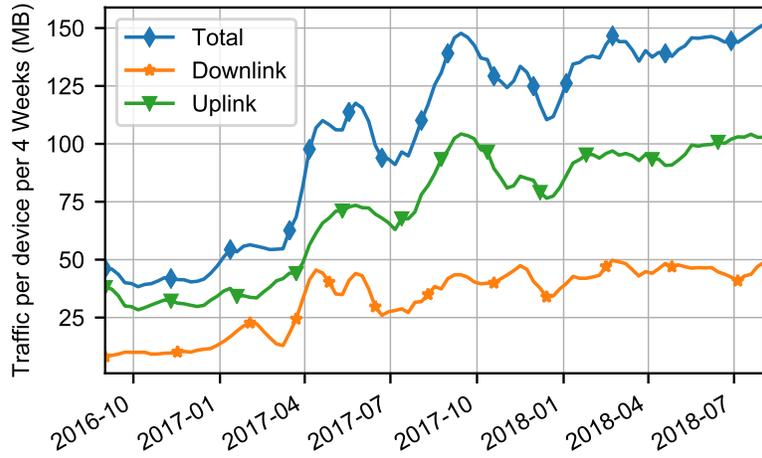Furthermore, we study uplink vs. downlink traffic volumes across industries.

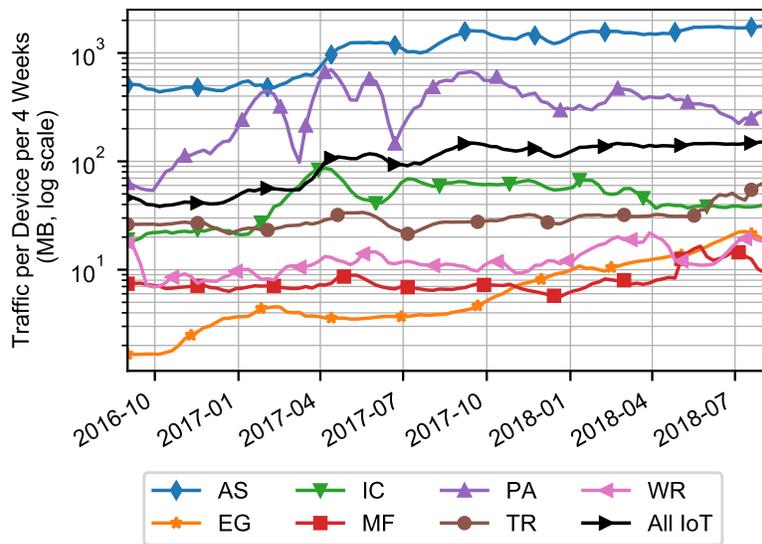Figure 1: Traffic per device per four week period



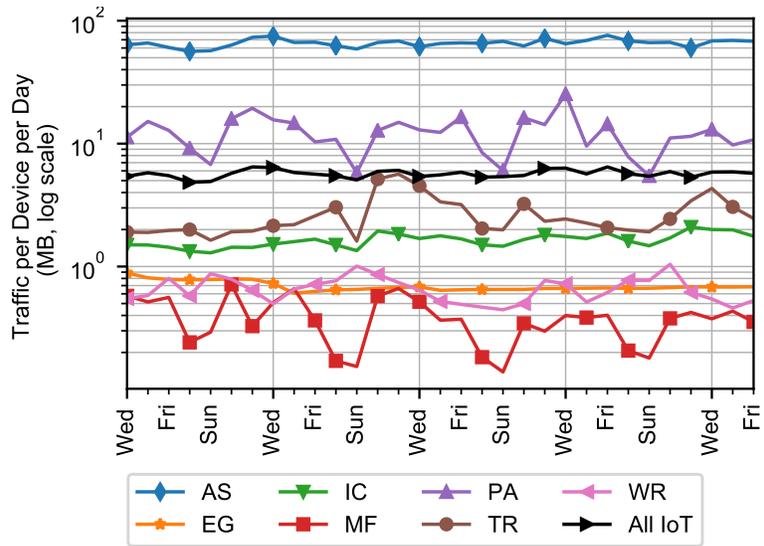Figure 2: Traffic per device per four week period for industries

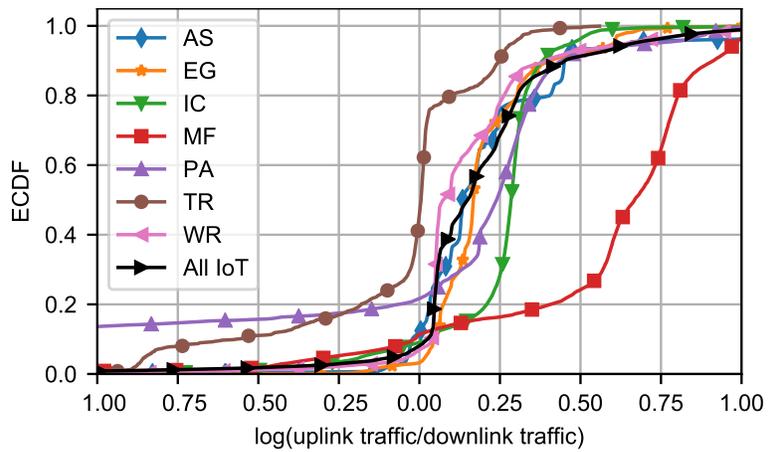Figure 3: Traffic per device per day for August 2018



Figure 4: ECDF of the log of uplink to downlink traffic ratio for August 2018

Figure 4 illustrates the empirical cumulative distribution (ECDF) of the log of uplink/downlink ratio for August 2018. Negative values indicate larger downlink traffic than uplink and positive values vice versa. As the figure shows, 92% of EIoT devices generated more uplink than downlink traffic, which is consistent with the finding of [4], but exceeds the observation of [5] by more than 30%. However, in some industries, particularly *Transportation* and *Professional activities*, the share of devices with greater uplink than downlink traffic is lower, around 54% and 78% respectively. Further, in *Manufacturing*, the uplink traffic is much larger than downlink traffic (compared to other industries), with a median ratio of 4.66 compared to 1.41 for all EIoT devices. Overall, the results illustrate both intra and inter-industry variation that helps illustrate the diversity of EIoT .

