Bayesian Estimation of Network-wide Mean Failure Probability in 3G Cellular Networks

Angelo Coluccia[1], Fabio Ricciato[1,2], and Peter Romirer-Maierhofer[2]

[1] University of Salento, Lecce, Italy
[2] FTW Forschungszentrum Telekommunikation Wien, Vienna, Austria
{ricciato,coluccia,romirer}@ftw.at

Abstract. Mobile users in cellular networks produce calls, initiate connections and send packets. Such events have a binary outcome — success or failure. The term "failure" is used here in a broad sense: it can take different meanings depending on the type of event, from packet loss or late delivery to call rejection. The Mean Failure Probability (MFP) provides a simple summary indicator of network-wide performance — i.e., a Key Performance Indicator (KPI) — that is an important input for the network operation process. However, the robust estimation of the MFP is not trivial. The most common approach is to take the ratio of the total number of failures to the total number of requests. Such simplistic approach suffers from the presence of heavy-users, and therefore does not work well when the distribution of traffic (i.e., requests) across users is heavy-tailed — a typical case in real networks. This motivates the exploration of more robust methods for MFP estimation. In a previous work [1] we derived a simple but robust sub-optimal estimator, called EPWR, based on the weighted average of individual (per-user) failure probabilities. In this follow-up work we tackle the problem from a different angle and formalize the problem following a Bayesian approach, deriving two variants of non-parametric optimal estimators. We apply these estimators to a real dataset collected from a real 3G network. Our results confirm the goodness of the proposed estimators and show that EPWR, despite its simplicity, yields near-optimum performance.

1 Introduction

The users of a third-generation (3G) mobile network generate various types of activity. Consider the following events: transmission of IP packets, opening of Transport- and Application-layer connections (i.e., envoy of TCP SYN packets and HTTP GET commands), activation of phone call, SMS envoy, data connection and signaling procedures (Attach Request, Location Area Update, Authentication Request, Paging etc.). All such events have in common two characteristics. First, each event is naturally *associated with an individual user* in the mobile network: the caller (for outgoing calls) or the callee (for incoming ones), the sender (for uplink packets) or the receiver (for downlink packets). Second, each event has a *binary outcome: success or failure*. The term "failure" can take on different meanings depending on the type of event: for packets, failure can represent the missed delivery (e.g. due to queue loss or corruption by link-level errors) or late delivery after a delay threshold.

Generally speaking, failures can be caused by the *unavailability* of some resource — e.g. due to exhaustion by too many concurrent requests — or by its *unreachability*. The involved resources can be shared — e.g., the GGSN or a Core Network link — or dedicated to individual users — e.g. a dedicated radio channel, or the receive buffer inside the terminal itself. The *failure probability* experienced by each individual user will be affected by the availability and reachability of both the shared and dedicated resource components.

Even if the status of shared resources is good, user-specific conditions can severely impact the individual failure probability: for example, a user with poor radio link will experience a high rate of packet losses, while a terminal configured with a wrong Access Point Name (APN, ref. [2]) will have all its PDP-context activations rejected. On the other hand, failures affecting multiple users at the same time often indicate a problem and/or overload in the shared section. Therefore, monitoring the incidence of failures across users is of great importance for the operation and troubleshooting of the network infrastructure.

Network equipments typically maintain logs and/or counters of attempts and failures, from which synthetic indicators are derived — often called Key Performance Indicators (KPI). Failures (and successes) can be also measured by a passive monitor either from the observation of an explicit failure indication (e.g. a Negative ACK or reject message) or by the absence of an explicit success indication (e.g. positive ACK). One of the most common KPI is trivially the ratio between the total number of failures and the total number of attempts (or requests). Here we show that such simplistic approach has some fundamental limitations and does not work well in scenarios of practical interest. The problem arises when the individual rate of requests (calls, procedures, packets) varies wildly across users — a case that is typically encountered in real networks when the distribution of user activity is often heavy-tailed. In this work we formulate the problem in terms of Bayesian estimation and provide (two variants of) an optimal estimator. Furthermore, we compare its performance to a simpler estimator developed in a previous work [1], discussing the trade-offs between optimality and simplicity.

2 Problem formulation

2.1 System model

To preserve generality, we will adopt the term "REQUEST" to refer to a general resource access attempt — to avoid specific terms like packet, call, connection. Unsuccessful requests are referred to as "FAILURES". Each request is associated with one mobile "USER".

For a generic measurement timebin (e.g. 1 minute), let I denote the total number of *active users* for which at least one request was observed. In operational networks I is often quite large, from thousands to millions depending on the timebin duration. For every user i ($i = 1 \dots I$) we introduce the following variables:

- n_i is the total number of requests associated to i ($n_i \geq 1$);
- m_i is the number of failures ($0 \leq m_i \leq n_i$);
- $r_i \stackrel{\text{def}}{=} \frac{m_i}{n_i}$ denotes the empirical failure ratio for i;

– a_i the (unknown) failure probability for i ($a_i \in [0,1]$).

