Understanding the Complexity of 3G UMTS Network Performance

Yingying Chen[†], Nick Duffield[*], Patrick Haffner[*], Wen-Ling Hsu[*], Guy Jacobson[*],
Yu Jin[*], Subhabrata Sen[*], Shobha Venkataraman[*], and Zhi-Li Zhang[†]
[*]AT&T Labs Research, [†]University of Minnesota

***Abstract*—With rapid growth in smart phones and mobile data, effectively managing cellular data networks is important in meeting user performance expectations. However, the scale, complexity and dynamics of a large 3G cellular network make it a challenging task to understand the diverse factors that affect its performance. In this paper we study the RNC (Radio Network Controller)-level performance in one of the largest cellular network carriers in US. Using large amount of datasets collected from various sources across the network and over time, we investigate the key factors that influence the network performance in terms of the round-trip times and loss rates (averaged over an hourly time scale). We start by performing the "first-order" property analysis to analyze the correlation and impact of each factor on the network performance. We then apply RuleFit – a powerful supervised machine learning tool that combines linear regression and decision trees – to develop models and analyze the relative importance of various factors in estimating and predicting the network performance. Our analysis culminates with the detection and diagnosis of both "transient" and "persistent" performance anomalies, with discussion on the complex interactions and differing effects of the various factors that may influence the 3G UMTS (Universal Mobile Telecommunications System) network performance.**

I. Introduction

The wide adoption of smart phones and other mobile devices such as smart tablets and e-readers has spurred rapid growth in mobile data. In order to meet user performance expectations and enhance user experiences, effectively managing cellular data networks is imperative: for example, quickly trouble-shooting performance issues as they arise, or adding additional capacity where the network elements are overloaded. Due to the sheer scale, complexity and dynamics of a typical large scale cellular data network, there is a myriad of diverse factors that may affect the performance of a cellular data network, from the types, geographical locations and coverage of various network elements (e.g., cell towers/NodeBs, radio network controllers (RNCs), IP gateways) within the network infrastructure, to the number of users served by each network element, the amount of traffic generated by the users, user usage patterns and behaviors (e.g., mixtures of dominant applications) across different times of days and weeks, to user handset side as well as the (application) server side issues. For instance, the numbers of NodeBs and sectors as well as the geographical coverage can vary significantly from one RNC to another, and the usage patterns and application mixes differ also markedly across the network and over time. Hence, teasing out each factor individually is not an easy task. Not only is there a vast array of diverse factors, but many of these factors are also intertwined, exerting disparate influences in different parts of the network. Furthermore, there may be *latent* factors that are not explicitly accounted for, or captured at all in the information and data that we can monitor and collect.

In this paper, utilizing the massive amount of performance and other data collected in one of the largest cellular network carriers in US, we set out to build *macroscopic* models that can help identify the various major factors that may (or may not) influence the network performance and assess their effects *across the network* and *over time* in a large UMTS cellular data network. This is as opposed to *microscopic* models that target specific network elements or users, e.g., to trouble-shoot (transient or persistent) performance issues experienced by certain network elements (e.g., cell towers) or users. Building effective microscopic models requires far more detailed, fine-grained (and often lower-level) measurement data. [1] Our approach is to first take several major classes of factors and perform network-wide correlation and other "first-order" analysis to understand how significantly each of these major factors *individually* influences the overall RNC-level network performance. This provides us with a baseline understanding of the individual effect (or lack thereof) that each of the major classes of factors has on the overall network performance.

Next, to examine the collective effect of various factors on the network performance and sort out their relative contributions, we apply RuleFit [1] – a powerful predictive learning method via rule-ensembles, combining both linear regression and decision trees. RuleFit provides both predictive and interpretative capabilities that are needed for our analysis. By aggregating as well as dividing the datasets along different dimensions (e.g., based on geographical locations such as states or along time such as days or weeks), we intelligently apply RuleFit to build macroscopic models to assess and dissect the various major factors that affect part or the whole network, as well as their impact on the network performance over time. For example, by comparing the model obtained using the datasets from all RNCs with those obtained from the RNCs located within certain geographical regions, we can identify the major factors that have persistent network-wide effect and uncover those factors that have marked impact at certain locales or at certain times. Furthermore, by examining the level of the overall contributions of all factors as well

[1]For instance, there are more than 500 low-level device counters associated with antennas, NodeBs, and RNCs. Making sense of these device counter and other low level statistics, especially how they relate to the overall network performance is a vast challenge. Apart from providing the network operators with a "big picture" view of network performance, our macroscopic models are developed also as a first step towards addressing this challenge. It serves as a guide to help the development of microscopic models using more detailed and low-level data for network element specific performance prognosis and problem diagnosis.

as the relative importance of each factor and how they differ across locales or over times, we can also infer whether there are potential latent factors in play, affecting the network performance in a way that cannot be quantified using the factors explicitly included in the models. Significant deviations from the model predictions may also signal anomalous events, and can be used for diagnosis purposes (see Sections V & VI).

While our study primarily focuses on *macroscopic* models, it facilitates the *microscopic* diagnosis of the network performance issues associated with certain part of the network or specific network elements by pointing to potential problematic features that are in play. Meanwhile, we believe that the methodology developed in this paper can be applied to other 3G network architectures, and possibly also the emerging 4G (LTE) cellular networks.

II. Background and DataSets

In this section, we give a brief overview of the architecture of the typical 3G UMTS network, and the datasets we collected for our study.