### 4.2. Mobility statistics

Concerning device mobility, we infer such mobility through the number of different cells visited[6] by devices. Figure 5 presents the ECDF of the number of unique cells visited by devices in August 2018. As the figure shows, about 40% of EIoT devices visited only a single cell indicating significant immobility. Furthermore, for a one week time frame (last week of August), we find an even higher fraction of 55% of devices visited only a single cell. In comparison, [4] found that 30% of devices visited only a single cell in their one-week dataset. The actual share of stationary devices (again given our definition from footnote 6) may be even higher since some devices at cell edges may jump between cells depending on signal strength fluctuations or cell load balancing.

We also observe differences in device mobility across industries. Overall around 95% of all EIoT devices and most devices in *Electricity and gas*, *Wholesale and retail trade*, and *Administrative and support* industries visited less than ten cells per month. Contrastingly and expectedly, devices in the *Transportation* industry are very mobile, with 90% having visited more than ten cells per month. Some industries, for example *Manufacturing*, include a mix of mobile and stationary devices.

### 4.3. Cell statistics

We analyze the distribution of devices and traffic over all the visited cells. Figure 6 illustrates the ECDF of EIoT traffic and devices[7] across EIoT -visited cells in August 2018. The traffic is highly concentrated spatially, with 10% of cells carrying about 93% of total EIoT traffic. Comparatively, [18] found 10% of cells carrying about 55% of total network traffic in a nationwide network in 2007. The high concentration of EIoT devices and traffic can be explained, given the typical centralization of company campuses compared to normal consumers.

---

[6]The definition of visit only includes cells where traffic was sent or received and thus it is a lower bound on the number of cells attached to by the device.

[7]We note that devices are only counted in their most visited cell (in terms of hours), though other definitions such as counting devices in all their visited cells produce similar results.
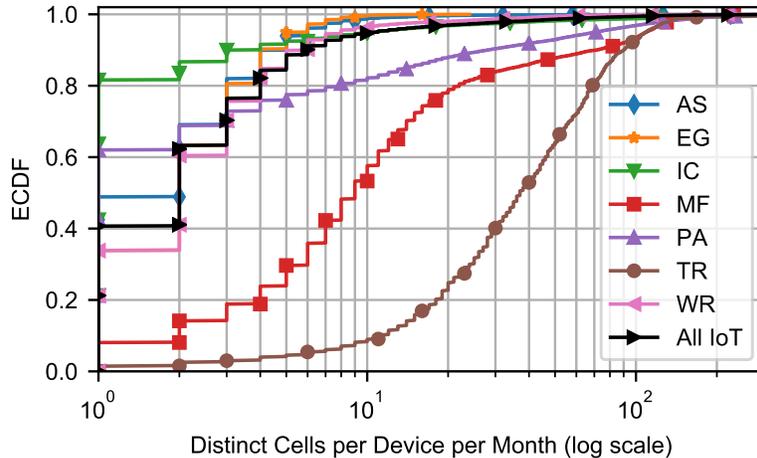
Figure 5: ECDF of EIoT device mobility in August 2018

This concentration is important for network planning because the deployment of EIoT -specific network features or optimizations would require changes to far fewer cells (and thus cost less) than for non-EIoT features. Finally, in terms of devices, we find that 10% of cells account for about 44% of devices while 50% of cells about 90% of devices, showing only moderate spatial concentration.

### 4.4. EIoT device population statistics

Through leveraging the additional device features dataset, we can analyze the feature and CM age evolution of the EIoT device population. Figure 7 shows the evolution of the mean CM age by industry, with age defined as the time since the release year of the CM model[8] (and not the manufacturing year of the CM). We observe that the mean CM age was over 8.5 years, as of August 2018. The *Electricity and gas* industry has the oldest CM population, with a mean age of more than ten years. Overall, the increase in mean population age for all industries illustrates the slow pace of new CM model deployment. We further analyzed the population of CMs deployed after September 2016 and found that the mean age in August 2018 was about seven years.

In terms of connectivity features, Figure 8 presents the penetration of 3GPP connectivity technologies among the EIoT device population. First, we observe the low penetration (and growth rate) of LTE of about 2% as of August 2018. This observation is in line with the significant age of the EIoT device CM population and contrasts with LTE penetration of 41% among non-EIoT devices in Europe in 2017 [19]. Additionally, the growth rate of LTE among Finnish smartphones was seven times as large as EIoT devices over a comparable yearly

---

[8]We assume that CM models are released on Jan. 1st. In other words, we overestimate the actual age, but this does not preclude tracking temporal dynamics and comparing industries.
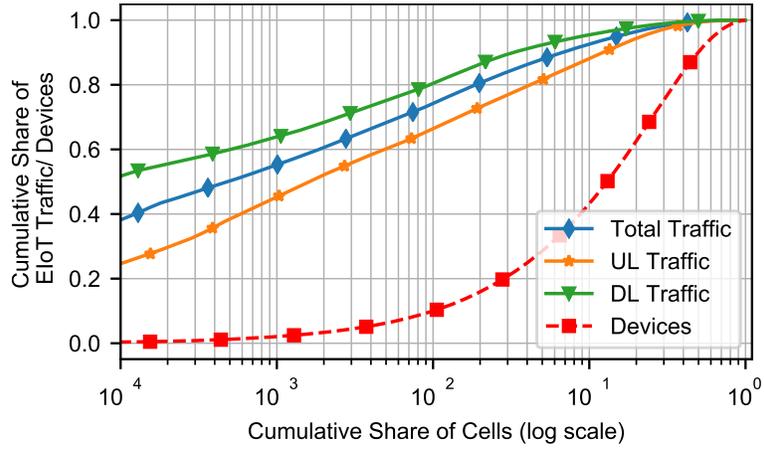
Figure 6: Distribution of traffic and devices among cells for August 2018
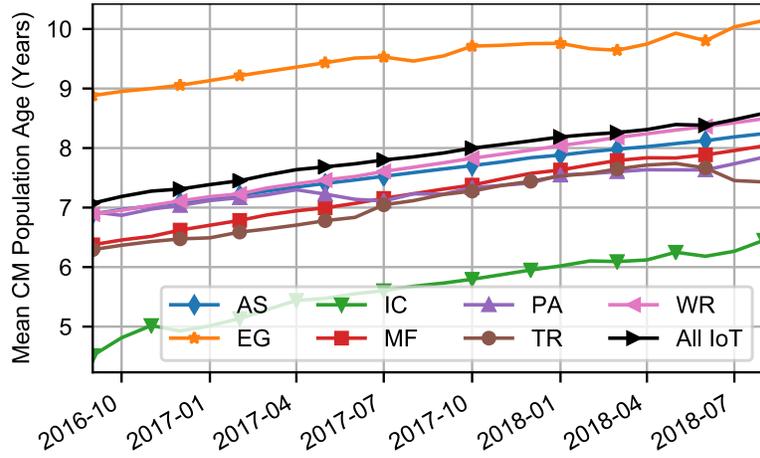