We denote by $N \stackrel{\text{def}}{=} \sum_i n_i$ and $M \stackrel{\text{def}}{=} \sum_i m_i$ respectively the total number of requests and failures across all terminals. To simplify the notation we will occasionally use the vectors $\boldsymbol{n} \stackrel{\text{def}}{=} [n_1 n_2 \cdots n_I]^T$, $\boldsymbol{m} \stackrel{\text{def}}{=} [m_1 m_2 \cdots m_I]^T$ and $\boldsymbol{a} \stackrel{\text{def}}{=} [a_1 a_2 \cdots a_I]^T$. For the sake of mathematical tractability we assume independence between failures: each request for user i fails with probability a_i independently from any other request of the same or other users. Therefore m_i is the sum of n_i Bernoulli trials with probability a_i, and all m_i's are independent Binomial random variables:

$$m_i \sim \mathcal{B}(n_i, a_i) \Rightarrow \begin{cases} \mathrm{E}[m_i] = n_i a_i \\ \mathrm{VAR}[m_i] = n_i a_i (1 - a_i). \end{cases} \tag{1}$$

Throughout the paper we assume that $\boldsymbol{m}$ and $\boldsymbol{n}$ have been measured in some way and serve as input for the problem at hand.

2.2 Goal definition

A central component of the model is that the (unknown) failure probabilities a_i's are regarded as i.i.d. random variables generated from a common underlying distribution $p(a)$ with mean value $\bar{a} \stackrel{\text{def}}{=} \mathrm{E}[a]$. Therefore it is natural to take $\bar{a}$ as the summary indicator representative of "network-wide failure probability". In other words, we are interested in obtaining an estimator of the Mean Failure Probability $\bar{a}$ from the measured vectors $\boldsymbol{n}$ and $\boldsymbol{m}$.

The goodness of such an estimator can be evaluated against the following criteria:

– **Optimality** We consider only unbiased estimators and take minimum variance as the optimality criterion. In fact, when the estimate of $\bar{a}$ is used for performance monitoring, lower variance (i.e., smaller statistical fluctuations) allows better discrimination of change-points and/or trends.
– **Generality** In passive monitoring the vector $\boldsymbol{n}$ is given and can not be controlled. Typically, the traffic volume is distributed very unevenly across users, often with long-tails. Moreover, the traffic distribution change across time, following daily or weekly cycles and long-term trends in user activity (see e.g. [4, §VI-A]). Therefore we seek a robust estimator that does not rely on specific assumptions about (the distribution of) $\boldsymbol{n}$ nor requires manual re-tuning when the distribution changes.
– **Simplicity** The ideal estimator should be easy to implement, fast to compute and conceptually simple — as a matter of fact, methods which can be understood straightforwardly by practitioners are more likely to be adopted in practice.

2.3 Resolution Approach

In a previous contribution [1] we focused on a particular class of weighted estimators and casted the problem in terms of constrained optimization. We derived a very simple sub-optimal estimator, hereafter called Empirical Piecewise-linear Weighted Ratio (EPWR), which showed excellent performance in simulations. The EPWR involves a

cut-off parameter θ to be set heuristically — we showed that it is not too sensitive to the exact value of θ as far as extreme settings (very small or very large) are avoided.

In this work we tackle the problem from a more theoretically-grounded perspective: following a formal Bayesian approach, we identify two estimators that are well suited for our purposes. Moreover, we compare the performance of the Bayesian estimators against the EPWR estimator derived in [1] plus two other simplistic estimators. The comparison is carried out on a sample dataset from a real operational network, based on data obtained with the METAWIN system [7].

3 Simple estimators

Hereafter we present two common estimators and highlight their limitations. In §3.3 we recall the EPWR estimator which was derived earlier in [1].

3.1 Empirical Global Ratio (EGR)

The Key Performance Indicator (KPI) most widely adopted by practitioners is simply the ratio between the total number of failures and the total number of requests across all users, formally:

$$S_{\text{EGR}} \stackrel{\text{def}}{=} \frac{\sum_{i=1}^{I} m_i}{\sum_{i=1}^{I} n_i} = \frac{M}{N}. \tag{2}$$

We refer to such quantity as the *Empirical Global Ratio* (EGR). It can be seen that S_{EGR} is indeed an unbiased estimator for the mean failure probability:

$$\mathrm{E}[S_{\text{EGR}}] = \frac{\sum_{i=1}^{I} \mathrm{E}[\mathrm{E}[m_i|a_i]]}{\sum_{i=1}^{I} n_i} = \frac{\sum_{i=1}^{I} n_i \mathrm{E}[a_i]}{\sum_{i=1}^{I} n_i} = \bar{a}$$

It is worth remarking that S_{EGR} is the simplest estimator to implement in practice. In fact, it does not require knowledge of the full vectors $\boldsymbol{n}, \boldsymbol{m}$ (each composed of I elements) but only of two global counters N and M. Therefore, it does not require the measurement platform to associate requests and failures to individual users.

