Making Sense of Customer Tickets in Cellular Networks

Yu Jin∗, Nick Duffield†, Alexandre Gerber†, Patrick Haffner†, Wen-Ling Hsu†, Guy Jacobson†,
Subhabrata Sen†, Shobha Venkataraman†, Zhi-Li Zhang∗

∗Computer Science Dept., University of Minnesota †AT&T Labs - Research

Abstract—Effective management of large-scale cellular data networks is critical to meet customer demands and expectations. Customer calls for technical support provide direct indication as to the issues and problems customers encounter. In this paper we study the customer tickets – free-text recordings and classifications by customer support agents – collected at a large cellular network provider, with two inter-related goals: i) to characterize and understand the major factors which lead to customers to call and seek support; and ii) to utilize such customer tickets to help identify potential network problems. For this purpose, we develop a novel statistical approach to model customer call rates which account for customer-side factors (e.g., user tenure and handset types) as well as geo-locations. We show that most calls are due to customer-side factors and can be well captured by the model. Furthermore, we also demonstrate that location-specific deviations from the model provide a good indicator of potential network-side issues. The latter is corroborated with the detailed analysis of customer tickets and other independent data sources (non-ticket customer feedback and network performance data).

I. INTRODUCTION

With the rapid growth in mobile voice and data services, effective management of large-scale cellular data networks is critical to meet customer demands and expectations. Due to the vast complexity involved, issues and problems may occur in a number of different places, e.g., mobile handsets, software and apps running on the handsets, or within the cellular network infrastructure – the latter itself spans large geographical regions, consisting of thousands of cell towers, radio spectrum access controllers, and a whole gamut of other network elements and servers, and supporting millions of users. Identifying and pinpointing – not to mention troubleshooting – these problems and issues can be an extremely challenging task. In addition, the increasing diversity and frequent roll-outs of new mobile handsets and devices (various “not-so-smart” and smart phones, e-readers, tablet computers, laptops, etc.), together with more complex and sophisticated software, apps and services running on them, further compound this task: from the customer perspective, added functionality in new mobile devices not only increases customer expectation of cellular services, but also requires increasing knowledge and sophistication on the customer side; from the cellular provider perspective, new mobile devices and services not only add new demands on the cellular network, but also complicate the task of separating issues and problems occurring at the customer side, e.g., due to handset or software issues, from those on the network side, e.g., network coverage or congestion related performance issues.

As a valuable source of information, customer-initiated feedback, e.g., calls to customer support lines, provides first-hand indication as to the issues and problems that customers encounter. These calls are typically recorded by customer (support) agents in the forms of customer tickets – free-text recordings of the conversations as well as classifications of call reasons and resolutions by customer agents. In this paper, we collect and systematically study the customer tickets over a 6-month time period at one of the largest cellular network service providers in the United States. Our goal is two-fold: i) to characterize and understand the major factors which lead customers to call and seek support – in particular, we are interested in separating customer-side factors from the network-side; and ii) to utilize such customer tickets to help identify potential network-side issues and problems.

As will be expanded on further in later sections, relying solely on the free-text description and/or customer agent classification contained in customer tickets to identify issues and problems, especially to separate those on the customer-side from the network-side, can be highly unreliable. Instead, we take a novel statistics-based, “semantic-free” approach to model and track customer call rates – the percentage of customers calling over an appropriately chosen time window, say, a week – over time, and to account for various factors affecting the call rates, e.g., such as customer-side factors (e.g., user tenure, service plans, and handset types) as well as geo-locations at various granularities (e.g., state, metro, radio network controllers (RNCs) or cell towers). The intuition here is that we use geo-locations (at various granularities) as proxies of network segments and elements: a location with a persistently high call rate can be a good indicator of potential chronic network-side issues and problems (e.g., congestion or poor coverage) at that location; on the other hand, an increase in call rates that are not location-specific is less likely network-related, and more likely caused by customer-side issues and problems (e.g., mobile devices, software, etc.). Mobility, however, poses a challenge in associating customers with locations. Customers often move around within the cellular networks, but the customer tickets themselves do not contain enough information to allow inference of which location the customer is complaining about. To circumvent this difficulty, we utilize another source of data collected within the cellular network (the GPRS Tunneling Protocol Control (GTP-C) messages, see Section V) to characterize the mobility of customers, and devise an effective method to associate customer tickets with locations where the reported problems may have happened.
Using the approach outlined above, we conduct a comprehensive study to analyze various customer-side factors, and correlate them with customer call rates at various locations. We build a statistical model to account for customer-side factors such as user tenure and device types. We show that most calls are due to customer-side factors and can be well captured by the model. Furthermore, we devise a novel method to detect locations with higher customer call rates that deviate from the model prediction. Through the detailed analysis of customer tickets as well as corroboration using other independent information sources: non-ticket customer feedback and network performance data (details in Section VI), we demonstrate that such location-specific deviations from the model are indeed excellent indicators of potential network-side issues.

