Signalling Storms in 3G Mobile Networks

Omer H. Abdelrahman and Erol Gelenbe

Department of Electrical & Electronic Engineering
Imperial College, London SW7 2BT, UK
Email: {o.abd06, e.gelenbe}@imperial.ac.uk

***Abstract*—We review the characteristics of signalling storms that have been caused by certain common apps and recently observed in cellular networks, leading to system outages. We then develop a mathematical model of a mobile user's signalling behaviour which focuses on the potential of causing such storms, and represent it by a large Markov chain. The analysis of this model allows us to determine the key parameters of mobile user device behaviour that can lead to signalling storms. We then identify the parameter values that will lead to worst case load for the network itself in the presence of such storms. This leads to explicit results regarding the manner in which individual mobile behaviour can cause overload conditions on the network and its signalling servers, and provides insight into how this may be avoided.**

I. Introduction

Mobile networks are vulnerable to *signalling attacks* which overload the control plane through traffic patterns that target the signalling procedures involved [1]–[4], by compromising a large number of mobile devices as in network Denial of Service (DoS) attacks [5], [6] or from outside the mobile networks (e.g. the Internet). Similarly software and apps on mobile devices [7], [8] can cause such disturbances through frequent traffic bursts. Such attackers can actively probe the network to infer the network's radio resource allocation policies [9], [10] and identify IP addresses in specific locations [11]. Indeed, a review of 180 cellular carriers around the world revealed that 51% of them allow mobile devices to be probed from the Internet by either assigning public IP addresses to mobile devices or allowing IP spoofing or device-to-device probing within the network [11], [12]. Signalling attacks may also be launched in conjunction with the presence of crowds in well identified locations such as sports arenas or concert venues [13].

Signalling attacks are similar to *signalling storms* caused by poorly designed or misbehaving mobile apps that repeatedly establish and tear down data connections [14], generating large amounts of signalling that may crash the network. Such signalling storms are a serious threat to the availability and security of cellular networks. While flash crowds last for a short time during special occasions such as New Year's Eve, signalling storms are unpredictable and tend to persist until the underlying problem is identified and corrected. This has prompted the industry to promote best practices for developing "network-friendly" mobile apps [15], [16].

A. Signalling Storms

Perhaps one of the most important features of smart phones and tablets is the "always-on" connectivity, which enables users to receive push messages, e.g. to notify of an incoming message or VoIP call. This is maintained by having the mobile device send periodic keep-alive messages to a cloud server. However, if for any reason the cloud service becomes unavailable, then the mobile device will attempt to reconnect more frequently generating signalling loads up to 20 times more than normal as reported in recent incidents [17]. In 2012 a Japanese mobile operator suffered a major outage [18] due to a VoIP app that constantly polls the network even when users are inactive. In another incident [19] the launch of a free version of a popular game on Android caused signalling overload in a large network due to frequent advertisements shown within the app. Also, many mobile carriers have reported [20] outages or performance issues caused by non-malicious but misbehaving apps, yet the majority of those affected followed a reactive approach to identify and mitigate the problem.

Signalling storms could also occur as a byproduct of large scale malware infections [21], such as botnets, which target mobile users rather than networks. A recent report by Kaspersky [22] revealed that the most frequently detected malware threats affecting Android OS are (i) SMS trojans which send costly messages without users' consent, (ii) adware which displays unwanted advertisements, and (iii) root exploits which allow the installation of other malware or the device to become part of a botnet. A sufficiently large number of users within a single network falling victims to such attacks, which involve frequent communications, could have a devastating impact on the control plane of the network.

The purpose of this paper is to analyse the effect of signalling storms, as well as of signalling attacks, and analyse in particular the manner in which such attacks can cause maximum damage to the radio and core networks. The approach we take is based on the development of a mathematical model of user signalling behaviour from which we derive some useful analytical results. While the literature [23]–[25] has focused on analysing signalling behaviour from an energy consumption perspective, we hope that this work can offer to mobile operators a greater understanding of bottlenecks and vulnerabilities in the radio signalling system, so that network parameters may be modified so as to mitigate for those effects that lead to network outages [26], [27].