Intuitively, the problem with S_{EGR} is the presence of few users with very high traffic (large n_i) that occasionally inflate the value of eq. (2), thus increasing the estimator variance. One possible approach to "correct" S_{EGR} is to pick users with large n_i as "outliers" and filter them out before computing the ratio in eq. (2). Such strategy has two drawbacks: for one, it requires an heuristically-tuned method to classify outliers. Second, filtering out "fat" users with large n_i means discarding their sample estimates for $\bar{a}$ — actually the most reliable ones — and for long-tailed n_i's the loss of information might not be negligible. In other words, the variance increase due to a lower number of reliable measurement samples will partially offset the reduction due to filtering.

3.2 Empirical Mean Ratio (EMR)

Note that the empirical loss ratio r_i for each user i is an unbiased estimator of the mean failure probability $\bar{a}$, formally $\mathrm{E}[r_i] = \mathrm{E}\big[\mathrm{E}[r_i|a_i]\big] = \mathrm{E}[a_i] = \bar{a}$. Therefore an intuitive

summary indicator can be obtained as the arithmetic mean across all users:

$$S_{\text{EMR}} \stackrel{\text{def}}{=} \frac{1}{I}\sum_{i=1}^{I} r_i = \frac{1}{I}\sum_{i=1}^{I} \frac{m_i}{n_i}. \tag{3}$$

We refer to such indicator as the *Empirical Mean Ratio* (EMR). From a system-level perspective, the implementation of S_{EMR} is more costly than S_{EGR}: in fact, it requires the measurement platform to count requests and failures separately for each user. The additional resource consumption (memory, processing) should be compensated by better accuracy of the estimation. However, S_{EMR} is a sub-optimal estimator for $\bar{a}$. The problem lies in the fact that the variance of r_i (conditioned to a_i) is inversely proportional to the number of packets n_i, i.e. $\text{VAR}[r_i|a_i] = a_i(1-a_i)/n_i$: intuitively, the larger the sample size, the better the accuracy of the estimate. Therefore, large variability of n_i's maps to large variability of the uncertainty (variance) of the individual estimates — a case of *heteroscedasticity*. In the simple arithmetic mean as in eq. (3), more accurate estimates (for large n_i) weight the same as poor ones (for small n_i), and if the number of low-traffic users is high they will drive up the overall variance.

Similarly to S_{EGR}, one possible "correction" for S_{EMR} is to filter out (discard) the observations associated to low-traffic users and compute the ratio in eq. (3) only on the samples above a minimum sample size $n_i \geq \gamma$. Again, the drawback of such strategy is that a certain amount of information available in the data is simply discarded, which is a grossly suboptimal approach to the problem.

3.3 Empirical Piecewise-linear Weighted Ratio (EPWR)

We have seen that both S_{EMR} and S_{EGR} suffer respectively from small and large users — the user size refers to its traffic volume n_i — and that a simplistic workaround would be to just discard the measurements associated with the smallest or biggest users for S_{EMR} and S_{EGR} respectively. This approach involves a certain loss of information. Moreover, the decimation of samples works *against* the goal of reducing the uncertainty of the estimate. Therefore we are set to find a more clever strategy: instead of selectively *discarding* measurement samples based on their size, one can simply *weight* them differently. Following this idea, in a previous work [1] we have derived a simple sub-optimal estimator that takes the form of a weighted average of the r_i's:

$$S_{\text{EPWR}} \stackrel{\text{def}}{=} \sum_{i=1}^{I} w_i r_i \tag{4}$$

with piecewise-linear weigths given by:

$$w_i = \frac{x_i}{\sum_{i=1}^{I} x_i}, \quad x_i = \min(n_i, \theta) \;\; \forall i \tag{5}$$

Such definition involves a single parameter θ that represents a sort of "cut-off" point dividing the users into two regions: those with $n_i < \theta$ are weighted proportionally to their size n_i, while those with $n_i > \theta$ are weigthed equally — proportionally to θ. The analysis reveals that the optimal value of the cut-off parameter is given by $\dot{\theta} = \frac{\bar{a}-\bar{a}-\sigma_a^2}{\sigma_a^2}$, i.e.

it depends on the first two moments of the distribution of $\boldsymbol{a}$. Notably the optimal setting does not depend on $\boldsymbol{n}$, which is an advantage in applications where $\boldsymbol{n}$ varies in time. Unfortunately, in practice the optimal value of $\dot{\theta}$ cannot be identified since $\bar{a}$ and σ_a^2 are unknown. However it can be shown that the performances of S_{EPWR} depend only weakly on the exact value of θ, provided that it falls in a "reasonable" intermediate range away from extreme values (very small or very large), indicating that an heuristically fixed value for θ would be sufficient to achieve near-optimal performance in most practical cases. This claim, supported by simulation results in [1], will be further confirmed by the numerical results presented below in §5.