The remainder of this paper is organized as follows: Section II overviews the cellular network architecture, and customer tickets that we use in the study. In Section III we motivate and argue for the semantics-free, statistical approach for characterizing customer tickets or call rates, and discuss the basic model and overall methodology. Section IV and Section V lay the foundations for the proposed technique by studying the correlation of call rates with various customer-side factors and characterizing customer mobility, respectively. Network-side problem detection using our model and its evaluation are presented in Section VI. Section VII discusses related works, and Section VIII concludes the paper.

II. BACKGROUND

We provide a quick overview of the cellular network under study, and briefly describe customer tickets. The datasets used in this paper are presented at the end.

Cellular Network Overview. The cellular network under study uses primarily UMTS (Universal Mobile Telecommunication System), a popular 3G mobile communication technology supporting both voice and data services. Fig. 1 depicts the key components in a typical UMTS network: When making a voice call or accessing a data service, a mobile device directly communicates with a cell tower or node-B, which forwards the voice/data traffic to a Radio Network Controller (RNC). In case of mobile voice, the RNC delivers the voice traffic toward the PSTN or ISDN telephone network, through a Mobile Switching Center (MSC) server. In case of mobile data, the RNC delivers the data service request to a Serving GPRS Support Node (SGSN), which establishes a tunnel with a Gateway GPRS Support Node (GGSN) using GPRS Tunneling Protocol (GTP), through which the data enters the IP network (and the public Internet). The UMTS network has a hierarchical structure: where each RNC controls and communicates with multiple node-Bs, and one SGSN serves multiple RNCs. UMTS offers a data downloading speed up to 2 Mbps. The UMTS network under study also interoperates with a number of network elements from the legacy 2G/2.5G GSM/GPRS/EDGE systems, which follow a similar architecture but have a much lower performance. Our study covers the entire network across all these technologies. For ease of exposition however, throughout the paper, we shall use 3G terminology to refer to all the elements in this network.

Customer Tickets. When a customer calls the customer service/technical support help line, the customer agent handling the call generates a customer ticket to record the conversation between the customer and the agent. The phone call may be transferred further to another department, e.g., equipment support or software support, for further problem resolution and a separate customer ticket is issued in this case. A ticket contains the time for the phone call and the entire conversation is recorded in a free-text format. In addition, each ticket is annotated with a call reason and a resolution summary, both of which are selected by the customer agent from a set of predefined categories, indicating the main problem reported/or the question asked by the customer and the resolution of the tickets given by that agent, respectively. Customers may call for a variety of reasons. A large majority of calls are non-technical related, e.g., questions about billing, service contracts, or how to use certain service or handset features, and so forth. Sometimes customers call when experiencing certain technical problems, e.g., unable to connect to the network, or access certain service features. These technical-related customer tickets are what we are interested in studying and making sense of.

Datasets. Our study is based on the customer tickets received and collected during a 6-month period. To assist our analysis, other relevant information such as customer tenure (i.e., how long a customer has been a subscriber), mobile device type, and so forth are also used – we emphasize here that no customer private information is used in our analysis and all customer identities are anonymized before any analysis is conducted. Similarly, to adhere to the confidentiality under which we had access to the data, at places, we present normalized views of our results while retaining the scientifically relevant bits. Additional information sources such as GTP control (GTP-C) messages at all GGSNs and passive network performance measurement data are used either for our analysis or for corroborating results of our network problem detection approach (see Sections V and VI for details).

III. CHARACTERIZING CUSTOM TICKETS: CALL RATES AND OVERALL METHODOLOGY

In this section, we first present a simple method to filter out obvious non-technical customer tickets, and argue why it may be hard to further classify the customer tickets based purely on customer agent classification or the free-text records.
The characteristics of mobility customer tickets. In both (a) and (b), the y-axis is normalized by dividing each value by \min(y).