Figure 7: Mean CM population age (in years) based on the CM model release year
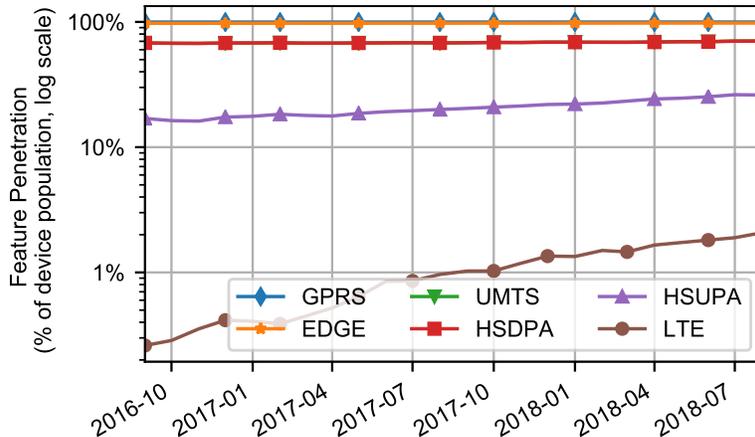
11

Figure 8: Penetration of EIoT device features

time-frame (specifically when re-basing both series to start at 0.1% penetration) [20].

We further find that although the penetration of LTE is growing in all industries, only in the *Transportation* industry has penetration exceeded 10%. Furthermore, we observe a significant difference in the penetration of HSDPA and HSUPA technologies of 70% and 26% respectively. This is surprising given the prevalence of uplink traffic in EIoT which suggests a stronger need for fast uplink technologies rather than downlink. Finally, we find the share of 2G only (GPRS and EDGE) devices is still about 30%. Therefore, discontinuing 2G service (for spectrum reuse purposes) would indeed affect a significant fraction of EIoT devices thus posing a problem for network operators.

We also examine the prevalence of NB-IoT/LTE-M capable devices in the population by using a publicly available list of such devices from GSMA[9]. However we find that these devices represent less than 0.05% of all devices and thus are too small of a sample for reliable analysis. Furthermore the devices currently using the network are likely primarily testing devices.

We examine the shares of CM vendors in the device population over the observation period[10]. Figure 9 illustrates the CM vendor shares over time and the Herfindahl-Hirschman index (HHI), a measure of market concentration. The figure indicates quite high concentration with only a small decrease in concentration as quantified by HHI (from 0.53 to 0.47) over the observation period. Though, we note that if the customer company with the most IoT devices is removed (to assess sensitivity), the HHI for August 2018 drops to 0.30, thus illustrating the potential for volatility given large customers.

---

[9]https://www.gsma.com/iot/mobile-iot-modules/

[10]We gather CM vendor information including public mergers, acquisitions, and divestment data (including dates) for all CM vendors that appear in the device population.
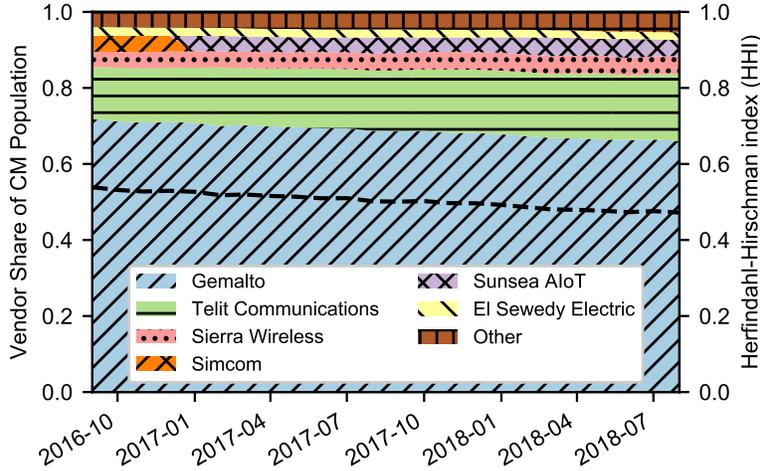
Figure 9: Vendor share of CM population and the Herfindahl-Hirschman index (HHI)

## 5. Modeling and Prediction Results

### 5.1. Temporal traffic spectrum and clustering analysis

To obtain additional temporal traffic insights, we perform several types of temporal analysis on different time scales. Specifically, our approach is to study the short timescale (hours, weeks) temporal patterns of three different months roughly evenly spaced over the two years: September 2016, August 2017, and August 2018. We always present the results from August 2018 and only present and note the results from the earlier months if substantially different.

First, we perform spectral analysis on a one-month traffic volume series of each device for uplink and downlink traffic. The spectral density of each series is estimated as the squared modulus of the discrete Fourier transform, in other words the periodogram. Then the peak power and corresponding period are extracted from each periodogram under the assumption that most EIoT devices will have a dominant timer-driven peak. Figure 10 illustrates the density of these (peak power, period) pairs for downlink traffic. The plot for uplink is almost identical. We find large fractions of devices have peaks at 24, 12, and 6 hour periods including devices with large and small peak traffic volumes (power). However, we also find other periods such as ∼13 hours, though this case is specific to only two large companies with similar device models. The reason for the use of a 13 hour period in these companies is unknown. We also note that some devices have peaks at one week thus reinforcing the patterns from Figure 3, however these devices tend to have small peak traffic volumes.