4 Bayesian estimators

Since $\boldsymbol{a}$ is regarded as a random vector, we can apply Bayesian techniques, which provide a theoretically well grounded approach to the estimation problem at the cost of somewhat higher complexity of the resolution procedure. In this section we discuss different possible approaches under the common framework of Bayesian inference.

The structure of our problem is that of a Bayesian hierarchical model [5]: the data $\boldsymbol{m}$ are described by a probability distribution whose parameters $\boldsymbol{a}$ are random variables themselves. The Bayesian approach requires to provide the *a priori* distribution $p(\boldsymbol{a})$, which is unknown in our case. If $p(\boldsymbol{a})$ were known, the optimal estimator for $\boldsymbol{a}$ would be obtained by following the classical Bayes' procedure, i.e. minimizing a Bayes risk. Two of the most common choices for the Bayes risk are the MAP (*Maximum A Posteriori*) criterion and the MMSE (*Minimum Mean Square Error*) criterion. In both cases the key element is the posterior distribution $p(\boldsymbol{a}|\boldsymbol{m})$, expressed by the Bayes' Theorem as:

$$p(\boldsymbol{a}|\boldsymbol{m}) = \frac{p(\boldsymbol{m},\boldsymbol{a})}{p(\boldsymbol{m})} = \frac{p(\boldsymbol{m},\boldsymbol{a})}{\int p(\boldsymbol{m},\boldsymbol{a})\,\mathrm{d}\boldsymbol{a}}$$

where $p(\boldsymbol{m},\boldsymbol{a}) = p(\boldsymbol{m}|\boldsymbol{a})p(\boldsymbol{a})$.

When $p(\boldsymbol{a})$ is unknown, as in our case, it is possible to choose a parametric distribution family with unknown parameters for the prior, and to resort to a procedure called *Empirical Parametric Bayes* [5]. The basic idea is quite simple: since the parameters of the prior distribution — called *hyperparameters* — are unknown, their estimates are used instead. It is then possible to derive a Bayesian estimator (MAP or MMSE) in the usual way. The estimation of the hyperparameters is preferably obtained via Maximum Likelihood (ML), although other techniques are sometimes used — e.g. Method of Moments [5].

In some applications it is not clear which family of distributions should be adopted. In such cases, a well-established practice is to use the so-called *conjugate prior*, i.e. the prior distribution $p(\boldsymbol{a})$ that results in a posterior distribution $p(\boldsymbol{a}|\boldsymbol{m})$ belonging to the same family. In our case, the conjugate prior corresponding to the Binomial distribution is the Beta distribution, with support in $[0,1]$, defined as:

$$p(a_i) = \frac{1}{\mathrm{B}(\alpha,\beta)} a_i^{\alpha-1}(1-a_i)^{\beta-1} \qquad i = 1,\ldots,I \tag{6}$$

where

$$\mathrm{B}(x, y) = \int_0^1 t^{x-1}(1-t)^{y-1}\,\mathrm{d}t \tag{7}$$

is the Beta function [10]. The Beta distribution is quite general and flexible: it has two positive shape parameters α, β which allow for a great variety of shapes, ranging from uniform to U-shape, convex or concave, symmetric or skewed. Moreover, consider that Bayesian estimation is usually found to be robust against deviations from the ideal choice of the prior distribution family. The mean of a Beta distribution is given by

$$X \sim \mathrm{Beta}(\alpha, \beta) \Rightarrow \mathrm{E}\left[X\right] = \frac{\alpha}{\alpha + \beta} \tag{8}$$

The ML estimates of the hyperparameters are given by:

$$\hat{\alpha}, \hat{\beta} \stackrel{\text{def}}{=} \arg\max_{\alpha,\beta} p(\boldsymbol{m}|\alpha, \beta) \tag{9}$$

where the marginal distribution $p(\boldsymbol{m}|\alpha, \beta)$ is obtained by marginalization:

$$p(\boldsymbol{m}|\alpha, \beta) = \int_0^1 \cdots \int_0^1 p(\boldsymbol{m}, \boldsymbol{a}|\alpha, \beta)\,\mathrm{d}a_1 \cdots \mathrm{d}a_I. \tag{10}$$

