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 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) and 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.**

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. 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 form 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.*

In this paper, 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 call rates, e.g., such as customer-side factors (e.g., user tenure 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 apply the proposed model to the 6-month ticket trace and detect locations with higher customer call rates that deviate from the model's prediction. Through detailed analysis of customer tickets as well as corroboration using non-ticket customer feedback (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 datasets that we use in the study. In Section III we motivate and argue for the semantics-free, statistical approach for characterizing customer call rates, and discuss the 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

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. The

UMTS network has a hierarchical structure: where each *Radio Network Controller* (RNC) controls multiple node-Bs, and one *Serving GPRS Support Node* (SGSN) serves multiple RNCs (see [1] for UMTS network details).

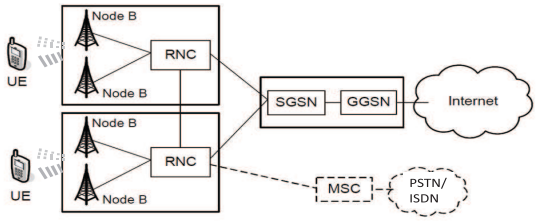

Fig. 1. UMTS network architecture.

Customer Tickets. When a customer calls the customer service help line, the customer agent handling the call generates a *customer ticket* to record the conversation with the customer. A ticket contains the time for the 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, etc. Sometimes customers call when experiencing certain *technical* problems, e.g., unable to connect to the network, etc. 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 *GPRS Tunneling Protocol Control* (GTP-C) messages at all *Gateway GPRS Support Nodes* (GGSNs) 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

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 can be difficult and unreliable.

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

To address the bias caused by repeat tickets from customers, our analysis is based on the time series of the *customer call rate*. 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, 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). 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 difference, which implies it might be the artifact of either network outage or potential chronic problems at these areas.

C. Basic Model and Overall Methodology

Our key idea is to model the 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

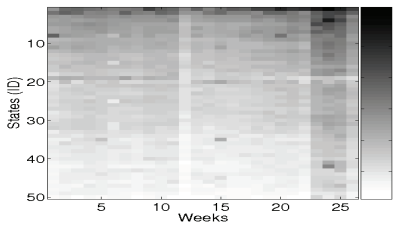

Fig. 2. Call rate breaks down by states.

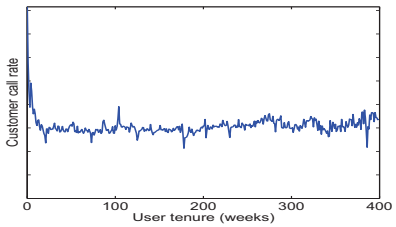

Fig. 3. User tenure vs. call rate.

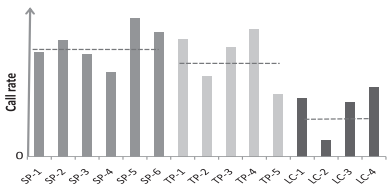

Fig. 4. Device vs. call rate.

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:

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

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.

$$P(\hat{c}|l) = \frac{1}{P(l)} \sum_{u} P(\hat{c}|f_t(u))P(l|u)P(u), \qquad (2)$$

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 δ 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, 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 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). Most of these resolutions are unique to new customers, such as porting from other ISPs and check service availability, etc. New customers also tend to ask questions regarding system configuration or to request for equipment exchange, etc[1].

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. 4, 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 than other laptop cards (e.g., 1/4 of the call rate of LC-1).

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

[1]We have also extracted annual and biennial patterns from Fig. 3, see [1].

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.

C. Modeling Call Rate Given Customer Related Factors

We have shown how customer related factors contribute to the variation of the call rate. From the history of tickets, we can directly model $P(\hat{c}|f_t(u))$ 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 $P(\hat{c}|f_t(u)) = g(f_t(u))$, 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.

A. Mapping Customers to Locations using GTP-C Messages

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 the 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 the 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.

B. Mapping Tickets to Locations

We manually investigate hundreds of tickets regarding connectivity problems, and find that the tickets are likely mapped to locations where customers spend more time (e.g., home location or work place), which we refer to as *primary locations* hereafter. Our technique for extracting primary locations for each customer is as follows.

Let $p_i := P(l_i|u)$ be the fraction of time that a customer u spends at location i, $1 \leq i \leq n$, where n is the total number of locations visited by the customer during the month, and hence $\sum_{i=1}^{n} p_i = 1$. We compute the *relative uncertainty* (RU) as $RU := -\sum_{i=1}^{n} p_i \log(p_i)/\log(n)$. A RU value above θ ($\theta = 0.8$ in the experiment) indicates that the customer spends roughly equal amount of time at all locations and hence no location is primary. Otherwise, we extract the location with the longest time usage as a primary location and compute the RU value for the remaining locations. The process iterates until all primary locations are extracted.