For a more granular temporal analysis, we perform temporal clustering on the averaged (over the month) and normalized 24-hour total traffic volume series of each device. The normalization is performed for each device over the 24-hour series such that the value for any given hour is the fraction of that device's total daily traffic in that hour. This normalization is required due to the order of
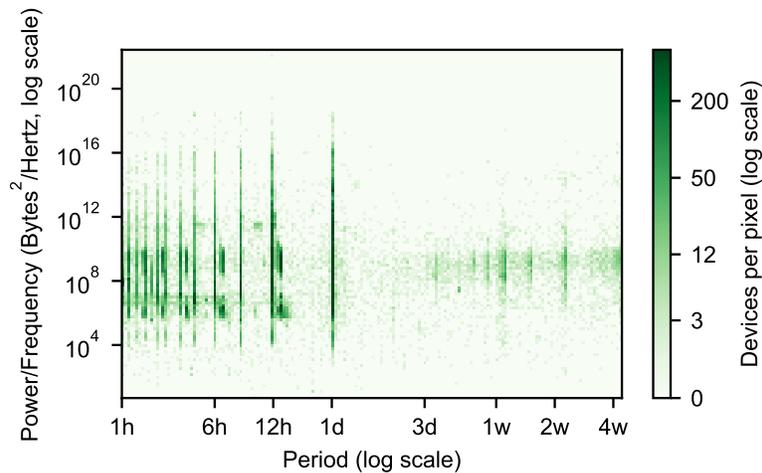
13

Figure 10: Density of spectrum peaks vs periods of devices for total traffic for August 2018

magnitude differences in traffic volumes between some devices. Each series is then transformed by a discrete wavelet transform (DWT) with a Daubechies-1 wavelet and a decomposition level of three.

The DWT coefficients are then clustered via bisecting k-means with the number of clusters chosen by the silhouette score. We use bisecting k-means because of the $O(n)$ run-time and ease of computational distribution. Comparatively, other approaches such as hierarchical clustering with ward linkage have a run-time of $O(n^2)$. Though for robustness, we also cluster a random sample of 2000 devices via hierarchical clustering with ward linkage and with the number of clusters chosen by the Davies-Bouldin score. This is the same clustering setup as in [4]. We find the same number of clusters as the full device clustering and virtually the same cluster centroids.

Regarding clustering results, we find that the optimal number of clusters is three. The clusters denoted 1, 2, and 3 encompass 25%, 41%, and 34% of devices respectively. The cluster centroids (in terms of time series rather than DWT coefficients) of the three clusters are illustrated in Figure 11. We find that clusters 1 and 3 have significant peaks at 0:00 and 2:00 respectively with over 80% of their traffic within that peak hour, while cluster 2 shows much steadier and flatter traffic throughout the day.

To better understand these patterns we look at the composition of the clusters by company ID and industries. Interestingly, 88% of cluster 1 devices belong to a single large company; thus this cluster is company-specific and not necessarily a general EIoT temporal pattern. Though [4] also found an outlier cluster with a peak at 02:00 that they attributed mainly to fleet management applications. For clusters 2 and 3, no single company represents more than 31% of devices and no single industry more than 50% of devices. The main industries for cluster 2 are *Wholesale and retail trade* (40%), *Electricity and gas*
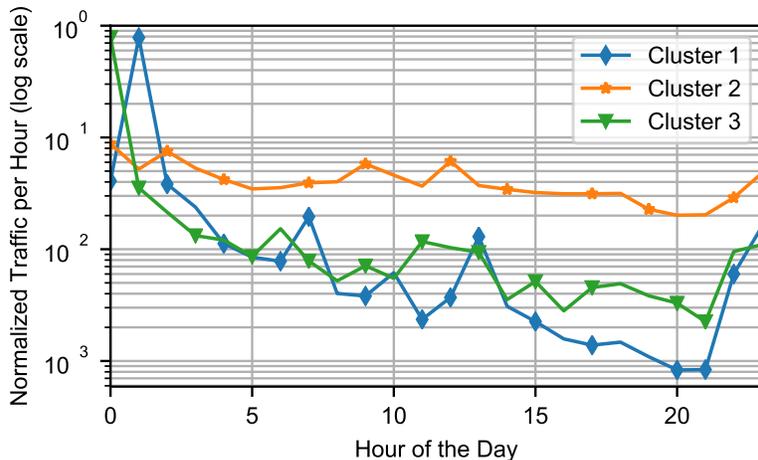
Figure 11: Cluster centroids of the three temporal clusters from August 2018

(22%), and *Information and communication* (14%), while the main industries for cluster 3 are *Wholesale and retail trade* (51%), *Electricity and gas* (30%), *Administrative and support services* (11%). This overlap in industries highlights diversity in use cases even within narrow industries such as *Electricity and gas*.

For an illustration of cluster separation, we plot the t-distributed stochastic neighbor embedding (t-SNE) of a random sample[11] of 4400 devices in Figure 12 with perplexity chosen as in [21]. The clusters appear to be well separated with only minimal overlap, especially the single-company dominated cluster 1, thus reinforcing the clustering results.

In terms of longer scale temporal phenomena, we did not find large differences in either temporal analysis method between the examined months. This suggests that EIoT phenomena change slowly; such behavior reinforces the previously identified slow change in, for example, device feature penetration.

### 5.2. Temporal traffic forecasting

Finally, we examine the possibility of EIoT traffic forecasting (a common network operator goal). Specifically, we evaluate the potential for medium-term forecasting of daily EIoT traffic through a state of the art forecasting model. This medium-term (from 30 to 365 day horizons) forecasting is the most practical given the length of our observation period (about two years) and granularity (hourly).