Due to independence, we can factorize the joint density function as follows:

$$p(\boldsymbol{m}, \boldsymbol{a}|\alpha, \beta) = \prod_{i=1}^{I} p(m_i, a_i|\alpha, \beta) = \prod_{i=1}^{I} p(m_i|a_i, \alpha, \beta)p(a_i)$$
$$= \prod_{i=1}^{I} \binom{n_i}{m_i} a_i^{m_i}(1-a_i)^{n_i-m_i} \cdot \frac{1}{\mathrm{B}(\alpha, \beta)} a_i^{\alpha-1}(1-a_i)^{\beta-1}$$

and recalling eq. (7) we can rewrite eq. (10) as follows:

$$p(\boldsymbol{m}|\alpha, \beta) = \int_0^1 \cdots \int_0^1 \prod_{i=1}^{I} p(m_i, a_i|\alpha, \beta)\,\mathrm{d}a_i = \prod_{i=1}^{I} \binom{n_i}{m_i} \frac{\mathrm{B}(\alpha + m_i, \beta + n_i - m_i)}{\mathrm{B}(\alpha, \beta)} \tag{11}$$

i.e., the marginal distribution of $\boldsymbol{m}$ is the product of I Beta-Binomial distributions $\mathrm{BetaBin}(n_i, \alpha, \beta)$. This expression can be used in eq. (9) to obtain the ML estimates of the hyperparameters. The solution cannot be expressed in closed-form and must be obtained numerically. Notably, the function (11) is unimodal (see [13] for a proof) and its negative logarithm is convex — a consequence of the log-concavity of the likelihood function for Beta [12] [8] — therefore standard numerical methods can be applied (e.g. the simplex method) which are likely to locate the optimum quickly and accurately.

Given the estimates of the hyperparameters, the posterior distribution

$$p(a_i|m_i, \hat{\alpha}, \hat{\beta}) = \frac{p(m_i, a_i|\hat{\alpha}, \hat{\beta})}{p(m_i|\hat{\alpha}, \hat{\beta})}$$

is obtained in a similar way, and as expected is a Beta distribution (conjugate prior):

$$a_i|m_i, \hat{\alpha}, \hat{\beta} \sim \mathrm{Beta}(\hat{\alpha} + m_i, \hat{\beta} + n_i - m_i) \tag{12}$$

The (empirical) Bayes estimator for the generic a_i is then derived in the classical way by minimizing a Bayes risk. Two of the most common criteria are MAP and MMSE.

The MAP estimator maximizes the *a posteriori* probability:

$$\hat{a}_i^{\text{MAP}} \stackrel{\text{def}}{=} \arg\max_{a_i} p(a_i|m_i, \hat{\alpha}, \hat{\beta}) = \frac{\hat{\alpha} + m_i - 1}{\hat{\alpha} + \hat{\beta} + n_i - 2} \tag{13}$$

for $i = 1 \dots I$, as can be easily verified by taking the derivatives of $\log p(a_i|m_i, \hat{\alpha}, \hat{\beta})$ (ref. eq. (6) and (12)). From the denominator of eq. (13) we observe that the MAP estimator may lead to inconsistent values for small n_i. This is a problem in network applications, where typically a considerable fraction of users generate only very few requests. Therefore the MAP estimator is not well suited for our purposes.

The MMSE estimator minimizes the mean squared error, and coincides with the conditional mean [5]. Recalling eq. (8) we can write:

$$\hat{a}_i^{\text{MMSE}} \stackrel{\text{def}}{=} \mathrm{E}\Big[a_i \Big| m_i, \hat{\alpha}, \hat{\beta}\Big] = \frac{\hat{\alpha} + m_i}{\hat{\alpha} + \hat{\beta} + n_i}$$

for $i = 1 \dots I$, and by taking the arithmetic mean we obtain the following estimator:

$$Q_{\text{MMSE}} \stackrel{\text{def}}{=} \frac{1}{I} \sum_{i=1}^{I} \hat{a}_i^{\text{MMSE}} = \frac{1}{I} \sum_{i=1}^{I} \frac{\hat{\alpha} + m_i}{\hat{\alpha} + \hat{\beta} + n_i}. \tag{14}$$

Otherwise, an alternative approach is to estimate directly $\bar{a}$ from eq. (8) with the ML estimates of the hyperparameters:

$$Q_{\text{HYP}} \stackrel{\text{def}}{=} \frac{\hat{\alpha}}{\hat{\alpha} + \hat{\beta}} \tag{15}$$

where the subscript "HYP" indicates that the estimator is obtained solely from the hyperparameters' estimates, without the posterior information $\boldsymbol{m}$.