A. UMTS Network Overview

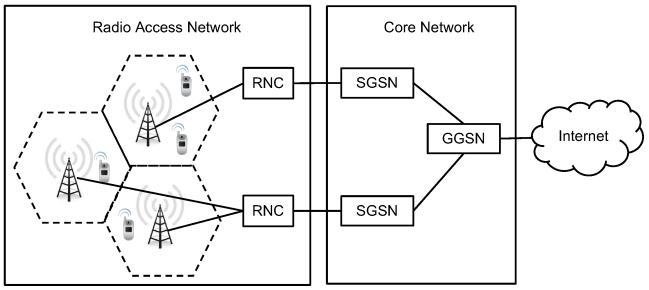

Fig. 1: UMTS network architecture

As illustrated in Fig 1, a typical UMTS network consists of two major entities, the UMTS Terrestrial Radio Access Network (UTRAN) and the core network. UTRAN consists of base stations (NodeB), and Radio Network Controllers (RNC) which are connected to and control a number of NodeBs. Each NodeB is typically configured with multiple sectors, e.g., 3 sectors (the common configuration) in three different directions, each covering 120 degree range. If there are more than three sectors associated with one NodeB, multiple sectors are overlapped on each direction, and are distinguished using different frequencies. Moreover, NodeB usually supports multi-carrier technology. Via software configuration, each sector can support multiple carriers in order to further increase its capacity. The core network is comprised of the Serving GPRS Support Nodes (SGSN) and the Gateway GPRS Support Nodes (GGSN). To connect to, say, a web server located in the outside Internet, a User Equipment (UE) will first contact the nearest NodeB. After receiving the user data access requests on one of its sectors, NodeB will handover the user requests to its upstream RNC, which further forwards the data service requests to an SGSN. In the core network, the SGSN establishes a tunnel with a GGSN, using the GPRS Tunneling Protocol. The data is carried as IP packets in the tunnel and finally reaches the web server in the external Internet.

B. Datasets

For this study, we combine datasets collected from a variety of sources in order to gain a more comprehensive view of the various factors that may influence the network performance. As mentioned in the introduction, these data sources include *static* information regarding the UMTS infrastructure, such as the GPS location (and zipcode) of each NodeB, the number of sectors (and carriers) per NodeB, the corresponding RNCs that NodeBs are associated with, the SGSNs that RNCs are connected to, the equipment type and vendor of each RNC, and so forth. To gain a sense of the population density and demographics in each base station/RNC coverage area, we also utilized the 2000 census data[2], which contains the land area coverage of each zip code. From these static data sources, we can estimate the geographic areas that are covered by each NodeB, RNC and SGSN, and get a sense of the population density and other demographic information (e.g., rural vs. small city vs. large metro, etc.). Therefore, these static factors can help us identify the impact from the geographical coverage, regional variations, the placement of the networking facilities, and other infrastructure-related issues.

There are two major sources of *dynamic* data that are used in our study. One is the IP traffic data collected periodically at each GGSN. All the measurements were computed at the RNC level, i.e. aggregated for all the users served by the same RNC, and at the granularity of an hour. The timestamp used throughout the paper refers to the local time of the RNC location. From the IP traffic data, we obtain the RTT and loss rate performance measurement data and the usage related statistics such as the number of bytes, flows, packets, average flow sizes, and so forth, as well as the application classifications and mixes (e.g., email, VoIP, streaming video, MMS, Appstore downloads) – traffic that cannot be classified is labelled *unknown*. RTT is measured for the end-to-end latency between the UE and the content provider that serves the content to the mobile device. Loss rate is estimated using the tcp-level packet retransmission rate. The other dynamic data source contains more lower level information, such as the total number of Radio Resource Control (RRC) [2] attempts served by each RNC at every hour. An RRC attempt indicates a connection establishment attempt between the UEs and the UTRAN, and therefore the #RRC attempts can be a good approximation of the #requests served at each RNC. The datasets collected span more than 6 months. However, as representative examples and for illustrative purposes, we will focus on the datasets collected during a two-week period in September 2011. To adhere to the confidentiality under which we had access to the data, at places, we present normalized views of our results while retaining scientifically relevant bits.

III. Problem Setting & Illustration

Due to the sheer scale, complexity and dynamics of the large 3G UMTS cellular network, there are a myriad of complex factors that may potentially affect the overall network performance. Some of these factors may depend on other factors, and interact with each other differently in different

[2]While the census data is almost 12 years old, the land area coverage per zip code does not change drastically over the years. Moreover, we do not make any direct conclusion based on the census data, but rather combine it with other data and techniques for our study.

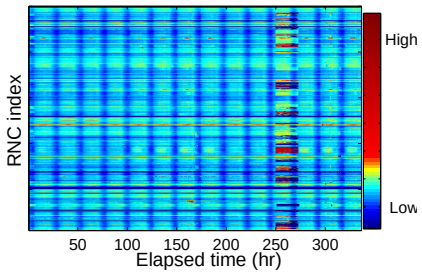

Fig. 2: Two-week RTT time series across all RNCs.

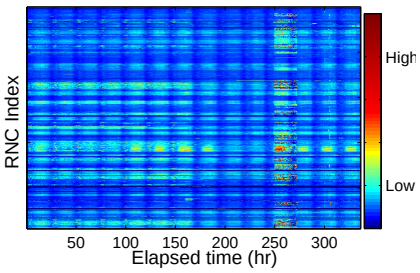

Fig. 3: Two-week loss rate time series across all RNCs.

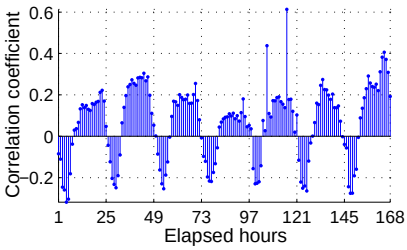

Fig. 4: Varying correlations between RTT and loss rate spatial variability.

parts of the networks. To identify the major factors and tease out the effect of each factor – and understand how they affect the overall network performance in different parts of the network and over time – are extremely challenging. In this paper, we consider RTT and loss rate (aggregated at the RNC-level and averaged hourly) as two key performance indicators. In Figs 2 & 3 (best viewed in color), we plot the RTT and loss rate performance of all RNCs in the network over a two-week time period. The x-axis is the elapsed time in hours since 12AM on Sep.2nd, 2011. The y-axis is the index of all the RNCs, which are grouped based on the state they (primarily) serve. The colorbar on the right is the indicator of the RTT or loss rate performance level. As we observe from the figure, not only that RNCs have large variability across times (the *temporal variability*), *e.g.*, a strong diurnal pattern, but also across different RNCs (the *spatial variability*). Moreover, the RTT and loss rate time series (averaged across all RNCs) are positively correlated in a considerable manner, i.e., correlation coefficient is greater than 0.7. This suggests that time dependent factors, *e.g.*, #bytes (byte count) and #flows (flow count), are possibly one of the main drivers for the performance variations over time.