Namely, we use the Prophet time series model [22] which is a decomposable additive regression model. Equation 1 details the high-level model formulation consisting of piece-wise (linear or logistic growth) trends $g(t)$, seasonality (weekly or yearly) $s(t)$, and holiday $h(t)$ components plus an error term $\epsilon_t$. The

---

[11]t-SNE has a run-time complexity of $O(n^2)$ and thus does not scale to large data.
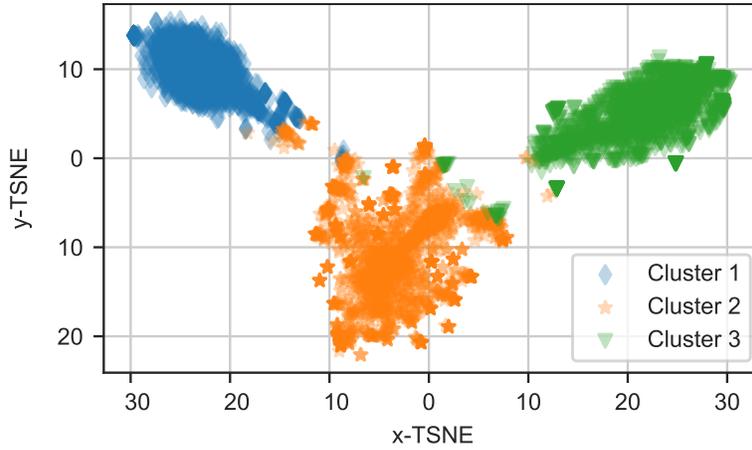
Figure 12: T-SNE of sample of 4400 devices from the three temporal clusters from August 2018

model fitting is flexible and automatically selects the appropriate trend change points and components. Specifically, the model fitting is performed through the probabilistic programming language Stan [23] via maximum a posteriori parameter estimation. The holiday component of the model uses the national holidays of Finland. The main justification for using Prophet over alternative models such as autoregressive integrated moving average (ARIMA) is that Prophet was designed for web event modeling and thus natively supports common network traffic properties such as the aforementioned piecewise trends, seasonality, and holidays.

$$y(t) = g(t) + s(t) + h(t) + \epsilon_t \tag{1}$$

Firstly, for reference Figure 13 illustrates the daily traffic per device series for all EIoT devices (hereafter All-EIoT ), the model fitted on that series for the entire observation period, and a one year forecast. The model illustrates two distinct trends with a visible change point at Oct. 2017 and a weekly seasonality that aligns with a weekday/weekend dichotomy. Visually, the model appears to provide a simple though reasonable fit of the series.

Next, to estimate the accuracy of forecasting we use a rolling window validation method known as simulated historical forecasts (SHF) [22]. In the SHF method forecasts are made for rolling historical horizons given a fixed training window size, a variable horizon window size, and a fixed period for shifting these windows within the full historical window. We use a training window size of 365 days, horizon window sizes from 30 to 180 days, and a period of 90 days. This allows the estimation of forecasting accuracy for time horizons ranging from 30 to 180 days. This estimation is performed for the all-EIoT daily traffic per device series and each daily industry series individually. For easy interpretation, accuracy is quantified by the mean absolute percent error (MAPE).
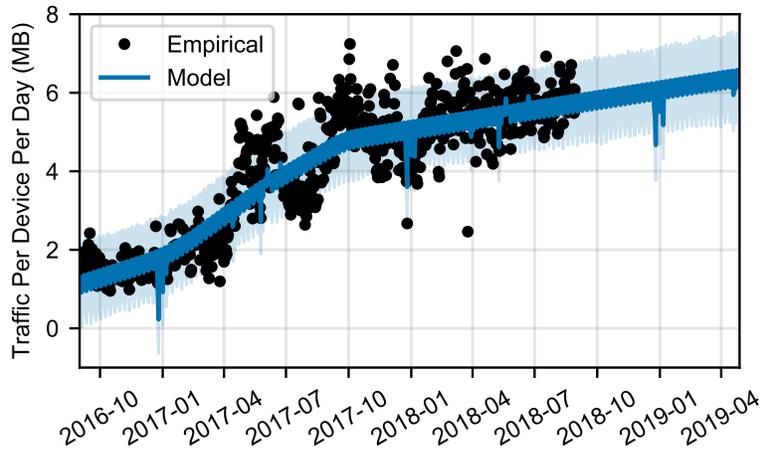
Figure 13: All-EIoT daily traffic per device and fitted model including one year forecast

In terms of results, Figure 14 details the estimated MAPE for forecasts over different time horizons for both the all-EIoT and industry-specific models. Interestingly, accuracy for the all-EIoT model is respectable with an error between 10-20%; whereas for the different industries the model accuracies vary more substantially. The low accuracy for the *Professional activities* industry is due to the high series variance and lack of a clear trend (see again Figure 2) suggesting that the industry may be too diverse to be a useful as an aggregation. Generally, as expected, accuracy decreases with longer time horizons. The overall moderate accuracy implies that network operators could use such models at least for general high-level planning of EIoT usage in the medium-term.

Additionally, forecasting of individual customer company traffic over time may be useful especially for large companies. To assess the viability of such forecasting, we perform the same procedure as previously but for the ten largest companies by number of devices. These companies are from six different industries and thus relatively diverse. We find the mean MAPEs (over the 30-180 day horizons) for the companies range between 2% to 55%; therefore illustrating similar diversity as on the industry level. This also reinforces that the traffic variation that impinges forecasting is both intra-industry and intra-company.

For research purposes we release publicly the all-EIoT and industry specific models as serialized Python pickle files[12]. These files can be imported into the Prophet library to allow forecasting and interrogation of the parameters.