Note that in both cases the $\hat{\alpha}, \hat{\beta}$ are estimated from the data vector $\boldsymbol{m}, \boldsymbol{n}$: the difference between the two estimators lies in the fact that Q_{MMSE} is suboptimal — because it heuristically adopts an arithmetic mean — but uses the posterior information $\boldsymbol{m}$ for both parameters and hyperparameters, while Q_{HYP} is optimal in the ML sense but uses the posterior information only for estimating the hyperparameters and not $\bar{a}$. Despite such difference, we found that Q_{MMSE} and Q_{HYP} always lead to extremely similar values when applied to our real datasets, as discussed later in §5.

Finally we remark that both estimators require a numerical procedure to be computed. However the shape of the objective function (unimodal, convex) allows the use of fast and accurate numerical methods — for the sake of space we do not provide here further details of the implemented numerical procedure. Nonetheless, the computational gap with S_{EPWR} remains large: in a MATLAB simulation with $I = 10^5$ users, the computation time of Q_{MMSE} and Q_{HYP} is about 10 seconds on a standard computer (Core2 Duo 2Ghz), against a few milliseconds of S_{EPWR}.

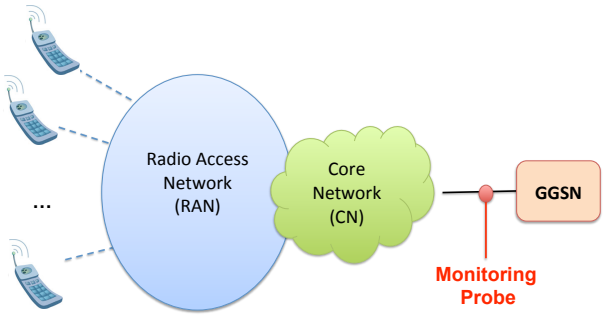

Fig. 1. Reference measurement scenario.

5 Numerical Results from Real Datasets

In the previous sections we have presented three sub-optimal estimators, very simple to implement and to compute — S_{EGR}, S_{EMR} and S_{EPWR} — and derived two Bayesian estimators — Q_{HYP} and Q_{MMSE} — which are more complex but optimal given the problem model at hand. Hereafter we compare their performance on a real dataset.

Our dataset is based on measurements collected from an operational 3G cellular network by the METAWIN system [7]. The measurement setting is sketched in Fig. 1: a passive monitor located on the Gn interface near the GGSN observes the TCP traffic in both directions (for more details on 3GPP network architecture refer e.g. to [2]). We aim at revealing congestion and/or other performance glitches in the network section between the monitoring point and the mobile terminals from the observation of TCP handshaking packets between mobile clients and Internet-side servers. To this purpose we collect two datasets: DATA:INV and DATA:RTT.

In DATA:INV we count for each mobile station i all SYNACK packets flowing in downlink (variable n_i) as well as the number of them which *failed to be unambiguously associated* to a corresponding uplink ACK (variable m_i). Such definition includes those cases where either the SYNACK or the corresponding ACK were lost in the network section from the monitoring point to the mobile terminal, but also other cases not necessarily related to loss events. For example, when two or more identical SYNACKs are observed but only one ACK (between the same end-points) the SYNACK-to-ACK association remains ambiguous, i.e. it is not possible to decide which one of the SYNACK triggered the ACK[3]. In other words, the SYNACK packet in downlink denotes a "request" event, and the presence [resp. lack] of an *unambigously corresponding* ACK packet in uplink denotes the "success" [resp. "failure"] event.

A second dataset DATA:RTT was obtained from the (semi-)RTT measurements: for each correctly (and unambigously) acknowledged SYNACKs, we measure the client-side RTT, i.e. the elapsed time between the timestamps of the SYNACK and the corresponding ACK. For this dataset, we mark a "failure" event when the RTT exceeds a

[3] In [3] we showed that such cases are not infrequent in GPRS/UMTS networks due to the presence of "early retransmitter" servers, with initial retransmission timeout set to sub-second values in the same order of the Round Trip Time (RTT) in these networks.