On the other hand, the spatial variability of RTTs and loss rates cannot be attributed to time dependent factors. Clearly some geographic conditions and location-dependent factors (e..g., varying usage patterns or different mixtures of applications served by different RNCs) may come into play here. In terms of their spatial variability, the RTT and loss rate have very weak correlation (in most hours), with values generally smaller than 0.4 and sometimes negatively correlated (see Fig 4). In other words, suppose RNC *A* has larger RTT than RNC *B* at a particular hour, it does not necessarily imply that *A* is also likely to have larger loss rate than *B* in the same hour. This suggests that there probably exists two separate sets of factors which contribute to the spatial diversity in the RTT and loss rate performance, respectively.

It is also interesting to observe that although their spatial variability is weakly correlated (i.e., not statistically significant enough to draw a causal explanation), there exhibits a distinct diurnal pattern, as clearly shown in Fig 4. The variation is also observed regarding the spatial variability of other usage metrics, such as the #bytes and #flows. These observations indicate that the overall network performance, whether RTT or loss rate, can be affected by a combination of many factors. Their relative importance or contribution to the network performance not only differs across the network, but also varies over time periodically.

IV. First-Order Property Analysis

In this section we start by looking into several major classes of factors that are expected to have likely influence on the RNC-level network performance. For each of them, we present metrics for characterization and estimation, and perform network-wide correlation and other "first-order" analysis to get a baseline understanding of the individual effect (or lack thereof) that each of the major classes of factors has on the overall network performance. Clearly, the complexity and diversity of the network making it almost impossible to exhaust all possible factors and perform similar analysis.

A. Usage Factors

User behaviors can affect the performance in a number of ways due to the varying traffic load and application mix over different time periods (e.g., the time of the day or the day of the week) and across different parts of the network (as represented by the RNCs). Foremost, it directly shapes the diurnal patterns seen in the performance. The usage statistics such as #bytes and #flows (per hour, averaged across all RNCs) have strong correlation with the performance, with the correlation coefficient being around 0.97 for RTT and 0.7 for loss rate. Besides, the traffic load also has a large variation across different RNCs. The largest byte count (and #RRC attempts) at an RNC per hour can reach up to 10 times the smallest among all RNCs. This large variation of traffic load across RNCs is an illustration of the huge diversity (in terms of users and their data access activity) that exists in the large UMTS network.

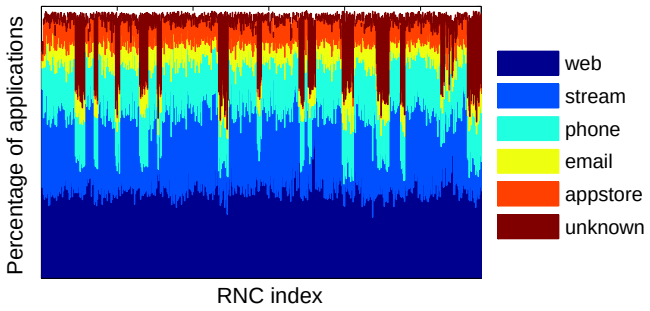

Fig. 5: Fractions of major applications on all RNCs.

In addition to the traffic load, application mix also exhibits large variation and diversity over time and across the RNCs. There are altogether 16 known categories of applications, and one unknown category. Most RNCs see a significant increase in streaming and appstore traffic during the nighttime, whereas during the daytime, web, email and other traffic tend to dominate. The breakdowns of the major applications such as web, streaming, smart phone, email, appstore, as well as the unknown traffic across different RNCs for one of the daytime hours in our dataset are shown in Fig 5 (best viewed in color). The x-axis is the index for all the RNCs, where the RNCs within the same state are grouped closer to each other. The composition of the traffic exhibits a clear geographical distinction. While most RNCs contain a large portion of streaming traffic, several clusters of RNCs contain much larger portions of unknown traffic. A closer look at the dataset reveals that they represent RNCs coming from 14 different states. As will be discussed later, these states are also among those that tend to persistently suffer the worst loss rate performance.

B. Infrastructural Factors

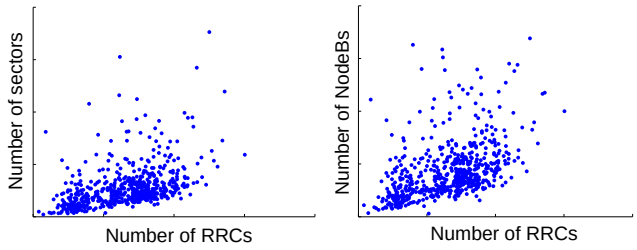

Fig. 6: Correlation of #RRC attempts served and #sectors, #NodeBs deployed at each RNC.

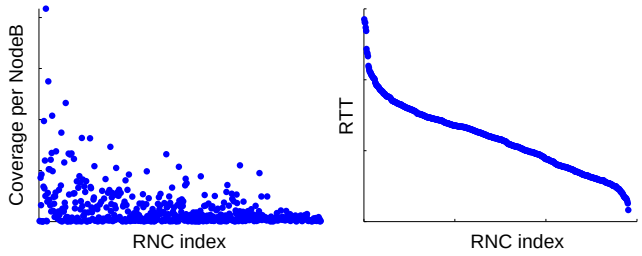

Fig. 7: The effect of NodeB coverage on RTT.

To meet the diversified user demands at different RNCs, the number of NodeBs, sectors (and carriers) associated with each RNC also vary accordingly, as shown in Fig 6. For those regions with a large or an increasing number of UMTS subscribers, multiple NodeBs are likely deployed in the same geographical location. Both the number of devices being placed and the geographical placement strategies used can play a crucial role in influencing the network performance experienced by the users in each region.