II. Modelling Signalling of a Single User

In the context of UMTS networks, bandwidth is managed by the *radio resource control* (RRC) protocol which associates a state machine with each *user equipment* (UE). There are typically four RRC states, in order of increasing energy consumption: IDLE, Paging Channel (cell_PCH), low bandwidth Forward Access Channel (cell_FACH), and high bandwidth Dedicated Channel (cell_DCH). We will refer hereafter to state cell_X as X. State promotions are triggered by uplink (UL) and downlink (DL) transmissions, and the move to FACH or DCH is determined by the size of the *radio link control* (RLC) buffer of the UE: if at any time the buffer exceeds a certain threshold in either direction, the state will be promoted to DCH. State demotions are triggered by inactivity timers.

Consider a UE that transitions from IDLE or dormant D to FACH, perhaps later to DCH, and then sometimes directly from D to DCH. We will let λ_L and λ_H be the rates at which low and high bandwidth calls[1] are normally made, and L and H be the corresponding states when the call is actually taking place in the sense that it is using the bandwidth of FACH and DCH. Furthermore, we will denote by η the state when a low bandwidth request is handled while the mobile is in DCH.

At the end of normal usage the call will transition from L to ℓ or from H, η to h, where ℓ and h are the states when the UE is not using the bandwidth of FACH and DCH respectively; thus, $\{L, \ell\} \in$ FACH and $\{H, \eta, h\} \in$ DCH. We denote the rates at which low and high bandwidth calls terminate by μ_L and μ_H. Since the amount of traffic exchanged in states L and η is usually very small (otherwise it will trigger a transition to H), we assume that their durations are independent but stochastically identical.

If the UE does not start a new session for some time, it will be demoted from h to ℓ or from ℓ to PCH which we denote by P. The UE will then return from P to D after another inactivity timer; however, because the mobile is not allowed to communicate in the P state, it will first move to FACH, release all signalling connections, and finally move to D. Let τ_H, τ_L and τ_P be the time-out rates in states h, ℓ and P, respectively.

We are considering signalling attacks (or misbehaving apps) which falsely induce the mobile to go from D, P to FACH or DCH, or from FACH to DCH, without the user actually having any usage for this request. The rates related to these malicious transitions will be denoted α_L and α_H. Since in these cases a transition to an actual bandwidth usage state does not take place, unless the user starts a new session, the timers will demote the state of the UE. Consequently, the attack results in the usage of network resources both by the computation and state transitions that occur for call handling, and through bandwidth reservation that remains unutilised.

In summary, the state of the UE at time t is described by the variable $s(t) \in \{\mathcal{N}, \mathcal{A}, \mathcal{S}(\mathcal{N}), \mathcal{A}(\mathcal{N})\}$ where:

- $\mathcal{N} = \{D, P, \ell, L, h, \eta, H\}$ represent the states occupied by the UE during or after a "normal" call.

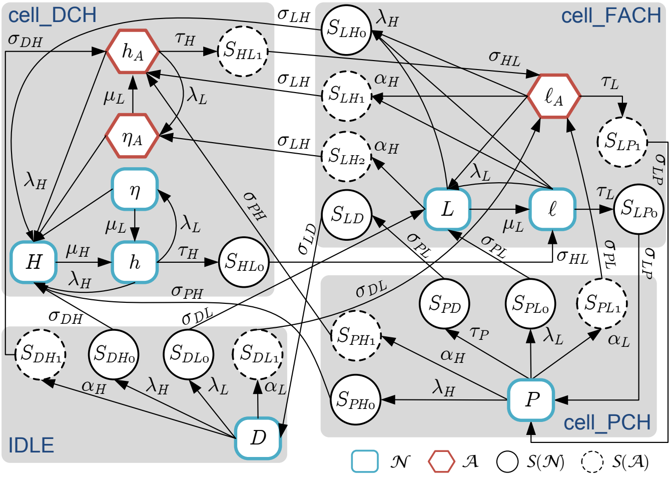

Fig. 1. The Markov model of a single user.