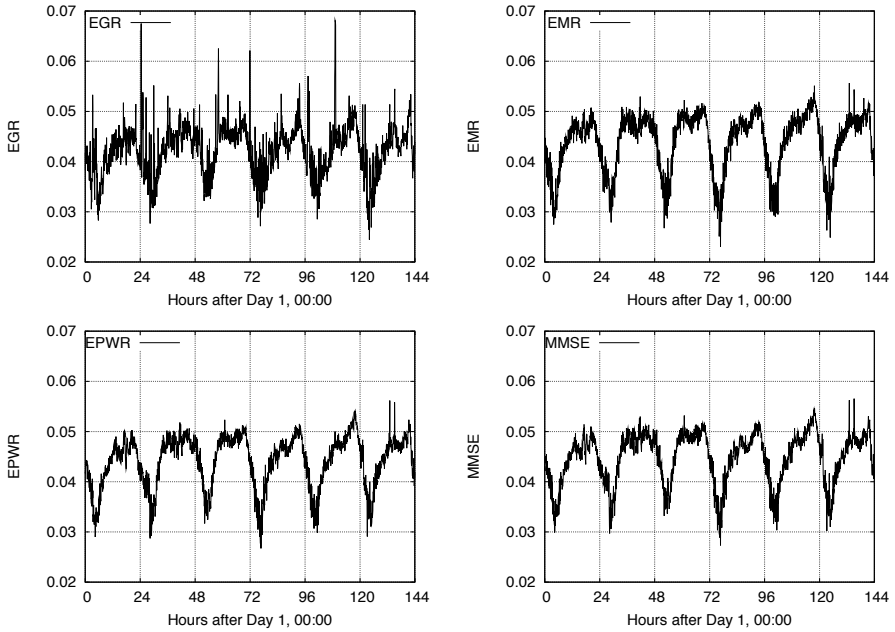

Fig. 2. Estimated mean failure probability for DATA:INV dataset (missing or ambiguous SYNACK/ACK associations).

fixed threshold T — we set $T = 0.5$ sec and $T = 1.5$ sec respectively for UMTS/HSPA and GPRS/EDGE. For more details about the measurement setting see [3].

The rationale for extracting such measurements is that congestion in the downlink path towards the terminals (e.g. at some SGSN or RNC) would expectedly result in an increase of downlink packet loss and/or delay, and therefore should be reflected in a increase of the "failure" rate in DATA:RTT and/or DATA:INV. Note that serious connectivity problems in the Radio Access Network might impede even the transmission of uplink SYN packets, which would translate into a reduction of SYNACK — i.e. we would observe *missing* rather than *unacknowledged* SYNACKs. While in principle such kind of events could be revealed by monitoring the absolute number of SYNs and corresponding SYNACKs, this aspect is left outside the scope of this work.

The measurements are binned in 5 minutes intervals. For this work we consider a measurement period of one week collected in August 2010. The analysis was conducted separately for UMTS/HSPA and GPRS/EDGE users, but for the sake of space we report only results for UMTS/HSPA. The time-series computed with different estimators are reported in Fig. 2 and Fig. 3 respectively for DATA:INV and DATA:RTT datasets. We found that Q_{HYP} and Q_{MMSE} always lead to almost identical results, with only negligible differences in a few timebins — for this reason and for the sake of space we report only the time-series of Q_{MMSE} and skip the graphs for Q_{HYP}.

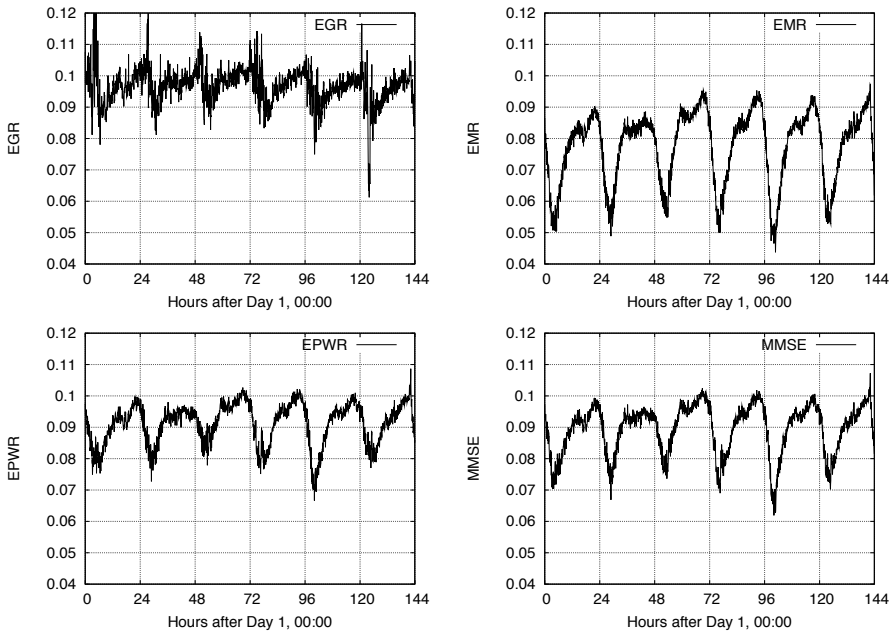

Fig. 3. Estimated mean failure probability for DATA:RTT dataset (unambiguous SYNACK/ACK pairs with semi-RTT exceeding 500 ms).