For example, a larger number of NodeBs within a geographical area of a fixed size can potentially improve the RTT performance in that it leads to a decrease in the coverage area per NodeB, and therefore smaller "last mile" network latency between UEs and NodeBs. To confirm and verify this intuition, we estimate the coverage area of each NodeB (which is not directly accessible from our data) using a combination of information from the census data and inference using the Voronoi Diagram (based on the GPS locations of NodeBs). The census data provides us with information such as the land area covered by each zip code. Along with the zip code information for each NodeB, we can approximately compute the coverage area per NodeB. However, since multiple NodeBs can be deployed even within the same zip code area, as such this approximation may overestimate the real coverage area of a NodeB. To overcome this problem, we also apply the Voronoi Diagram to demarcate the serving area per NodeB, based on the GPS locations of the NodeBs. The limitation with this method is that not all US land areas (e.g., deserts and other uninhabitable areas) are covered by the NodeBs. Therefore, both methodologies may overestimate the coverage area in some way. For this reason, we infer the coverage area per NodeB by taking the minimum of the results obtained by both methods. In Fig 7, the impact of our estimated NodeB coverage area on the RTT performance is plotted. RNCs are ordered in descending order of their RTT performance for both plots. As expected, RNCs with NodeBs that have smaller coverage areas tend to have smaller average RTT performance.

Similar to the location and placement of NodeBs, the number of RNCs deployed in a geographical area and the coverage areas of RNCs are likely driven by the user/subscriber population and the load they generate. The same also holds true for SGSNs, but to a lesser degree. In general, due to their smaller numbers, RNCs and SGSNs are not as geographically dispersed as NodeBs, but rather placed in a few selected locations for ease of management. There is far less variability in the distance between RNCs and SGSNs than there is between NodeBs and RNCs. The latter highly depends on the geographical coverage of each RNC, and varies from region to region. Nonetheless, both distances can potentially have a significant impact on the network performance, especially the RTT performance.

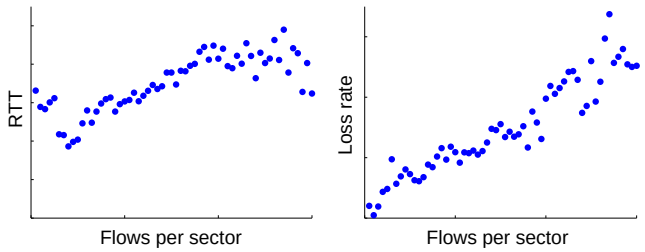

Fig. 8: The impact of the number of flows per sector on RTT and loss rate performance.

In addition, the more sectors (or carriers) are configured on each NodeB, the less load (#flows or #bytes) will be incurred on each sector. The impact of the number of flows per sector on the network performance using one-day data is shown in Fig 8. For a better illustration of the overall trend, the x-axis bucketizes the values of flows per sector, and y-axis averages the RTT or loss rate over all the data points within each bucket. As shown in the figure, there is a clear linear degradation trend on the performance as the flows per sector increases.

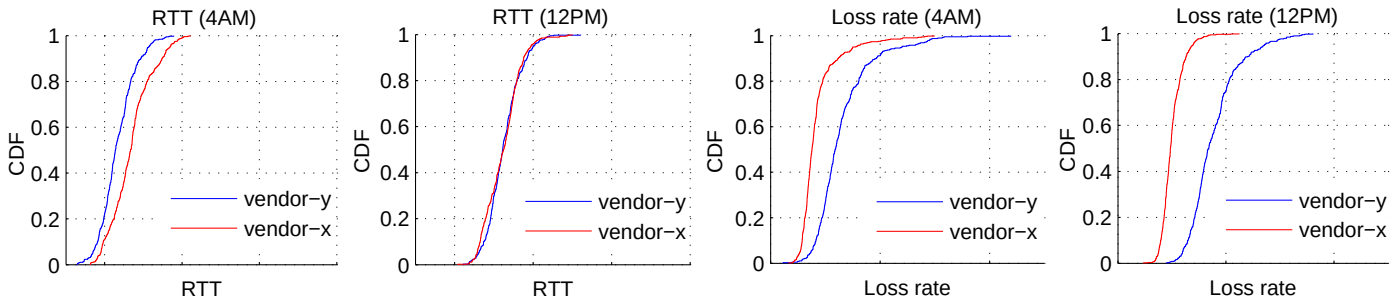

Fig. 9: Comparing the performance for two major types of network elements.

C. Network Element Factors

In addition to the aforementioned factors, the choice of vendor for the network elements such as NodeBs and RNCs may also have an effect on the network performance. This may be due to the fact that different vendors can have fairly different hardware specifications [3], resulting in different data processing speeds and capacities.

In the large cellular network studied in this paper, two major types of network elements are used, vendor-x and vendor-y. The type of NodeB is usually chosen to be the same as the RNC it connects to, most likely due to the hardware level compatibility issue. As an illustration, the comparison of the RTT and loss rate performance distributions of these two types of network elements at two different hours is shown in Fig 9, one for 12PM and the other for 4AM. While the RTT performance distribution is quite similar for both types for most of the 24 hours, all RNCs of vendor-x outperform those from vendor-y in terms of their loss rate. Moreover, such contrasting performance is persistently observed across all 24 hours. More detailed investigation reveals that the maximum capacity for NodeBs of vendor-x is 6 sectors and 12 carriers, whereas NodeBs of vendor-y only have 3 sectors and 6 carriers. To increase the capacity, additional carriers are configured on vendor-y's NodeBs through software configuration. However, as the overall data processing capacity of the network element is still constrained by the hardware capacity, NodeBs of vendor-y tend to incur more packet losses, especially during high loads.

V. Performance Modeling

To understand the complex interaction of various factors and assess their relative importance and contribution to the network performance, we apply RuleFit [1] – a powerful predictive learning method via rule-ensembles in our study. RuleFit provides both predictive and interpretative capabilities that are useful in a number of ways. First, it provides better modeling accuracy, especially for such a large pool of factors by combining linear regression and decision trees. Second, Rulefit ranks the factors based on their importance to the output performance, thereby enabling the relative importance analysis. Third, by analyzing the fitting accuracy of the model across states and hours, or comparing the model obtained using the datasets from all RNCs with those obtained from the RNCs located within certain geographical regions, we can identify the major factors that have persistent network-wide effect as well as uncover those factors that have marked impact at certain locales or at certain times. Furthermore, by examining the level of the overall contributions of all factors as well as the relative importance of each factor and how they differ across locales or over times, we can also infer whether there are potential latent factors in play, affecting the network performance in a way that cannot be quantified using the factors explicitly included in the models. Significant deviations from the usual model predictions may also signal anomalous events, and can be used for diagnosis purposes.