- $\mathcal{A} = \{\ell_A, h_A, \eta_A\}$ are similar to $\{\ell, h, \eta\}$ but forced by malicious traffic. Note that a transition to state η_A happens either from L because of an attack that forces an ongoing low bandwidth call to communicate over DCH, or from h_A because of a new normal low bandwidth call that could have been handled through FACH.
- $\mathcal{S}(\mathcal{N})$ and $\mathcal{S}(\mathcal{A})$ are the signalling states for respectively normal and attack conditions, which capture the non-negligible overhead needed in order to establish and release network resources during state promotions and demotions. We denote by σ_{XY}^{-1} the average transition delay from state X to Y, where $X, Y \in \{D, P, L, H\}$ and the subscripts L and H are used here to represent both normal and attack states in FACH and DCH.

Fig. 1 shows the different states, signalling phases and transitions of the Markov model. The stationary equations for the states in $\mathcal{N}$ are given by:

$$\begin{aligned}
&\pi(D)[\lambda_H + \alpha_H + \lambda_L + \alpha_L] = \pi(P)\tau_P, \\
&\pi(P)[\lambda_H + \alpha_H + \lambda_L + \alpha_L + \tau_P] = [\pi(\ell) + \pi(\ell_A)]\tau_L, \\
&\pi(\ell)[\lambda_H + \alpha_H + \lambda_L + \tau_L] = \pi(L)\mu_L + \pi(h)\tau_H, \\
&\pi(L)[\lambda_H + \alpha_H + \mu_L] = [\pi(D) + \pi(P) + \pi(\ell) + \pi(\ell_A)]\lambda_L, \\
&\pi(h)[\lambda_H + \lambda_L + \tau_H] = \pi(H)\mu_H + \pi(\eta)\mu_L, \\
&\pi(\eta)[\lambda_H + \mu_L] = \pi(h)\lambda_L, \\
&\pi(H)\mu_H = \sum_{s \in \{\mathcal{N}, \mathcal{A}\}, s \neq H} \pi(s)\lambda_H,
\end{aligned}$$

while the equations for the attack states $\mathcal{A}$ are:

$$\begin{aligned}
&\pi(\ell_A)[\lambda_H + \alpha_H + \lambda_L + \tau_L] = [\pi(D) + \pi(P)]\alpha_L + \pi(h_A)\tau_H, \\
&\pi(h_A)[\lambda_H + \tau_H + \lambda_L] = \sum_{s \in \{D, P, \ell, \ell_A\}} \pi(s)\alpha_H + \pi(\eta_A)\mu_L, \\
&\pi(\eta_A)[\lambda_H + \mu_L] = \pi(h_A)\lambda_L + \pi(L)\alpha_H.
\end{aligned}$$

We can express the normalisation condition as a weighted sum of the probabilities of the states $\{\mathcal{N}, \mathcal{A}\}$, i.e. $1 =$

[1] A call refers to any UL/DL activity, e.g. data session, location update, etc.

$\sum_{s\in\{\mathcal{N},\mathcal{A}\}} \pi(s)w_s$ or:

$$1 = \underbrace{\pi(D)[1+\frac{\Lambda_H}{\sigma_{DH}}+\frac{\Lambda_L}{\sigma_{DL}}]}_{\text{Pr[user in IDLE]}}$$
$$+ \underbrace{\pi(P)[1+\frac{\Lambda_H}{\sigma_{PH}}+\frac{\Lambda_L}{\sigma_{PL}}+\tau_P(\frac{1}{\sigma_{PL}}}_{\text{Pr[user in PCH]}}+\frac{1}{\sigma_{LD}})]$$
$$+ \underbrace{(\pi(\ell)+\pi(\ell_A))[1+\frac{\Lambda_H}{\sigma_{LH}}+\frac{\tau_L}{\sigma_{LP}}]+\pi(L)[1+\frac{\Lambda_H}{\sigma_{LH}}] + (\pi(h)+\pi(h_A))[1+\frac{\tau_H}{\sigma_{HL}}]+\pi(\eta)+\pi(\eta_A)+\pi(H)}_{\text{Pr[user in DCH]}} \quad (1)$$