### 5.3. Mobility modeling and analysis

Next, we perform mobility modeling to capture the typical patterns EIoT devices follow across cells. To do so we utilize a finite mixture model of Markov

---

[12]The models can be found at (link omitted for double blind review purposes).
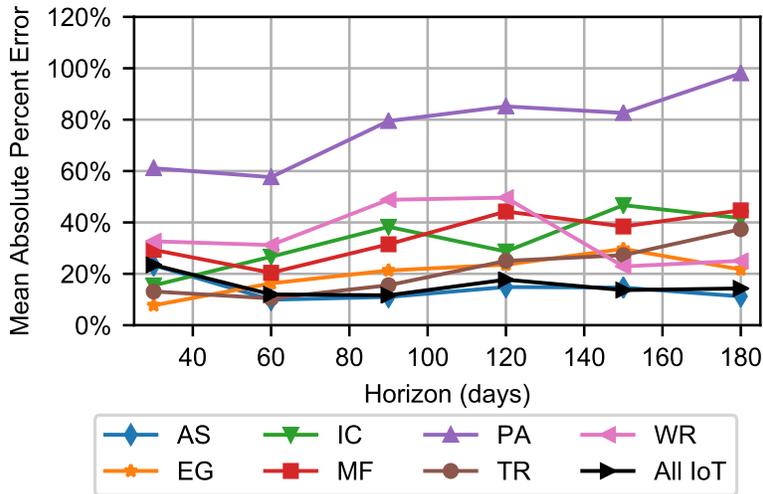
Figure 14: Estimated MAPE of forecasts over varying time horizons (30 to 180 days) for all EIoT and industry-specific models

models fit through a library known as Clickclust [24]. Each Markov model represents the transitions between a fixed set of categories (specifically cells).

More specifically, Clickclust [24] estimates a finite mixture model of Markov models for a set of categorical sequences where the probability distribution of the finite mixture model is

$$f(\gamma|\vartheta) = \sum_{k=1}^{K} \alpha_k f_k(\gamma|\vartheta_k)$$

where $K$ is the number of component distributions $f_k(\cdot|\vartheta_k)$ with parameter vectors (and mixing proportions $\alpha_1, \ldots, \alpha_{K-1}$) subject to restrictions $\alpha_k > 0$ and $\sum_{k=1}^{K} \alpha_k = 1$. Each component distribution is a first order Markov model representing a cluster of similar cell sequences. The number of component distributions $K$ is selected through agglomerative clustering to minimize the Bayesian information criterion (BIC) via a two-stage iterative procedure with an expectation-maximization (EM) algorithm.

Due to computational complexity issues, the assumption of diurnal patterns, and the desire to model mobile (rather than mostly stationary) devices, we perform some initial processing and filtering. Specifically, we only model the cell sequences of devices from September 30 to 31, 2018. We also remove runs of the same cell in the cell sequences and thus focus only on the notion of mobility. We then select a random sample of 2000 devices with a cell sequence length of at least five (i.e., sent or received in at least five hour-cell combinations) and a cell sequence with between three and 50 distinct cells. Finally, we normalize each cell sequence by encoding the most frequent cell as 0, the next most frequent as 1, and so on.

In terms of results, a simple application of the EM algorithm suggests an optimal mixture model with one component, as shown in Table 2 (where BIC is minimized at $K = 1$). This result suggests potential model over-parameterization and order underestimation, which is possible with a large Markov state space (our state space is 50 due to the distinct cell limit from our filtering). Fortunately, Clickclust contains a forward state selection (FSS) algorithm, allowing for the aggregation of Markov states into equivalence blocks considering their transition probabilities. The application of the FSS procedure estimates an optimal mixture model with three components with a BIC of 66014.44. Table 2 provides a performance summary for both the EM and FSS procedure[13] including the number of equivalence blocks $d$ for FSS.

Table 2: Clickclust mixture model fitting BIC scores for different numbers of Markov models (in parentheses is the number of equivalence classes in FSS)

| Method | $K = 1$ | $K = 2$ | $K = 3$ | $K = 4$ |
|---|---|---|---|---|
| EM | **88203.11** | 104810.3 | 121983 | 139020.2 |
| FSS | 66961.61 (9) | 66236.89 (8) | **66014.44** (7) | 66177.28 (6) |

The optimal solution provided by the FSS procedure has the mixing proportions $\alpha_1, \alpha_2, \alpha_3$ of 0.67, 0.04, 0.29, indicating an unbalanced weight distribution for the components. The FSS procedure aggregated the 50 distinct cells into seven equivalence blocks, as shown in Table 3. We observe that the number of distinct cells per block increases when including less visited cells, indicating that transition probabilities among less visited cells are more similar than among more visited cells.

Table 3: Mapping of distinct cells to equivalence blocks in FSS

| Cell Order | Equivalence Blocks |
|---|---|
| 1st-25th most visited | 4 5 6 7 3 3 3 3 2 2 2 2 2 2 2 2 2 2 1 1 1 1 1 1 1 |
| 25th-50th most visited | 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 |

Figure 15 illustrates the transition probabilities between equivalence blocks for each component of the mixture model. Component 1, which has the highest weight, mainly models the bidirectional transitions between the two most visited cells (equivalence blocks 4 and 5, as shown in Table 3). Component 2, which has the smallest weight, primarily models the transitions between the least visited cells (blocks 1 and 2) and other blocks. Finally, component 3 models transitions between multiple other blocks excluding block 1. The high transition probabilities between the top cells may partly be the result, as discussed earlier, of devices at cell edges that jump between cells depending on signal strength fluctuations or cell load balancing. Remember also that sequences with less

---

[13]For reference, the EM and FSS procedures use the following parameters iter = 3, eps = 1e-8, r = 50, min.gamma = 1e-2, and min.beta = 1e-2. The computation time for FSS for K=4 was approximately ten hours on eight Intel Xeon X5650 CPUs with 80GB total RAM.

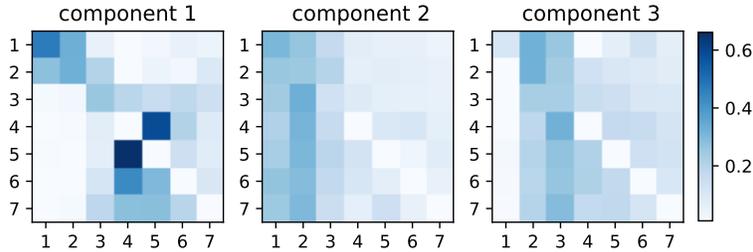than three unique cells were removed from the mobility modeling.



Figure 15: Transition matrices between equivalence blocks for each component of the FSS model

For research purposes we release publicly the 3-cluster FSS mixture model and script[14] (as serialized R objects and files) thus allowing for model interrogation and the synthetic generation of cell sequences. Transition matrices can also be found within the script.