In this section, we start with an overview of Rulefit, and a description of the input features we use in Rulefit. By focusing on a generalized model generated using one-day data, important factors associated with the performance are revealed, Next, the fitting accuracy as well as using it as a mechanism to detect persistent performance anomalies is presented. The predictability of the network performance is also discussed.

A. Rulefit Overview

Rulefit is a supervised rule-based ensemble learning technique. Given a set of input variables $\mathbf{x}$, the ensemble prediction is a linear combination of the predictions of each ensemble members, which takes the form,

$$F(\mathbf{x}) = a_0 + \sum_{i=0}^{m} a_m f_m(\mathbf{x}) \tag{1}$$

where a_i is the linear combination parameters, and $f_m(\mathbf{x})$ is a set of base learners. They are different functions of the input variables possibly derived from different parametric families. In applying Rulefit, we choose base learners to be a set of simple rules, along with a set of linear functions of the input variables. The rules are generated using decision trees. Each rule base learner is in the form of a conjunctive statement,

$$r(\mathbf{x}) = \prod_{j} I(x_j \in s_j) \tag{2}$$

where s_j is a subset of all the possible values of the input variables. $I(.)$ is an indication function, and the rule takes value 1 only if all the arguments hold true. The parameters a_i of the linear regression model are estimated with a lasso penalty. The importance of rule r_k is measured using

$$I_k = |a_k|\sqrt{s_k(1-s_k)} \tag{3}$$

where s_k is the support of the rule on the training data. An input variable is considered important if it defines important rule (or linear) predictors. The relative importance of an input variable x_l is defined as

$$J_l = I_l + \sum_{x_l \in r_k} I_k/m_k \tag{4}$$

where I_l is the importance of the linear predictor, and m_k is the total number of input variables that define the rule r_k.

B. Predictor Variables

The input predictor variables to Rulefit as discussed in Section IV are summarized into the following 35 metrics. Due to space limitation, we refer the readers to [4] for a detailed discussion of the mutual information among these variables.

Usage factors: 1) $nRRC$: number of RRC establishment attempts, 2) $nByte$: number of bytes, 3) $nFlows$: number of flows, 4) $Bpflow$: number of bytes per flow, 5) - 21) fractions of 17 application categories, including misc, web, ftp, instant messaging (IM), P2P, navigation (Nav), VPN, email, Voip, game, Appstore, video optimization (opt), multimedia messaging service (mms), push notification (push), smartphone applications, streaming, and the unknown category.

Infrastructural factors: 22) $B2RDist$: distance between NodeB and RNC, 23) $R2SDist$: distance between RNC and SGSN, 24) $nSector$: number of sectors, 25) $nCarrier$: number of carriers, 26) $RRCpsec$: number of RRCs per sector, 27) $RRCpcar$: number of RRCs per carrier, 28) $nNodeB$: number of NodeBs, 29) $CovpNB$: land coverage per NodeB, 30) $Bpsec$: number of bytes per sector, 31) $Bpcar$: number of bytes per carrier, 32) $Fpsec$: number of flows per sector, 33) $Fpcar$: number of flows per carrier.

Network element factor: 34) $device$: network element device type.

Besides, we also include geographical 35) $state$ as one of the input variables to capture any state specific features. All features are considered numerical, except $device$ and $state$, which is in the form of an unorderable categorical variable.

C. Training & Prediction Sets

For the varying purpose of modeling, prediction, and diagnosing the network performance, we organize our whole datasets into subsets of data in three different ways.

The most intuitive way to organize the data is based on the date they were collected. Each subset of the datasets is a one-day data, containing 24 hourly aggregated measurements for all RNCs. This will facilitate our modeling and prediction of the datasets on a daily basis. Starting from Sep.2nd, 2011, each such dataset is labeled as *TS-D-*{*mm/dd*}. However, this organization of the datasets mixes the measurements performed at different hours as well as from RNCs within different states. To better understand the modeling accuracy at specific hours or at particular RNCs, as well as help diagnose the problem specific to certain hours or RNCs, we further reorganize the datasets based on their hour *TS-H-*{*hour*}, and state information. For simplicity, we perform this finer granularity analysis only on the datasets collected over the first 10 days which observe less anomalies compared to the rest of the days. By comparing the model and important factors from different angles, we are potentially able to diagnose the performance and narrow down the issues to certain days, hours, or states.

D. Generalized Model

To gain insights of the important factors associated with the performance, a generalized Rulefit model is generated and discussed using the first day data TS-D-09/02. The model is generalized in the sense that it is able to model the major performance behaviors for RNCs within different states, across different hours, and predict the performance in the near future.

Important factors: As one of the outputs of our model, input predictor variables are ranked based on their relative importance to the performance. The top 10 most important factors contributing to RTT and loss rate for the generalized model are listed in Tables I and II, respectively. The number below each factor is the *relative* importance level of that factor, where 100 is most important and 0 is least important. Both RTT and loss rate prove to be highly state dependent, likely due to the varying geographical conditions, localized user interest or application mix, as well as other state specific latent factors.

Beyond the state factor, other important factors suggest that RTT is mainly dependent on such architectural factors as the coverage of NodeB, the distance between NodeB and RNC, etc. To better improve the RTT performance, a better placement strategy of the network elements is critical. For instance, replacing NodeBs covering huge areas with larger number of NodeBs covering smaller area and transmitting at a lower power might be more cost-effective. Moreover, instead of using a few selected locations for the placement of RNCs, spreading them out may help reduce the distance between NodeB and RNC.

On the other hand, loss rate is more dependent on such factors as the network element type, the application mix, and flow size (Bpflow). As we discussed earlier, network elements of vendor-x exhibit persistently better loss rate performance than those of vendor-y. Application mix and flow size are highly dependent though. The streaming application is ranked as the most important factor is due to the fact that its flow size is much larger than all other applications.