**A. “Technical” Customer Tickets**

Most “non-technical” tickets can be easily identified and filtered, using customer agent classification (call reason and/or resolution summary, e.g., billing, contract, usage, etc.) and/or certain key words contained in the free text, e.g., plan, payment, bill and account. For the remaining tickets, further separation based solely on customer agent classification or keywords in the free text can be difficult and unreliable. For example, a customer may call and complain about network connectivity problems; however, the true problem may later be found to be certain service features not properly configured at the user handset. However, the customer agent may still simply classify the call reason and resolution as is, namely, “network connectivity” problem. The free text in the ticket may not help much either.

As an example, consider the keyword connect, which might, on the first thought, suggest problems on the network-side. We extract all tickets containing this key word, and classify them based on call reasons and call resolutions, respectively. Besides the tickets generated for customers calling about the “connectivity” issues containing the keyword connect, tickets generated for customers calling about handset (equipment), feature and software issues often also contain the same keyword. More interesting, apart from a portion of these calls being further transferred to next-levels of technical support as the resolution, the remaining top 10 resolutions for these tickets are usage (e.g., instructions on properly using some equipment), equipment (e.g., request to exchange a handset), or feature (e.g., activation or proper configuration of a service feature). This example also illustrates that it is difficult to depend upon just the call reason, resolution, or keywords in the free-text in tickets to reliably identify and separate the customer-side problems (e.g., handset malfunction or misconfiguration of software or apps) from the network-side problems (e.g., poor network coverage or network congestion). Clearly, the usefulness and accuracy of the free-text describing the problem a customer encounters and the customer agent classification (call reason and resolution) hinges on the sophistication of customers and expertise of the customer agents.

Instead of relying on the customer agent classification or keywords in the free-text in the tickets and examining individual tickets, in this paper we take a statistics-based, “semantics-free” approach: we study and characterize statistical properties of tickets across different factors (e.g., geographical locations, device types, etc.), and build statistical models to help understand the correlations between customer call rates and these factors – in particular, we use them to help identify potential network-side problems. We use the “semantics” of tickets (call reasons, resolutions or keywords in the free-texts) only for the purpose of corroboration and validation of our results. For the remainder of the paper, we consider the collection of tickets after only removing those “non-technical” tickets that can be easily and reliably identified. For convenience, we refer to this collection as “technical” tickets.

**B. Call Rates and Their Distributions across Geolocations**

Having identified technical tickets, we now mine for meaningful patterns over time and locations. We then prune out patterns which can be easily explained, so that we can focus on hard-to-detect network problems.

We show the number of technical tickets received each day in Fig. 2(a) over the six months observation period. The weekly average of the tickets is shown as a dashed curve\(^1\). We observe that tickets have a strong weekly pattern, where more tickets are received during the business days (Monday to Friday) and much fewer tickets come over the weekends. The number of received tickets remains stable up to 120 days and shows an ascending trend afterwards. To eliminate the day-of-week effect, we look at the weekly ticket aggregate. Fig. 2(b) displays the number of tickets received (dashed curve) and the number of unique customers who have issued tickets (solid curve) each week. We observe that the number of tickets is always larger than the number of customers, implying multiple tickets from some customers. Further investigation of

\(^1\)In both Fig. 2(a) and (b), the drop of tickets is caused by missing tickets from one particular day in the 12th week.
the tickets from these customers reveals that around 40% of the tickets come within 15 minutes after the immediate preceding ticket and around 60% of the tickets arrive no more than 1 day after. From the free-text content in the tickets, we find that most of these tickets are repeat tickets, e.g., customers call multiple times until the problem is resolved, calls get disconnected due to various reasons, or calls are transferred among different departments. The number of repeat tickets depend heavily on customers, e.g., if the customer needs the service immediately, she may complain multiple times in a short time period to get problem resolved earlier. To address the potential bias introduced by these repeat tickets, in this paper we track the time series of the customer call rate instead of the number of tickets. The customer call rate is defined as the proportion of customers who have issued at least one ticket within an observation time period $T$, where we set $T = 1$ week to address both the daily and weekly effect.

To understand the increase in call rate at certain weeks, we investigate on call rates at different states in the US (Fig. 2(c), Section V explains the details of mapping customers to different locations using GTP-C messages), where the $x$-axis shows the time (weeks) and the $y$-axis represents the ID of the 50 states. Each point $(x, y)$ stands for the $x$-week’s call rate at state $y$. A darker color corresponds to a higher call rate. For ease of visualization, we number the states on the $y$-axis in decreasing order of the average state-level call rate.