with $\Lambda_H = \lambda_H + \alpha_H$ and $\Lambda_L = \lambda_L + \alpha_L$. Writing $\Lambda = \Lambda_L + \Lambda_H$, $q_L = \frac{\lambda_L}{\lambda_H+\mu_L}$, $\rho_L = \frac{\lambda_L}{\Lambda_H+\mu_L}$, and $q_H = \frac{\lambda_H}{\mu_H}$, the solution to the above set of equations becomes:

$$\begin{aligned}
\pi(D) =& \ \frac{\tau_P\tau_L}{(\Lambda+\tau_P)(\Lambda+\tau_L)} \ G,\\
\pi(P) =& \ \frac{\Lambda\tau_L}{(\Lambda+\tau_P)(\Lambda+\tau_L)} \ G,\\
\pi(L) =& \ \rho_L \ G,\\
\pi(H) =& \ q_H[\frac{q_L\rho_L\alpha_H}{\lambda_L}+(1+\rho_L)(\frac{\Lambda_H}{\tau_H}[1+q_L]+1)] \ G,\\
\pi(h) =& \ \frac{\mu_H}{\lambda_H[1+q_L]+\tau_H} \ \pi(H),\\
\pi(\eta) =& \ q_L \ \pi(h),\\
\pi(\ell) =& \ \frac{1}{\Lambda_H+\lambda_L+\tau_L}[\mu_L\rho_L G+\frac{\mu_H\tau_H\pi(H)}{\lambda_H[1+q_L]+\tau_H}],\\
\pi(h_A) =& \ \frac{\alpha_H}{\lambda_H[1+q_L]+\tau_H}[1+\frac{q_L\rho_L\mu_L}{\lambda_L}] \ G,\\
\pi(\eta_A) =& \ \frac{\alpha_H q_L}{\lambda_H[1+q_L]+\tau_H}[1+\frac{\lambda_H+\tau_H+\lambda_L}{\Lambda_H+\mu_L}]G,\\
\pi(\ell_A) =& \ \frac{1}{\Lambda_H+\lambda_L+\tau_L}[\frac{\alpha_L\tau_L}{\Lambda+\tau_L}+\frac{\alpha_H\tau_H(1+\frac{q_L\rho_L\mu_L}{\lambda_L})}{\lambda_H[1+q_L]+\tau_H}]G,
\end{aligned}$$

where G can be obtained from (1) yielding:

$$G^{-1} = [1+\rho_L][q_H+\frac{\Lambda_H}{\tau_H}\{(1+q_L)(1+q_H)+w_h-1\}]+$$
$$\frac{\frac{\tau_L}{\Lambda+\tau_P}[\tau_P w_D+\Lambda w_P]+\Lambda w_\ell}{\Lambda+\tau_L}+\rho_L[w_L+\frac{q_L}{\lambda_L}(1+q_H)\alpha_H].$$

A. Signalling Load on the RNC and SGSN

Let n_{XY} denote the number of signalling messages sent or received by the *radio network controller* (RNC) when a transition occurs from state X to state Y, then the signalling rate generated by a single user due to both normal and malicious traffic can be computed as:

$$\begin{aligned}
\gamma_r \ =& \ \pi(D)[\Lambda_H n_{DH}+\Lambda_L n_{DL}]+\pi(P)[\Lambda_H n_{PH}+\Lambda_L n_{PL}]\\
&+[\pi(\ell)+\pi(\ell_A)+\pi(L)]\Lambda_H n_{LH}\\
&+[\pi(h)+\pi(h_A)]\tau_H n_{HL}\\
&+[\pi(\ell)+\pi(\ell_A)]\tau_L\{n_{LP}\mathbf{1_{L\to P}}+n_{LD}\mathbf{1_{L\to D}}\}\\
&+\pi(P)\tau_P n_{PD}\mathbf{1_{L\to P}}, \qquad (2)
\end{aligned}$$

where the characteristic function $\mathbf{1_{X\to Y}}$ takes the value 1 if the transition $X \to Y$ is implemented and 0 otherwise. Note that the mobile network operator may not use PCH state, e.g. when the vendor does not support it or it is disabled in order to extend the battery life of mobile devices. In this case, σ_{PL}, σ_{LP} and τ_P are set to ∞ so that the user is moved directly from FACH to IDLE after an inactivity timer.