In addition to uncovering important factors associated with the network performance, the list of rules generated by the model can also be interpreted as a set of possible actionable steps that can be taken by the operators towards improved performance (see [4] for more details).

Fitting accuracy and detecting *persistent* performance anomalies: Due to the fact that the performance, as well as the traffic load and application mix have large variations across different RNCs and hours, our generalized model may not fit equally well for all these scenarios. As depicted in Fig 10, the performance at early morning hours, from 2AM to 4AM, behaves fairly different from the rest of the hours. In fact, it improves during those hours due to the decrease of the overall traffic load and the change of application mix. While most traffic originates from web, email, etc. during daytime hours, streaming traffic that contains much larger flows becomes more prevalent at early morning hours, leading to much better loss rate performance than transmitting huge number of short-lived flows. The improvement of loss rate performance at early morning hours also partially explains the negative correlation of RTT and loss rate as observed in Fig 4. Whereas loss rate is susceptible to drastic change during early morning hours, RTT is relatively stable as it is more dependent on such factor

TABLE I: Top 10 factors and their relative importance to RTT performance.

TS \ Rank	1	2	3	4	5	6	7	8	9	10
TS-D-09/02	state 100	CovpNB 39.72	mms 37.04	B2RDist 15.98	jabber 13.20	Fpsec 5.96	nFlow 4.65	email 4.41	Bpflow 4.24	Bpcar 4.07
TS-D-09/12	state 100	nByte 77.24	CovpNB 53.19	nFlow 33.28	Fpcar 22.64	B2RDist 19.06	Bpcar 17.85	mms 13.81	Fpsec 12.58	RRCpsec 9.34
TS-H-10	state 100	CovpNB 43.30	B2RDist 17.02	R2SDist 10.73	Bpflow 8.29	jabber 6.33	RRCpsec 5.53	nNodeB 4.00	nCarrier 3.76	mms 3.74

TABLE II: Top 10 factors and their relative importance to loss rate performance.

TS \ Rank	1	2	3	4	5	6	7	8	9	10
TS-D-09/02	state 100	stream 61.74	device 46.30	Bpflow 36.43	Appstore 28.57	mms 19.13	Voip 15.75	CovpNB 9.84	R2SDist 9.09	web 7.91
TS-D-09/12	Bpflow 100	state 71.45	unknown 45.72	smartphone 22.23	web 21.08	nByte 18.09	email 15.25	nFlow 13.82	RRCpcar 11.97	Fpcar 11.94
TS-H-10	Bpflow 100	state 93.09	device 43.57	stream 20.68	email 15.79	misc 14.57	jabber 12.20	Appstore 11.77	mms 8.18	unknown 7.01

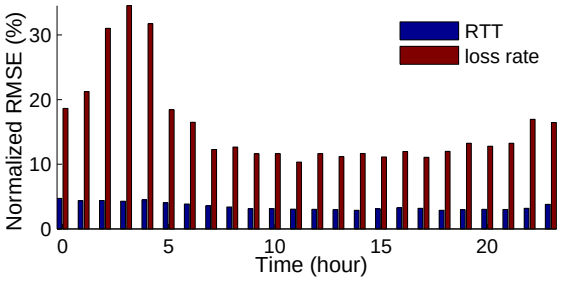

Fig. 10: Hourly fitting accuracy on Sep.2nd.

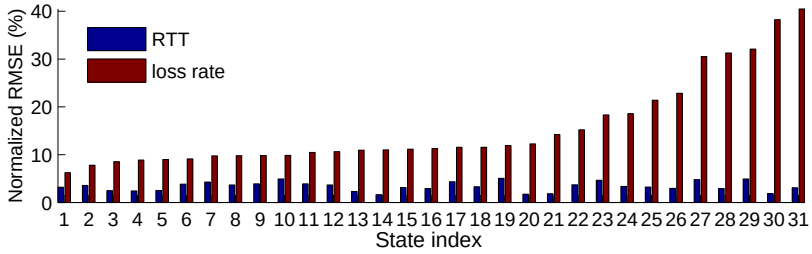

Fig. 11: Fitting accuracy across states on Sep.2nd.

as the coverage of NodeB and the distance between NodeB and RNC.

As both RTT and loss rate are highly state dependent, the fitting error across different states using the same generalized model also observes a large variation, as shown in Fig 11, The x-axis is the state index. In particular, the performance observed in states labeled as 29, 30, 31 is much worse than other states. These performance anomalies (improving or regressing) exist *persistently*, rather than persisting for just one or two days. Therefore, by comparing the fitting accuracy across states and hours (or RNCs if datasets at a finer granularity time scale are available), Rulefit is able to help us detect these anomalies. More importantly, it can also provide the operators with a high-level guide in performing microscopic diagnosis regarding the performance issues specific to certain states, hours, (and RNCs) (see Section VI).

It is also worth noting that loss rate has persistent worse fitting accuracy than RTT, which suggests possible missing latent factors in our model. Loss rate performance is therefore far more complicated and could depend on lots of other factors that were not captured in our dataset.

Predictions: It is also interesting to understand the predictability of the network performance of the 3G cellular network. Some of the factors are normally stable, and therefore can

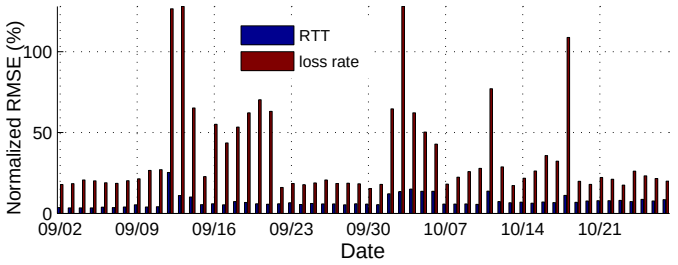

Fig. 12: Prediction of a two-month duration.

be easily predicted from historical data (and prepared for future needs when necessary), whereas other factors are more dynamic leading to poor predictability of the performance.