The call rates show significant variation across states (see Fig. 2(c)). We observe an universal increase in call rate across all states from the 23rd week to the 25th week. Investigation into the tickets reveals that a new version of a very popular smartphone device was released at the beginning of the 23rd week and the increase of the call rate was mainly caused by customers who received this device. In addition, some states show high call rate at certain weeks (dark points on the plot) and a few states (the top rows on the graph) exhibit persistently higher call rates than the rest of states. No customer side factors could be identified as responsible for such regional differences, which implies it might be the artifact of either network outage or potential chronic problems at these areas.

In summary, we identified as a robust measure of customer complaints for a given location the weekly customer call rates. A first attempt to explain the variations in these call rates suggests that customer related factors are well-defined, easily quantified and provide simple explanations to temporal variations. On the other hand, network problems which could explain variations over locations are latent and hard to detect.

C. Basic Model and Overall Methodology

Our key idea is to model customer call rate purely using customer related factors. If the model does not fit the observed call rate well, the difference between the model and the real call rate can be explained by potential network problems. As we have observed in Fig. 2(c), network problems cause unexpected fluctuations in the call rate as a function of the location, not the time. Therefore, our model will fix week $t$ and examine the call rate given the location for that week.

Let $\mathcal{U}_t$ denote the set of customers in a cellular network at the beginning of week $t$ and $\mathcal{L}$ denote the locations across the network. We note that a location $l \in \mathcal{L}$ may refer to a real-world geographic location, such as a city or a state, but it may also correspond to a network element, such as a node-B or an RNC. The observed customer call rate $P(c|l)$ at $l$ can be expressed as follows:

$$P(c|l) = \frac{1}{P(l)} \sum_{u \in \mathcal{U}_t} P(c|u,l)P(l|u)P(u),$$

where $P(c|u,l)$ is in fact an indicator function with $P(c|u,l) = 1$ if $u$ has issued ticket regarding $l$ and $P(c|u,l) = 0$ otherwise. $P(l|u)$ stands for the proportion of time that $u$ spends at $l$. $P(l)$ stands for the expected number of customers that appear at $l$ and $P(u) = 1/|\mathcal{U}_t|$ is a prior identical for all users. We note that in Eq. 1, the customer related factors and the network related problems are both captured by $P(c|u,l)$.

Our second model assumes that only customer related factors determine the customer call rate: $P(c|u,l) = P(\hat{c}|u)$. While this location-independence assumption is obviously not true, our goal is precisely to pinpoint the situations where it is broken. $P(\hat{c}|u)$ can be further approximated as $P(\hat{c}|f_t(u))$ where $f_t(u)$ are the customer related factors associated with $u$ at week $t$ (user tenure, device, etc.). Replacing $P(c|u,l)$ in Eq. 1 gives us the expected call rate at $l$ given the location-independence assumption.

$$P(\hat{c}|l) = \frac{1}{P(l)} \sum_{u} P(\hat{c}|f_t(u))P(l|u)P(u),$$

Comparing these two models helps us understand how various customer related factors affect the call rate. If $|P(\hat{c}|l) - P(c|l)| < \delta$, where $\delta$ is a small constant, the network related problems contribute little to the customer call rate. On the other hand, we identify a network related problem (e.g., an outage problem) if $P(c|l)$ is significantly larger than $P(\hat{c}|l)$, and the problem is likely to be chronic if such difference is persistent over a long time period.

In Section IV, we present a comprehensive study of dominant customer related factors and interpret how they affect the call rate. This study provides us a way to estimate $P(\hat{c}|u) \approx P(\hat{c}|f_t(u))$ using these dominant factors. Due to the fact that customers are moving around in the cellular network, we present a method for estimating $P(l|u)$ in Section V by tracking GTP-C messages. Combining these results, we use the model comparison to detect locations with chronic network problems and present our results in Section VI.

IV. Correlating Call Rate with Customer Factors

Among all the available customer related factors in the customer profile dataset, we have identified two factors that have significant correlation with the call rate: user tenure and device. In this section, we interpret these factors and show how they impact the customer call rate.
A. Impact of User Tenure on Call Rate

User tenure is defined as the number of weeks a customer stays in the service since the registration time. Taking one particular week $T$, we show the user tenure vs. the call rate in Fig. 3(a), where the $x$-axis represents the specific user tenure and the $y$-axis stands for the call rate of all the customers with a tenure $x$ weeks at the beginning of the week $T$.