TABLE I
NETWORK PARAMETERS

Transition	**Triggering Event**	$\mathbf{n_{XY}}$	$\mathbf{m_{XY}}$	$\sigma^{-1}_{\mathbf{XY}}$
IDLE → FACH	Low bandwidth UL/DL traffic	15	5	0.75
PCH → FACH	(e.g. location update, keep-alive	3	–	0.15
	messages)			
IDLE → DCH	High bandwidth UL/DL traffic	20	5	1.0
PCH → DCH	(e.g. VoIP calls, video	10	–	0.5
FACH → DCH	streaming, web browsing)	7	–	0.35
DCH → FACH	inactivity timer $\tau_H^{-1} = 2-10$s	5	–	0.25
FACH → PCH	inactivity timer $\tau_L^{-1} = 2-10$s	2	–	0.1
PCH $\xrightarrow{\text{FACH}}$ IDLE	inactivity timer $\tau_p^{-1} = 5-20$min	6	2	0.3

On the other hand, the core network is more protected from signalling attacks since only transitions to/from state D trigger signalling with the core. Let $m_{XY} \leq n_{XY}$ be the number of control plane messages exchanged between the RNC and the *serving GPRS support node* (SGSN) during such transitions, then the signalling load on the core network from a single user becomes:

$$\begin{aligned}
\gamma_c \ =& \ \pi(D)[\Lambda_H m_{DH}+\Lambda_L m_{DL}]+\pi(P)\tau_P m_{PD}\mathbf{1_{L\to P}}\\
&+[\pi(\ell)+\pi(\ell_A)]\tau_L m_{LD}\mathbf{1_{L\to D}}. \qquad (3)
\end{aligned}$$

Table I summarises the state transition model along with parameter values used in the numerical results: (i) the number of signalling messages exchanged during state transitions are obtained from the UMTS standards documentation, and can also be found in the literature (e.g. [28]); (ii) typical values for the inactivity timers τ_H^{-1} and τ_L^{-1} are in the range $2-10$ seconds, while τ_P^{-1} should be significantly longer (in the order of minutes); and (iii) the average transition times are assumed to be proportional to the number of signalling messages involved, and normalised with respect to the transition IDLE → DCH which is assumed to take 1 second.

III. MAXIMISING THE IMPACT OF AN ATTACK

If an attacker succeeds in inferring the radio network configuration parameters (e.g. through active probing [9]–[11]), then it is easy to monitor the user's behaviour in order to estimate $\lambda_L, \lambda_H, \mu_L$ and μ_H. The attacker can then maximise the impact on the radio or core network by choosing the rate of malicious traffic bursts α_L and α_H so as to maximise (2) or (3). This is illustrated in Fig. 2 where we plot the average rate of signalling messages that a misbehaving user generates on the RNC and SGSN assuming $\alpha_L = 0$ and different values of α_H. The results indicate that there is indeed an optimum value of α_H which maximises the load on the *core network*, while the load on the radio network increases monotonically with the attack rate up to a maximum level.