From Fig. 2 and Fig. 3 we first observe that S_{EGR} exhibits larger fluctuations (higher variance) and occasionally large spikes due to the sporadic presence of heavy users (high n_i) with many failures[4] (high m_i). Second, both S_{EMR} and S_{EPWR} perform quite close to the optimal reference Q_{MMSE}. To dig further, we resorted to the inspection of the scatterplots $\langle S_{\text{EMR}}, Q_{\text{MMSE}} \rangle$ and $\langle S_{\text{EPWR}}, Q_{\text{MMSE}} \rangle$ (not shown here) which revealed that S_{EPWR} correlates slightly better than S_{EMR} to the reference Q_{MMSE} values: in DATA:INV the Pearson the correlation coefficient is 0.997 and 0.954 respectively for $\langle S_{\text{EMR}}, Q_{\text{MMSE}} \rangle$ and $\langle S_{\text{EPWR}}, Q_{\text{MMSE}} \rangle$, while in DATA:RTT S_{EMR} exhibits a larger bias.

6 Conclusions

So called Key Performance Indicators (KPI) play an important role in the operation of real mobile networks, as they provide a synthetic view of the network-wide status and quality. A large class of network events can be modeled as binary REQUESTS associated to USERS (e.g. origin or destination entity) and having one of two possible outcomes, i.e. SUCCESS or FAILURE. For these, a very popular KPI is the ratio between the total number of failures to the total number of requests, regardless of per-user associations. We have shown that such simplistic KPI — referred to as EGR in this

[4] The "spikes" in DATA:INV are due to mobile terminals receiving very high rate of SYNACKs to which they do not respond: they are likely involved in TCP scanning or SYN flooding.

work — suffers from the presence heavy-users. The problem is of practical relevance in real networks, where the distribution of requests across users is often heavy-tailed. In some cases, the variability of EGR makes it useless for any practical exploitation.

To overcome the limitations of EGR, network operators should adopt more robust KPIs based on separate counts of success and failures per individual users. The problem is then how to make the best possible use of such data.

We have introduced a system model that motivates the adoption of the Mean Failure Probability (MFP) as a natural KPI. Since MFP is unknown, it must be estimated from the observed data. We have shown that EGR can be considered an unbiased estimator for MFP, but not the optimal one. In a previous work we had derived a more robust estimator, namely the EPWR, very simple to implement, that involves a free parameter to be tuned heuristically. In this paper we have formalized the problem in terms of Bayesian estimation, deriving two Bayesian estimators which are provably optimal (in two different senses) given the system model. The analysis of two real datasets from an operational 3G mobile network has confirmed that all the proposed estimators are considerably more stable than EGR, and that EPWR performs very closely to the optimal reference provided by the Bayesian solution.

Our estimators can be adopted by network operators and/or equipment vendors as robust KPI. Owing to the generality of the system model, and to the abstract definition of the notions of "request", "failure" and "user" therein, the concepts and estimators proposed in this work can be applied to a wide range of different measurements, in communication networks and other application domains.

References

1. A. COLUCCIA, F. RICCIATO, P. ROMIRER: On Robust Estimation of Network-wide Packet Loss in 3G Cellular Networks, *5th IEEE Broadband Wireless Access Workshop (BWA'09), Honolulu, Nov. 2009.*
2. H. KAARANEN *et al.*: UMTS Networks — Architecture, Mobility and Services, 2nd ed., *Wiley*, 2005.
3. PETER ROMIRER *et al.* Network-Wide Measurements of TCP RTT in 3G, *Proc. of TMA'09 workshop*, Aachen, LNCS vol. 5537, May 2009
4. A. D'ALCONZO, A. COLUCCIA, F. RICCIATO, P. ROMIRER: A Distribution-Based Approach to Anomaly Detection for 3G Mobile Network, *IEEE GLOBECOM '09*
5. E.L. LEHMANN, G. CASELLA: Theory of Point Estimation, *Springer*, 1998
6. K.M. WOLTER: Introduction to Variance Estimation, *Springer Series in Statistics*, 2007.
7. METAWIN and DARWIN projects `http://userver.ftw.at/~ricciato/darwin/`
8. T. P. MINKA: Estimating a Dirichlet distribution, *Microsoft Technical Report*, 2003
9. H. ROBBINS: An Empirical Bayes Approach to Statistics, *Proc. Third Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1, Univ. of California Press, 1956, 157-163.
10. M. ABRAMOWITZ, I. A. STEGUN: Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables, *Dover Publications*, New York, 1972
11. A. YEREDOR: The Joint MAP-ML Criterion and its Relation to ML and to Extended Least-Squares, *IEEE Trans. on Signal Processing*, vol. 48, no. 12, Dec. 2000
12. E. I. GEORGE, U. E. MAKOV, A. F. M. SMITH: Conjugate Likelihood Distributions, *Scandinavian Journal of Statistics*, vol. 20, no. 2, 1993, pp. 147–156
13. B. KEVIN. J. REEDS: Compound Multinomial Likelihood functions are unimodal: proof of a conjecture of I. J. Good, *The Annals of Statistics*, vol. 5, no. 1, 1977, pp. 79–87.