We also perform an alternative mobility analysis using an information theory framework. Specifically, we quantify for all devices the potential for next cell prediction and optimization by estimating the information theoretical upper bound of predictability (hereafter $\Pi^{max}$) similar to [25]. $\Pi^{max}$ denotes the theoretical maximum percentage of cell visits that could be predicted given the entropy of the cell sequence. Though in contrast to [25] we only estimate the predictability of cell attachments with data transfer (as this is our definition of cell visit) rather than of all cell attachments (which would infer the full mobility as in [25]). This formulation avoids a common missing data problem from prior work in that cell attachments without data transfer do not generate CDRs or DDRs and are therefore often absent from mobile network datasets.

The entropy of a sequence is estimated through a Lempel-Ziv compression based estimator detailed in Equation 2 where $n$ is the length of the sequence and $\Lambda_i$ is the length of the longest subsequence starting from $i$ and not seen earlier from 1 to $i-1$. This estimator quickly converges to the true entropy rate as $n \to \infty$. This entropy is then used to numerically solve for $\Pi^{max}$ through Equation 3 (which is derived from Fano's inequality [25]) where $N$ is the number of distinct symbols (i.e., cells) in the sequence.

$$H_{rate} = \left( \frac{1}{n} \sum_{i=1}^{n} \frac{\Lambda_i}{\log_2 n} \right)^{-1} \tag{2}$$

$$H_{rate} = -\Pi^{max} \log_2 \left( \Pi^{max} \right) - \left( 1 - \Pi^{max} \right) \log_2 \left( 1 - \Pi^{max} \right) + \left( 1 - \Pi^{max} \right) \log_2 \left( N - 1 \right) \tag{3}$$

---

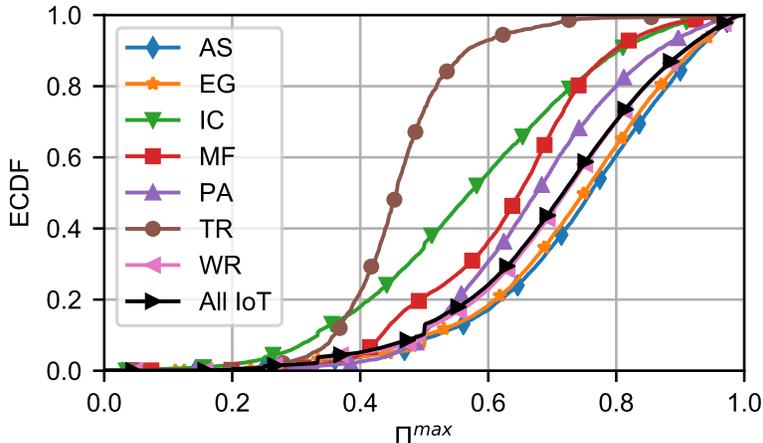[14]The models can be found at (link omitted for double blind review purposes).

Figure 16: ECDF of $\Pi^{max}$ (upper bound of predictability) for devices by industry for period of March to August 2018.

We examine the single cell sequence covering from March to August 2018 (last six months of the observation period). As before we remove runs of the same cell in the cell sequence. We also limit the analysis to devices with a cell sequence length of at least 20. The longer observation period and length limit are necessary because the entropy estimator requires a reasonable length for accurate estimation[15]. This limitation removes lower activity, mostly stationary devices (about 39% of devices).

Figure 16 illustrates the ECDF of $\Pi^{max}$ for devices by industry for the period of March to August 2018. The results illustrate that a non-trivial fraction of EIoT devices have unpredictable cell mobility dynamics; especially in the *Transportation* and *Information and communication* industries. This suggests that network operators could first focus their predictive optimization efforts on industries with high predictability to gain quick wins.

## 6. Limitations

We next discuss two major study limitations. First, the study is not fully representative since the dataset is from only a single MNO in a single country and includes only EIoT and not consumer IoT. However, as previously mentioned, we hope that similar studies from other countries and MNOs can help to build up a wide-ranging and practical understanding of IoT dynamics. Second, the time resolution of one hour means that the study might have missed

---

[15]Specifically, the variance and bias decrease proportionally as $1/n$ and $1/log_2(n)$ respectively. The threshold of 20 is somewhat arbitrary, unfortunately determining the sequence length required for a specific entropy estimation accuracy is non-trivial and current methods (i.e [26]) assume i.i.d. and Zipfian symbol probabilities.

more granular phenomenon on the minute and second timescales. However, we note that potentially more intrusive data collection methods would be required for those timescales, thus potentially hindering collection. Therefore, we leave such granular work for future studies.

## 7. Conclusions

Overall, this work presented an analysis of cellular EIoT traffic and mobility patterns over several different timescales for a major Finnish MNO. The analysis includes trends over a two-year span thus allowing a view of the evolution of EIoT . Moreover, trends were broken down by industry, and the penetration of device features in the EIoT device population was analyzed. Finally, the analysis evaluated EIoT traffic forecasting and mobility modeling. Overall the analysis provided a diverse set of results of which we highlight a few.

For example, we found that EIoT traffic per device tripled over the last two years; however, the mean age of CM models in the device population also increased significantly to over eight years. Furthermore, the penetration of LTE-enabled EIoT devices is very low (2%) and growing very slowly. Also we found significant variation between devices of different industries with orders of magnitude differences in traffic volume and mobility. Furthermore, we illustrated that total daily EIoT traffic can be accurately forecast (~15% error) over a medium-term (30 to 180 day) horizon. Finally, we presented that a non-trivial fraction of EIoT devices have inherent unpredictability in terms of their mobility.

The results have implications for mobile ecosystem players. We note the following example implications:

- MNOs should be cautious in discontinuing 2G service (for spectrum re-farming purposes) since a large fraction of EIoT devices likely still use GPRS and EDGE and the EIoT device life cycle is lengthy.

- Network managers should consider EIoT spatio-temporal traffic patterns when defining mesoscale (on order of hours or days) IoT network configurations and optimizations (such as the powering down of certain BSs for energy saving purposes) or when developing ML models that perform such optimizations. For example, given the very high spatial traffic concentration and thus inter-base station variability, the use of cell level or multi-level (rather than global) ML models is likely to be important.