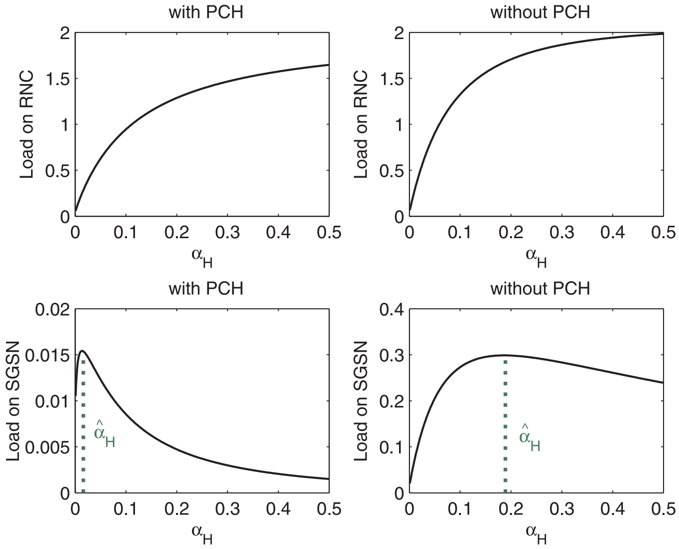

Fig. 2. The average signalling load (msg/s) on RNC and SGSN with and without PCH state versus attack rate α_H when $\alpha_L = 0$, normal traffic is characterised by $\lambda_L^{-1} = 600, \mu_L^{-1} = 5, \lambda_H^{-1} = 1800, \mu_H^{-1} = 120$, and using the parameters of Table I with $\tau_H^{-1} = \tau_L^{-1} = 5$s and $\tau_P^{-1} = 5$min. $\hat{\alpha}_H$ in (6) provides a good estimate of the optimum attack rate.

The effect of PCH state is also examined in Fig. 2 showing a significant reduction (about 95%) in the amount of control plane traffic reaching the core network as compared to the case where the user is moved directly from FACH to IDLE. In fact, as the value of the timer τ_P^{-1} gets larger, an attacker would find it extremely difficult to overwhelm the SGSN with signalling load unless a very large number of UEs are compromised. This feature also results in up to 30% drop in the amount of signalling load traversing the radio network.

A. Radio Network

Numerical investigations suggest that the load on the radio network increases with the frequency of the malicious bursts up to a maximum level reached when *either* α_H or α_L tends to infinity, depending on the parameters of the network as well as the user's traffic characteristics. If PCH is enabled then the attacker could either induce the transition FACH → DCH as soon as the channel is released, or take a two-step approach to first move from PCH to FACH immediately after the timer τ_L^{-1} expires then trigger another transition to DCH some time later. Note that any other attack policy would be slowed down by the long timer τ_P^{-1} and thus would not succeed in creating a more severe impact. To investigate both policies, let us set $\alpha_L \to \infty$ so that the transition PCH → FACH is triggered repeatedly, creating a load on the radio network given by:

$$\begin{aligned}\gamma_r \; = \; & \frac{n_{LH} + n_{HL}}{\theta_{LH} + \frac{\tau_L(\Lambda_H+\mu_L)}{\Lambda_H(\Lambda_H+\lambda_L+\mu_L)}\;\theta_{PL}} \\ & + \frac{n_{PL} + n_{LP}}{\theta_{PL} + \frac{\Lambda_H(\Lambda_H+\lambda_L+\mu_L)}{\tau_L(\Lambda_H+\mu_L)}\;\theta_{LH}}, \qquad \alpha_L \to \infty,\end{aligned}$$

where $\theta_{XY} = \sigma_{XY}^{-1} + (1+q_L)(1+q_H)\tau_Y^{-1} + \sigma_{YX}^{-1}$. Now if we maximise the above expression with respect to α_H, we obtain the following interesting result:

$$(\alpha_L^*, \alpha_H^*) = \begin{cases} (\infty, 0), & \text{if } \frac{n_{LH}+n_{HL}}{\theta_{LH}} \le \frac{n_{PL}+n_{LP}}{\theta_{PL}}, \\ (0, \infty), & \text{otherwise.} \end{cases} \tag{4}$$

Therefore, the load on the radio network can be maximised through low (resp. high) bandwidth bursts that repeatedly induce the transition PCH → FACH (resp. FACH → DCH) if the condition $[n_{LH}+n_{HL}]\theta_{LH}^{-1} \le [n_{PL}+n_{LP}]\theta_{PL}^{-1}$ is (resp. is not) satisfied. When PCH state is not used, we obtain similar results, but the attack is maximised by continuously triggering IDLE → FACH or FACH → DCH depending on whether the condition $[n_{LH} + n_{HL}]\theta_{LH}^{-1} \le [n_{DL} + n_{LD}]\theta_{DL}^{-1}$ is satisfied or not, respectively. The worst case load on the RNC is then:

$$\gamma_r^* = \max\left[\frac{n_{XL} + n_{LX}}{\theta_{XL}}, \frac{n_{LH} + n_{HL}}{\theta_{LH}}\right],$$

$X = D$ or P depending on which transition $L \to X$ is used.