Since customers mainly stay at the primary locations and most tickets are regarding these primary locations, we only consider primary locations while mapping the tickets to locations. In particular, let $1 \leq i \leq K$ be the K primary locations associated with a customer. When a ticket is received from the customer, we consider the chance that the ticket is related to location j (p_j) equals $p_j/\sum_{i=1}^{K} p_i$ if j is among one of the K primary locations, and p_j equals 0 for a non-primary location.

VI. Detecting Chronic Network Problems

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 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 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 $median(P(c|l) - P(\hat{c}|l))$ as an indicator for identifying locations with a persistently higher observed call rate (than the expected call rate), or, equivalently, locations with potential chronic problems. This score provides us a means of ranking all network locations based on the likelihood of having chronic problems. To ensure that each location we examine has a sufficiently large population, we only focus on the node-Bs and RNCs which are among the primary locations for at least 100 customers.

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. We use the normalized Pearson's residuals for ranking the call reasons with the highest correlations with the call rate.

In Table I, 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%

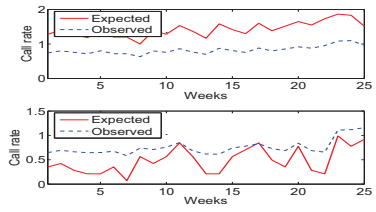

Fig. 5. Node-Bs of good/bad condition.

TABLE I
DOMINANT CALL REASONS ASSOCIATED WITH LOCATIONS OF POTENTIAL CHRONIC NETWORK PROBLEMS.

Dominant call reasons	Percentage of calls	
Network problem?	Yes	No
connectivity	20%	5%
equipment	47%	10%
feature	17%	3%
other	16%	82%

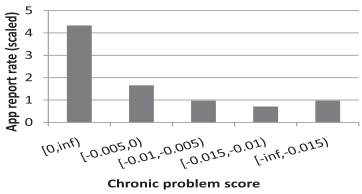

Fig. 6. Evaluation using App rate.

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 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. 6 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.

VII. Related Work

There is a rich literature in detecting and troubleshooting network problems in large networks. A majority of work focus 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] or machine-learning techniques [3], [4], or via inference of dependency among network elements, entities and events [5], [6], or correlating bursts of customer tickets with other network events [7]. Our work differs 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.

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 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 corroborate that these identified locations are associated with certain network related problems, which inevitably lead to a persistent high call rate at these areas.

Acknowledgement

The work is supported in part by the NSF grants CNS-0721510, CNS-0905037, CNS-1017647, the DTRA grant HDTRA1-09-1-0050, and the AT&T VURI grant.

References

[1] Y. Jin, N. Duffield, A. Gerber, P. Haffner, W.-L. Hsu, G. Jacobson, S. Sen, S. Venkataraman, and Z.-L. Zhang, "Making Sense of Customer Tickets in Cellular Networks," AT&T research, Tech. Rep., 2010.

[2] G. Khanna, M. Y. Cheng, P. Varadharajan, S. Bagchi, M. P. Correia, and P. J. Veríssimo, "Automated rule-based diagnosis through a distributed monitor system," *IEEE Trans. Dependable Secur. Comput.*, 2007.

[3] I. Cohen, M. Goldszmidt, T. Kelly, J. Symons, and J. S. Chase, "Correlating instrumentation data to system states: a building block for automated diagnosis and control," in *OSDI'04*, 2004.

[4] B. Aggarwal, R. Bhagwan, T. Das, S. Eswaran, V. N. Padmanabhan, and G. M. Voelker, "Netprints: diagnosing home network misconfigurations using shared knowledge," in *NSDI'09*, 2009.

[5] P. Bahl, R. Chandra, A. Greenberg, S. Kandula, D. A. Maltz, and M. Zhang, "Towards highly reliable enterprise network services via inference of multi-level dependencies," in *SIGCOMM'07*, 2007.

[6] S. Kandula, R. Mahajan, P. Verkaik, S. Agarwal, J. Padhye, and P. Bahl, "Detailed diagnosis in enterprise networks," in *SIGCOMM'09*, 2009.

[7] A. Mahimkar, Z. Ge, A. Shaikh, J. Wang, J. Yates, Y. Zhang, and Q. Zhao, "Towards automated performance diagnosis in a large iptv network," in *SIGCOMM'09*, 2009, pp. 231–242.