We observe in Fig. 3(a) that a customer tends to have a much higher call rate when she initially enrolls in the service. We summarize the top resolutions that are associated with these new customers (with a tenure of no more than 4 weeks) in the row marked "beginning service" in Table I. Most of these resolutions are unique to new customers, such as porting from other ISPs and check service availability, etc. In addition, new customers also tend to ask questions regarding system configuration or to request for equipment exchange, etc.

**TABLE I**

<table>
<thead>
<tr>
<th>Patterns</th>
<th>Dominant resolutions</th>
</tr>
</thead>
<tbody>
<tr>
<td>beginning</td>
<td>porting, check eligibility, configuration, equipment return/exchange</td>
</tr>
<tr>
<td>annual</td>
<td>voice and data feature inquiries and modification</td>
</tr>
<tr>
<td>biennial</td>
<td>SIM change, unlock device, IMEI change</td>
</tr>
</tbody>
</table>

In addition to the beginning-of-the-service pattern, we observe other patterns that contribute to the periodic increase of the call rate. We apply Discrete Fourier Transform (DFT) to the curve on Fig. 3(a) and demonstrate the result in Fig. 3(b). The $x$-axis represents the length of the period and the $y$-axis shows the significance of the sinusoidal curve with a period of $x$ from the DFT result. The annual (around 50 weeks) and biennial (around 100 weeks) periods appear to be more significant from the DFT result (marked with red ovals). Let $P(r|x)$ be the fraction of tickets from customers with a tenure $x$ that are annotated with a particular resolution $r$. To explain the reasons for the annual pattern, we study the linear correlation between $P(r|x)$ and the sinusoidal curve with a period of 50 weeks. The linear correlation is measured using correlation coefficient ($\text{corr}$). Between two variables $u$ and $v$, the $\text{corr}$ is defined as follows.

$$\text{corr}_{u,v} := \frac{E[(u - \bar{u})(v - \bar{v})]}{\sqrt{\text{var}(u) \cdot \text{var}(v)}}$$

$\text{corr}_{u,v} > 0$ if the two variables are positively correlated and $\text{corr}_{u,v} = 0$ if $u$ and $v$ are independent. We rank resolutions based on the values of the corresponding correlation coefficients, and use the top ones to explain the annual pattern. Similarly, we explain the biennial pattern by correlating $P(r|x)$ with the sinusoid of a 100-week period.

We summarize the top resolutions that contribute to the annual pattern and the biennial pattern in Table I (row 3 and 4). The annual pattern is mainly caused by customer inquiries and modification to different kinds of service features, such as road side assistant and sophisticated IP based voicemail, etc. This reveals the fact that when a customer stays in the service for a relatively long time period (e.g., a year), she is likely to become interested in trying new features provided by the ISP. In comparison, the biennial patterns are mainly due to the typical two-year contract between the ISP and the mobile customers. When a customer reaches the end of the contract, she is more likely to request for a free key to unlock the device or change the current device or SIM card.

B. Impact of Customer Device on Call Rate

The second customer related factor that we analyze is the customer device. For our study, we choose the top 15 devices from 3 different categories: 6 smartphones (denoted as “SP-X”, all with mandatory data plans), 5 traditional phones (denoted as “TP-X”, all with optional data plans) and 4 laptop card devices (denoted as “LC-X”, all with mandatory data plans). In Fig. 3(c), we illustrate the call rates corresponding to different devices over one-week period across all customers. The dotted lines show the average call rates for the three categories of devices, respectively.

We observe that the call rate varies across different categories of devices and also within each category. For example, smartphones and traditional phones show high average call rates, which are 2 times larger than that of the laptop cards. Among all laptop cards, LC-2 has a much smaller call rate compared with other laptop cards (e.g., 1/4 of the call rate of LC-1). In the following, we investigate on the reasons behind the very different call rates associated with these laptop cards.

We search for significant correlations between devices $d$ and ticket resolutions $r$. Our basic idea is to identify the $(d, r)$ pairs such that $P(d, r) > P(d)P(r)$ and the difference is statistically significant. In other words, the correlation is defined as the (positive) deviation from the independent assumption on
We use the normalized Pearson’s residual to extract such correlations, which is defined as:

\[ e_{d,r} := \frac{n_{d,r} - \hat{\mu}_{d,r}}{\sqrt{\hat{\mu}_{d,r}(1 - P(d))(1 - P(r))}}, \]

where \( n_{d,r} \) represents the number of observed tickets associated with resolution \( r \) and device \( d \) and \( \hat{\mu}_{d,r} := N \cdot P(d) \cdot P(r) \) represents the expected number of tickets associated with \( d \) and \( r \) given the assumption that \( d \) and \( r \) are independent. \( N \) stands for the total number of tickets received during the observation time period. Because \( e_{d,r} \) follows the standard normal distribution, a significant correlation between \( d \) and \( r \) is identified if \( e_{d,r} > 3 \) (with a \( P \)-value less than 0.05).