For this purpose, we conduct a prediction of the performance for a two-month duration using the observations in the first week (i.e., Sep.2nd to Sep.8th). Considering the day of week pattern of the overall performance, we use the same day within the first week to predict the same day performance of future weeks. The daily prediction errors as characterized by normalized root mean square errors(NRMSE) are shown in Fig 12. Note that the prediction error for the first week is the indication of the fitting accuracy of the model generated by Rulefit for the real data.

Our observations are two-fold. First, most of the future RTT performance is quite predictable mainly due to the fact that it is more dependent on infrastructure related factors. Second, similar to the poor fitting accuracy of the generalized model to the loss rate performance, the prediction error is much larger in loss rate than in RTT. The "complexity" of loss rate performance is partially attributed to its dynamic nature of the user related factors such as the usage statistics and application mix as well as the possible dynamic latent factors such as the handoff rate that were not captured in our datasets.

VI. Performance Diagnosis

In this section, we discuss the utility of our macroscopic models in performing microscopic diagnosis of performance anomalies, both *transient* and *persistent*, and revealing the possible existence of latent factors.

A. Transient Performance Regression

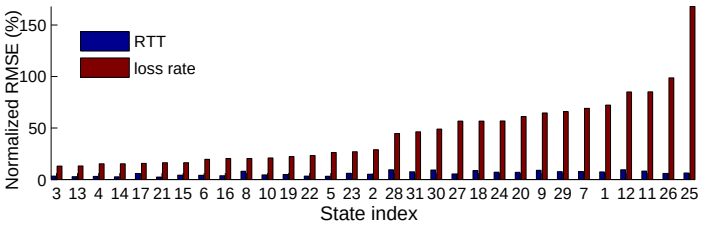

Fig. 13: Prediction error across states on Sep.12th.

As shown in Figs 2 & 3, the RTT and loss rate performance on Sep.12th exhibit extraordinary large values. This type of performance regressions is *transient* in the sense that it appears irregularly and disappears after sometime. To diagnose such anomalies, we utilize the macroscopic model generated using Rulefit to guide us in identifying the attributable factors and entities that are in play. To be more specific, we extract the key factors associated with the anomaly by applying the relative importance analysis, and tag the problem to a subset of states, hours (or RNCs if possible) by comparing the fitting accuracy across different states as well as over time. To gain insights of the underlying behavioral change associated with the factors as suggested by Rulefit, we combine our *macroscopic* model with *microscopic* diagnosis whenever possible.

As an illustration, we diagnose the anomaly on Sep.12th by comparing the important factors generated from Rulefit using TS-D-09/02 and TS-D-09/12. As summarized in Tables I and II, the #bytes starts to play a more critical role in RTT performance, while the #bytes per flow also replaces the state factor and becomes the most important in loss rate, both suggesting the possible issue related to the change of traffic load on that day. By comparing the fitting accuracy across states, as manifested in Fig 13 (using the same state index as in Fig 11), as well as across different hours, we conclude that the anomaly happened at all hours on that day, however only within a subset of states (as indicated by the large fitting errors). Based on these heuristics, we speculate that the problem is likely due to the drastic change of traffic load in this subset of states.

As one step further beyond our high-level macroscopic model, we compare the number of bytes served on Sep.2nd and Sep.12th for each individual state. The results indicate that those states with larger fitting errors also experienced a drastic drop of traffic on that particular day. Therefore, our assumption is indeed borne out by our investigation of the measurement data itself. A more detailed analysis reveals that the drop of traffic load not only happened to a subset of RNCs within those states, but rather, all RNCs within those states. Moreover, these states are by chance all located on the east coast. Based on these induction, we believe that the degradation of the performance was not attributed to holidays or particular events, which normally incur more traffic than usual. Rather, it could be an issue associated with the network carrier itself, such as the failure of SGSN or GGSN that can have a large impact on all the RNCs on east coast.

B. Persistent Performance Anomalies

To diagnose the performance regressions that persist within a subset of problematic states as discussed earlier, we intend to adopt the same methodology in diagnosing the transient anomalies. Similarly, we apply Rulefit to the training sets collected in different states. Due to the elimination of state factor, the RTT fitting accuracy gets improved for all states. However, it does not hold true for loss rate. The explained variance of the loss rate model improves for those states with good performance while degrades to only 50% for those states with poor loss rate performance. This suggests that there may exist possible latent factors that play a critical role but were missing out in our feature list. This phenomenon is also partly revealed by our previous observation as shown in Fig 5, where proportions of the "unknown" category of traffic in those problematic states are much larger than the rest of the states. The missing latent factors might be quite correlated with the unknown type of traffic, and therefore suggesting a finer-granularity investigation of this type of traffic by the operators. Other latent factors might include the state specific geographical and demographic conditions. For instance, in big metro area, the large number of tall buildings can lead to fairly poor performance. In addition, large user density can lead to the placement of many microcells, i.e. base stations of smaller sizes and transmitting at a lower power [5]. As a result, users traverse more quickly through different cells. The poor handoff process therefore may also lead to the interruption of service.

Similarly, to better understand the varying loss rate performance at different hours, we apply the same techniques for training sets collected at different hours. The important factors for one of the daytime hours TS-H-10 are shown in Tables I and II. Unlike RTT, whose important factors at daytime hours turn out to be similar to the generalized model, loss rate is highly dependent on Bpflow during the daytime hours. This is consistent to our observation that most applications prevalent during those hours have smaller flow sizes (as compared to the prevalent streaming traffic with larger flow sizes in the early morning hours), resulting in a larger number of smaller flows, and therefore degraded loss rate performance.