B. Core Network

Signalling between the UE and core network happens for a number of different reasons, but with respect to the RRC state machine, it usually occurs when the UE moves from/to the IDLE state. The attack against the core network can then be launched more effectively by causing a transition to FACH, rather than DCH, immediately after the user becomes IDLE so as to avoid the timer τ_H^{-1} and the associated demotion delay. Thus, optimally $\alpha_H^* = 0$, and the attack rate that maximises the load on the core network can be shown to be:

$$\alpha_L^* = \sqrt{c^2 + \frac{b - ca}{\theta_{PLH}}} - c - \lambda_L, \tag{5}$$

where:

$$\begin{aligned}&\theta_{PLH} = \theta_{PL} + (1 + q_L)\tau_L^{-1}\lambda_H\theta_{LH}, \\ &a = \lambda_H[2\theta_{PLH} + \sigma_{DH}^{-1} - \sigma_{PL}^{-1} - \sigma_{LH}^{-1}] \\ &\quad + \tau_P[\theta_{PLH} + \sigma_{DL}^{-1} + \sigma_{LD}^{-1}] + (1 + q_L)(1 + q_H + \lambda_H\theta_{LH}), \\ &b = \lambda_H^2[\theta_{PLH} + \sigma_{PH}^{-1} - \sigma_{PL}^{-1} - \sigma_{LH}^{-1}] \\ &\quad + \lambda_H\tau_P[\theta_{PLH} + \sigma_{DH}^{-1} + \sigma_{LD}^{-1} - \sigma_{LH}^{-1}] \\ &\quad + (\lambda_H + \tau_P)(1 + q_L)(1 + q_H + \lambda_H\theta_{LH}), \\ &c = \lambda_H \; \frac{m_{DH} + m_{PD}}{m_{DL} + m_{PD}}.\end{aligned}$$

Obviously, the attack is worst when there is no background high bandwidth user traffic, in which case we end up with:

$$\alpha_L^* = \sqrt{\frac{\tau_P[1 + \frac{\lambda_L}{\mu_L}]}{\theta_{PL}}} - \lambda_L, \qquad \lambda_H = 0$$

and consequently the maximum possible load that an attacker can impose on the SGSN is:

$$\gamma_c^* = \frac{m_{DL} + m_{PD}}{\sigma_{DL}^{-1} + \sigma_{LP}^{-1} + \sigma_{PL}^{-1} + \sigma_{LD}^{-1} + (1 + \frac{\lambda_L}{\mu_L})(\frac{1}{\tau_L} + \frac{1}{\tau_P} + \frac{2}{\Lambda_L^*})}$$

with $\Lambda_L^* = \alpha_L^* + \lambda_L$. When $\tau_P \to \infty$, we get the intuitive result $\alpha_L^* = \infty$, i.e. the attacker should send a low-bandwidth traffic burst as soon as the timer τ_L^{-1} expires, leading to $\gamma_c^* = [m_{DL} + m_{LD}]\theta_{DLH}^{-1}$ where $\theta_{DLH} = \theta_{DL} + (1 +$