We summarize the resolutions that show significant correlations with different laptop card devices in Table II. From other information sources, we know that LC-2 and LC-3 are essentially the same device with different names. LC-2 is mainly provided for business customers and LC-3 is used by non-business customers. The difference in the user population results in striking differences in the dominant ticket resolutions. The LC-3 device had a software problem and many customers called for technical support on installation and configuration of the connection manager software. Though we expect that LC-2 should also exhibit a similar problem, the software related resolutions show no dominance for LC-2. This is because most companies maintain their own technical support team which resolves such software issue for their employees. Therefore, the dominant resolutions associated with LC-2 are service cancellation and SIM card change due to employment changes, since a customer often has to terminate the contract if she switches jobs.

### Table II

**Dominant Resolutions for Different Laptop Card Device.**

<table>
<thead>
<tr>
<th>Device</th>
<th>Dominant resolutions</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>LC-1</td>
<td>equipment exchange</td>
<td>poor device quality</td>
</tr>
<tr>
<td>LC-2</td>
<td>cancel service or change SIM card due to unemployment</td>
<td>enterprise device</td>
</tr>
<tr>
<td>LC-3</td>
<td>download/install/use/configure connection manager software</td>
<td>software related issue</td>
</tr>
</tbody>
</table>

The laptop card LC-1 had a similar call rate to LC-3, however, the dominant resolutions were also quite different. Unlike the software problem reported from LC-3 customers, the LC-1 customers mainly called for equipment exchanges. Exploration of customer reviews for LC-1 shows that this particular device tended to have a poor quality and hence a high exchange rate.

### C. Modeling Call Rate Given Customer Related Factors

So far, we have shown how various customer related factors contribute to the variation of the call rate. From the history of tickets, we can directly model \( P(\hat{e} | f_i(u)) \) using a multinomial distribution, by counting the call rate given all combinations of the values of these customer related factors. However, the model constructed in this way contains too many parameters and hence does not generalize. Instead, in this paper, we construct a much simpler discriminative linear model \( P(\hat{e} | f_i(u)) = g(f_i(u)) \), where \( g \) is a linear function which combines various customer related factors with different weights. We use the Adaboost algorithm combined with logistic calibration [1] to automatically learn the function \( g \) from tickets, which has the advantage of automatically determining the best partition of continuous variables, e.g., user tenure.

### V. User Mobility and Customer Tickets

In order to learn \( p(l | u) \), we need to study the user mobility patterns in the network. The analysis in this section serves the purpose of mapping customers to locations where the reported problems are likely to have occurred. Such a mapping, though crucial for problem diagnosis and troubleshooting, is not recorded in customer tickets and has to be retrieved from other data sources.

#### A. Mapping Customers to Locations using GTP-C Messages

The most accurate way to associate customers with locations is to place probes at network elements closer to the mobile devices, e.g., node-Bs and RNCs, and track the elements that customers connect to. However, such information is often unavailable given the large number of low-level network elements and such a feature is not guaranteed on all network devices. In this paper, we instead use GTP-C messages as a proxy to infer the approximate mapping between customers and network elements.

When a customer wants to access the cellular network data service, a **GTP Create** message is sent to the GGSN (recall Fig. 1) to establish a GTP tunnel for the current GTP session, which contains the Location Area Code (LAC) and Cell ID (CID) of the node-B that is currently serving the customer. A **GTP Update** message will be sent to GGSN to update the latest LAC and CID when the customer travels beyond a certain distance and a SGSN handover happens. When the customer finishes using the data service, a **GTP Delete** message is sent to GGSN to remove the GTP tunnel and hence terminate the GTP session. By tracking the GTP-C messages, we are able to associate customers with locations.