VII. Related Work

There have been quite a few efforts regarding the evaluation or improvement of celluar network performance, such as MobiPerf [6] and existing online tools, e.g., [7], [8]. Compared with our datasets, their measurement studies are all based on

client side factors, without enough insights pertaining to the carrier network itself. A recent study [9] studied the important factors affecting the application performance. However, the performance modeling and diagnosis, as well as the performance differences across the network were not fully understood. Besides, unlike previous studies, which either consider the performance indicator for a specific set of applications such as VoIP and streaming [10]–[12], or consider the performance impact from only a few factors [13]–[15], we evaluate the RTT and loss rate performance that are critical to most of the applications and try to catch up as many factors as we can in modeling the performance. Other performance evaluation works include a study of the impact of variable 3G wireless link latency and bandwidth on the TCP/IP performance [16], the design of a better handoff decision algorithm using the mobile terminal location and area information [17], performance evaluation of cellular networks using more realistic assumptions [18], and a remodeling of handoff interarrival time [19]. However, they are all simulation based work, and the studies were performed under certain restricted assumptions. In terms of methodology, a comparative study of various analysis methods of network performance was performed in [20]. However, none of them was applied to real data and verified to be an effective approach in detecting and diagnosing the performance issue. To our best knowledge, our work is the first attempt to identify a large set of factors in affecting the performance across the network as well as over time, and perform modeling and diagnosis based on them.

VIII. Conclusions

In this paper we have studied the RNC-level performance in a large UMTS cellular network, with the goal to provide a "big picture" understanding of the various major factors that may influence the overall network performance across the network and over time. Our major contributions are the following. *i) Large scale data collection:* Our study utilizes massive data periodically collected from diverse sources over more than six months within one of the largest UMTS cellular network carriers, whose coverage spans the whole United States. *ii) Identification of a rich set of factors:* We identify a rich set of features or factors along different dimensions. While the features gathered in our study are clearly not exhaustive, new factors can be easily accommodated in our models. *iii) Macroscopic modeling and anomaly detection:* The macroscopic models developed in our study provide a better understanding of how various factors may have a differing effect on the network performance across the network or over time. By comparing the fitting accuracy of the models across networks and over time, persistent anomalies specific to certain locales or hours can be detected. *iv) Diagnosing anomalies and unveiling potential latent factors:* We illustrate how the operators can use macroscopic models as a guide to trouble-shoot performance anomalies. Our models can also help unveil potential latent factors, whose effect can be inferred, for example, by comparing the explained variances produced by models feeding on different (sub)sets of data and by performing the relative importance analysis.

Acknowledgments

This work was supported in part by the National Science Foundation under Grant CNS-0905037 and Grant CNS-1017647, by the Defense Threat Reduction Agency under Grant HDTRA1-09-1-0050 and an AT&T VURI (Virtual University Research Initiative) grant.

References

[1] J. Friedman and B. Popescu, "Predictive learning via rule ensembles," *The Annals of Applied Statistics*, pp. 916–954, 2008.

[2] "3gpp specification: 25.331," http://www.3gpp.org/ftp/Specs/html-info/25331.htm.

[3] "Umts/3g nodeb/bts comparison," http://www.umtsworld.com/technology/ran/ran.htm.

[4] "Understanding the complexity of 3g umts network performance," http://www-users.cs.umn.edu/~yingying/techreport.php.

[5] A. Goldsmith, *Wireless communications*. Cambridge Univ Pr, 2005.

[6] J. Huang, C. Chen, Y. Pei, Z. Wang, Z. Qian, F. Qian, B. Tiwana, Q. Xu, Z. Mao, M. Zhang *et al.*, "Mobiperf: Mobile network measurement system," Technical report). University of Michigan and Microsoft Research, Tech. Rep., 2011.

[7] "Speedtest.net," http://speedtest.net/.

[8] "Consumer broadband test," http://www.broadband.gov/qualitytest/.

[9] J. Huang, Q. Xu, B. Tiwana, Z. Mao, M. Zhang, and P. Bahl, "Anatomizing application performance differences on smartphones," in *Proceedings of the 8th international conference on Mobile systems, applications, and services*. ACM, 2010, pp. 165–178.

[10] B. Orstad and E. Reizer, "End-to-end key performance indicators in cellular networks," Ph.D. dissertation, Masters thesis, Information and Communication Technology, Agder University College, Faculty of Engineering and Science, Grimstad, Norway, 2006.

[11] S. Liu, W. Jiang, and J. Li, "Architecture and performance evaluation for p2p application in 3g mobile cellular systems," in *Wireless Communications, Networking and Mobile Computing, 2007. WiCom 2007. International Conference on*. IEEE, 2007, pp. 914–917.

[12] Q. Zhang, W. Zhu, and Y. Zhang, "Network-adaptive scalable video streaming over 3g wireless network," in *Image Processing, 2001. Proceedings. 2001 International Conference on*, vol. 3. IEEE, 2001.

[13] A. Haque, M. Kyum, M. Al Sadi, M. Kar, and M. Hossain, "Performance analysis of umts cellular network using sectorization based on capacity and coverage," *International Journal*, 2011.

[14] J. Fajardo, F. Liberal, and N. Bilbao, "Impact of the video slice size on the visual quality for h. 264 over 3g umts services," in *Broadband Communications, Networks, and Systems, 2009. BROADNETS 2009. Sixth International Conference on*. IEEE, 2009, pp. 1–8.

[15] P. Svoboda, F. Ricciato, W. Keim, and M. Rupp, "Measured web performance in gprs, edge, umts and hsdpa with and without caching," in *World of Wireless, Mobile and Multimedia Networks, 2007. WoWMoM 2007. IEEE International Symposium on a*. IEEE, 2007, pp. 1–6.

[16] M. Chan and R. Ramjee, "Tcp/ip performance over 3g wireless links with rate and delay variation," *Wireless Networks*, 2005.

[17] A. Markopoulos, P. Pissaris, S. Kyriazakos, and E. Sykas, "Cellular network performance analysis: handoff algorithms based on mobile location and area information," *Wireless Personal Communications*, 2004.

[18] Y. Fang, "Performance evaluation of wireless cellular networks under more realistic assumptions," *Wireless Communications and Mobile Computing*, vol. 5, no. 8, pp. 867–885, 2005.

[19] S. Dharmaraja, K. Trivedi, and D. Logothetis, "Performance analysis of cellular networks with generally distributed handoff interarrival times," in *Symposium on Performance Evaluation of Computer and Telecommunication Systems (SPECTS 2002)*, 2002.

[20] P. Lehtimäki, "Data analysis methods for cellular network performance optimization," Ph.D. dissertation, Helsinki University of Technology, 2008.