- Network planners should consider the specific requirements of those industries targeted by business development plans given the significant inter-industry traffic and mobility differences. An example of relevant planning could be the optimal placement of resources for edge computing. Specifically, in the administrative and support industry, uplink capacity could be saved by enabling the edge processing of video streams from CCTV systems with no built-in ML object detection.

- Industry customers, MNOs, and EIoT device manufacturers should collaborate with multiple CM vendors to avoid the risk of over-reliance on a single vendor (given the potential for high CM market concentration).

## 8. Acknowledgments

## References

[1] B. Finley, A. Vesselkov, Cellular iot traffic characterization and evolution, in: 2019 IEEE 5th World Forum on Internet of Things (WF-IoT), 2019, pp. 622–627. `doi:10.1109/WF-IoT.2019.8767323`.

[2] Analysys Mason, Circle Research, Vodafone IoT, Vodafone IoT Barometer 2017/18 (2017).

[3] J. Marjamaa, A measurement-based analysis of machine-to- machine communications over a cellular network, Master's thesis, Aalto University (2012).

[4] M. Z. Shafiq, L. Ji, A. X. Liu, J. Pang, J. Wang, Large-scale measurement and characterization of cellular machine-to-machine traffic, IEEE/ACM Transactions on Networking 21 (6) (2013) 1960–1973. `doi:10.1109/TNET.2013.2256431`.

[5] P. Romirer-Maierhofer, M. Schiavone, A. D'Alconzo, Device-Specific Traffic Characterization for Root Cause Analysis in Cellular Networks, in: M. Steiner, P. Barlet-Ros, O. Bonaventure (Eds.), 7th International Workshop on Traffic Monitoring and Analysis, Springer, Cham, Barcelona, 2015, pp. 64–78. `doi:10.1007/978-3-319-17172-2_5`.

[6] OECD Broadband Portal (12 2017).

[7] C. E. Andrade, S. D. Byers, V. Gopalakrishnan, E. Halepovic, D. J. Poole, L. K. Tran, C. T. Volinsky, Connected cars in cellular network, in: Proceedings of the 2017 Internet Measurement Conference on - IMC '17, ACM Press, New York, New York, USA, 2017, pp. 235–241. `doi:10.1145/3131365.3131403`.

[8] A. Baer, P. Casas, P. Fiadino, L. Golab, M. Mellia, E. Schikuta, DBStream: A holistic approach to large-scale network traffic monitoring and analysis, Computer Networks 107 (2016) 5–19. `doi:10.1016/J.COMNET.2016.04. 020`.

[9] A. Baer, P. Svoboda, P. Casas, MTRAC - discovering M2M devices in cellular networks from coarse-grained measurements, in: 2015 IEEE International Conference on Communications (ICC), IEEE, 2015, pp. 667–672. `doi:10.1109/ICC.2015.7248398`.

[10] M. Laner, P. Svoboda, M. Rupp, Detecting M2M traffic in mobile cellular networks, in: Systems, Signals and Image Processing (IWSSIP), 2014 International Conference on, IEEE, Dubrovnik, 2014.

[11] M. Kim, D. Kotz, S. Kim, Extracting a mobility model from real user traces, in: Proceedings IEEE INFOCOM 2006. 25TH IEEE International Conference on Computer Communications, 2006, pp. 1–13.

[12] J.-K. Lee, J. C. Hou, Modeling steady-state and transient behaviors of user mobility: formulation, analysis, and application, in: Proceedings of the 7th ACM international symposium on Mobile ad hoc networking and computing, 2006, pp. 85–96.

[13] Y. Pigné, G. Danoy, P. Bouvry, A vehicular mobility model based on real traffic counting data, in: International Workshop on Communication Technologies for Vehicles, Springer, 2011, pp. 131–142.

[14] S. K. Korotky, Semi-empirical description and projections of internet traffic trends using a hyperbolic compound annual growth rate, Bell Labs Technical Journal 18 (3) (2013) 5–21.

[15] N. Vlachos, Internet traffic volumes characterization and forecasting, Ph.D. thesis, Heriot-Watt University (may 2016).

[16] Eurostat, NACE Rev. 2. Statistical classification of economic activities in the European Community (2008).

[17] International Organization for Standardization, ISO 8583-1:2003 Financial transaction card originated messages - Interchange message specifications (2003).

[18] U. Paul, A. P. Subramanian, M. M. Buddhikot, S. R. Das, Understanding traffic dynamics in cellular data networks, in: 2011 Proceedings IEEE INFOCOM, 2011, pp. 882–890.

[19] GSM Association, The Mobile Economy 2018, Tech. rep., GSM Association (2018).

[20] A. Vesselkov, A. Rikkonen, H. Hämmäinen, Mobile handset population in finland 2005-2013, Tech. rep., Aalto University, Department of Communications and Networking (01 2014).

[21] Y. Cao, L. Wang, Automatic selection of t-sne perplexity, arXiv preprint arXiv:1708.03229 (2017).

[22] S. J. Taylor, B. Letham, Forecasting at scale, The American Statistician 72 (1) (2018) 37–45. `doi:10.1080/00031305.2017.1380080`.

[23] B. Carpenter, A. Gelman, M. D. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. Brubaker, J. Guo, P. Li, A. Riddell, Stan: A probabilistic programming language, Journal of statistical software 76 (1) (2017).

[24] V. Melnykov, Clickclust: An r package for model-based clustering of categorical sequences, Journal of Statistical Software, Articles 74 (9) (2016) 1–34. `doi:10.18637/jss.v074.i09`.

[25] C. Song, Z. Qu, N. Blumm, A.-L. Barabási, Limits of predictability in human mobility, Science 327 (5968) (2010) 1018–1021.

[26] A. D. Back, D. Angus, J. Wiles, Determining the number of samples required to estimate entropy in natural sequences, IEEE Transactions on Information Theory 65 (7) (2019) 4345–4352. `doi:10.1109/TIT.2019.2898412`.