We note that the location mapping is only an approximation in that when a customer does not travel far enough to trigger a SGSN handover, no GTP Update message will be sent and the customer will always be mapped to the location where the GTP session starts. However, there are situations when new GTP sessions will be initiated, for example, a customer reboots the device or the current GTP session is cut off due to certain reasons (e.g., the customer drives through a long tunnel). In addition, when the customer switches between 2G and 3G, a new GTP session is created. Moreover, due to the wide deployment of public WiFi hot spots, a customer will switch more frequently between 3G and WiFi, resulting in new GTP sessions. All these factors provide us with a relatively large number of GTP-C messages to ensure good accuracy in the location mapping.
primary locations for individual customers is as follows. The locations where customers spend more time, which we refer to as primary locations, when a ticket is issued by the customer, how can we associate with a customer. When a ticket is received from the customer, we consider the chance that the ticket is related to a specific location. In our case, we focus on customer tickets. We show the distribution of scores for these node-Bs and RNCs in Fig. 5(b) and (c), respectively. We observe that node-Bs, where the first node-B (top plot) has a higher observed call rate over time, indicating a potential chronic problem at that node-B. In comparison, the bottom plot in Fig. 5(a) shows a node-B in a relatively good condition, where the observed call rate is always below the expected call rate.

A. Experimental Results

We compute the observed call rate and the expected call rate for various network locations at different granularities, using Eq. 1 and Eq. 2 (Sec. III). Fig. 5(a) shows two example node-Bs, where the first node-B (top plot) has a higher observed call rate over time, indicating a potential chronic problem at that node-B. In comparison, the bottom plot in Fig. 5(a) shows a node-B in a relatively good condition, where the observed call rate is always below the expected call rate.

We define a chronic problem score as $\text{median}(P(c|l) - P(\hat{c}|l))$ as an indicator for identifying network locations with a persistently higher observed call rate (than the expected call rate), or, equivalently, locations with potential chronic problems. This score is designed to provide us a means of ranking all network locations based on the likelihood of having chronic problems. To ensure that each location we examine has sufficiently large population, we only focus on the node-Bs and RNCs which are among the primary locations for at least 100 customers. We show the distribution of scores for these node-Bs and RNCs in Fig. 5(b) and (c), respectively. We observe that only a small portion of node-Bs and RNCs are candidates for locations with potential chronic problems (with chronic problem scores greater than 0). In the following, we evaluate the detection results using a number of different data sources –
based on ticket semantic information, a second independently-generated customer feedback data set, and finally, network performance measurements.

B. Evaluations using Customer Side Datasets

Evaluation using ticket call reasons. Our first evaluation is based on the dominant ticket call reasons associated with the locations whose chronic problem scores are greater than 0. Using a technique similar to the approach used for Table II, we use the normalized Pearson’s residuals for ranking the call reasons with the highest correlations with the call rate.

<table>
<thead>
<tr>
<th>Dominant call reasons</th>
<th>Percentage of calls</th>
</tr>
</thead>
<tbody>
<tr>
<td>Network problem?</td>
<td>Yes</td>
</tr>
<tr>
<td>connectivity</td>
<td>20%</td>
</tr>
<tr>
<td>equipment</td>
<td>47%</td>
</tr>
<tr>
<td>feature</td>
<td>17%</td>
</tr>
<tr>
<td>other</td>
<td>16%</td>
</tr>
</tbody>
</table>

In Table III, we display the composition of the top 4 call reasons for the locations with and without potential chronic network problems. We observe that, at these detected locations, there are far more dominant call reasons related to network connectivity issues (20%), when compared to 5% for the other locations. In addition, over 47% of the call reasons are related to equipment problems at these detected locations, compared to 10% for the rest of the locations. As we have pointed out earlier in the paper, it is sometimes difficult for the customer agents to differentiate equipment related problems from network related problems. We analyzed 100 randomly selected tickets related to equipment problems and around 30% of them were due to problems of sending/receiving data/SMS or frequently dropped calls. We suspect that such equipment problems may also be caused by chronic network problems at these detected locations.

Evaluation using App messages. Our second evaluation is based on the messages from a customer side application (referred to as App) on one of the most popular smartphone devices (referred to as SP−M). App serves as an independent way other than customer tickets for customers to report problems. When a customer encounters a certain problem, she can select from one of the predefined problem categories and send a message to the ISP to report the problem. The message contains the serving LAC and CID when the message is sent, which enables an accurate mapping between a problem and the related network locations. We collect all the App messages received during the same 6 month period as the customer ticket dataset. We note that, compared to the App messages, our detection technique based on customer tickets has a much wider coverage, since it is not restricted by the SP−M population at different locations.

A key difference between App messages and customer tickets is that it is very easy for customers to impulsively send App messages, even in non-primary locations. In fact, we find that only 20% of the App messages are regarding primary locations. For this reason, to ensure a fair comparison between the call rate and the App message rate at a particular location, we only select App messages whose associated network location are among the senders’ primary locations. In addition, since it is difficult to estimate the customers who have installed App, we use the entire SP−M customer population as the base of App users, effectively making the assumption that App users are uniformly distributed among all SP−M customers across different locations.

We can now calculate the App call rate (specifically, a call rate that is proportional to the true App call rate) as the percentage of SP−M customers who have sent at least one App message given an observation time period $T$. A location is considered to have a potential network problem if the corresponding App call rate is higher than other locations. Again, we only look at locations which are among the primary locations for at least 100 SP−M customers. Fig. 5(d) demonstrates the correlation between the App call rate and the chronic problem scores for RNCs. We divide locations according to the scores into equal-sized bins. For RNCs inside each bin, we report the median of their App call rates.

We observe a strong correlation between the App call rate and the chronic problem score. For RNCs with scores greater than 0, we find the corresponding App call rate is around 3 times the App call rate for the rest of the RNCs. In addition, the median App call rate drops as the score becomes lower.

C. Evaluation using Network Data

We extracted 20 RNCs with the highest chronic problem scores to analyze with standard network measurements. On examining these RNCs manually, we noticed that most of them –16 of the 20 – turned out to be part of the previous generation EDGE network, instead of the UMTS network. We also find that a significant fraction of the user devices used at these RNCs were 3G devices, and these devices generated a substantial amount of the traffic of these RNCs, on the EDGE network. It is thus likely that these calls were the consequence of customers who own 3G devices and therefore, expect to receive speeds typical of the UMTS network, but instead receive lower speeds and worse performance because their device only uses the EDGE network.

To get a more quantitative measure of their relative performance, we also examined round trip times (RTTs) of the devices at these RNCs over a period of 4 days. We use passively monitored RTTs of all the TCP flows across the entire network (as is standard, we define RTT to be the time between the SYN, SYN-ACK and ACK packets), and these RTTs are mapped back to the different RNCs. We found that at nearly all of these RNCs, devices experienced extremely high RTTs – in particular, the median RTTs measured at these RNCs were in the highest 15th percentile of median RTTs, even with respect to other RTT measurements on the EDGE network, i.e., 90% of the RNCs on the EDGE network have lower median RTTs than these RNCs. This indicates, again, that these RNCs do indeed seem to experience significantly...
worse performance than most of the RNCs on the cellular network, which supports our experimental results.

VII. RELATED WORK

There is a rich literature in detecting and troubleshooting network problems in large networks. A majority of work focuses on detecting, locating or trouble-shooting wired/wireless IP data network problems using passive or active network measurement data, e.g., via expert rule-based inference [2]–[4] or machine-learning techniques [5]–[12], or via analysis and inference of correlation and dependency among network elements, entities and events [13]–[17]. In some cases, other data sources are also used. For instance, GIZA [18] tracks bursts of customer tickets and correlates these ticket bursts with other network events to locate the problem. MyExperience [19] combines passive logging of device usage, user context, and environmental sensor readings with in situ customer feedback to support studies of mobile technology usage and evaluation. Our work differs from these earlier studies in that we directly analyze and characterize customer tickets to understand the major factors affecting customer call rates, and develop novel statistical models and approaches to explicitly account for and separate customer-side and network-wide factors. Furthermore, we demonstrate that location-specific deviations from model prediction can help point to and locate network problems. Comparing to network measurement data, the volume of customer tickets is far less. Hence our method provides a useful tool to help zero in on network problems that actually affect customers, and when combining with network measurement data, can help diagnose and trouble-shoot these problems more quickly and effectively — this is part of an on-going work that we are conducting.

VIII. CONCLUSION

In this paper, we presented comprehensive analyses of customer tickets received in a large cellular network. We showed that the probability of a caller reporting a particular problem is affected by various customer-side factors such as user tenure and device type and network side problems. By explicitly addressing these customer-side factors and taking into account user mobility in the cellular network, we devised a novel approach to use customer tickets as a front-line to pinpoint locations with potential chronic network problems. Evaluation using independent data sources from both customer side and network side corroborate that these identified network locations are associated with certain network related problems, which inevitably lead to a persistent high call rate at these areas. Our future work will mainly concentrate on correlating the detection results with network performance metrics to troubleshoot these chronic problems.

REFERENCES