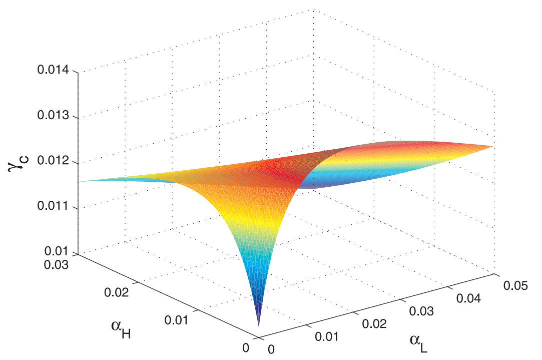

Fig. 3. The average signalling load (msg/s) on SGSN versus the attack rates α_H, α_L, when normal traffic profile is $\lambda_L^{-1} = 300, \mu_L^{-1} = 5, \lambda_H^{-1} = 600, \mu_H^{-1} = 180$, and the timers are $\tau_H^{-1} = \tau_L^{-1} = 5$s and $\tau_P^{-1} = 5$ min.

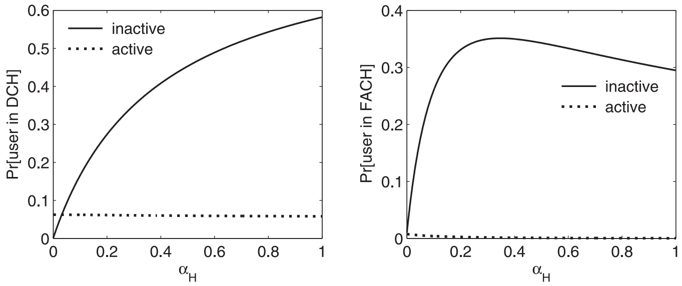

Fig. 4. The fraction of time the UE spends in DCH and FACH waiting for a timer or state transition (solid line) and while using the bandwidth (dotted line) as a function of α_H, when $\alpha_L = 0, \lambda_L^{-1} = 600, \mu_L^{-1} = 5, \lambda_H^{-1} = 1800, \mu_H^{-1} = 120, \tau_H^{-1} = 2$s, $\tau_L^{-1} = 5$s and $\tau_P^{-1} = 10$ min. Large inactive times indicate anomalous signalling behaviour.

$q_L)\tau_L^{-1}\lambda_H\theta_{LH}$. In Fig. 3 we plot γ_c versus the attack rates, and the numerical results indicate that $(\alpha_L^*, \alpha_H^*) = (0.02, 0)$ which coincide with the prediction of (5).

In practice, however, mounting an attack based solely on low bandwidth bursts may not be feasible. To begin with, it may be difficult to accurately estimate the RLC buffer's thresholds which determine whether a session will be handled through the low or high speed channel, and also the thresholds could differ from one RNC to another. Furthermore, many operators choose to move users directly into DCH or use very small RLC thresholds such that even keep-alive messages are sent over the high speed channel [10]. Thus, a more practical approach for an attacker is to assume that the majority of data transmissions are handled through DCH, and in turn compute an attack rate $\hat{\alpha}_H$ that maximises the load on the SGSN under such circumstances, i.e.:

$$\hat{\alpha}_H = \arg\max_{\alpha_H} \quad \gamma_c, \qquad \text{when} \quad \Lambda_H >> \Lambda_L,$$

yielding:

$$\begin{aligned} \hat{\alpha}_H \; = \; & \sqrt[3]{-\frac{B}{2} + \sqrt{\frac{B^2}{4} + \frac{A^3}{27}}} + \sqrt[3]{-\frac{B}{2} - \sqrt{\frac{B^2}{4} + \frac{A^3}{27}}} \\ & - \frac{b}{6a} - \lambda_H, \end{aligned} \tag{6}$$

where:

$$A = -\frac{b^2}{12a^2}, \quad B = \frac{b^3}{108a^3} - \frac{c}{2a},$$

$$a = \sigma_{LH}^{-1} + [1 + \frac{\lambda_H}{\mu_H}]\tau_H^{-1} + \sigma_{HL}^{-1},$$

$$b = \tau_L(\sigma_{PH}^{-1} + [1 + \frac{\lambda_H}{\mu_H}][\tau_H^{-1} + \tau_L^{-1}] + \sigma_{HL}^{-1} + \sigma_{LP}^{-1}) + \tau_P a,$$

$$c = \tau_L\tau_P[1 + \frac{\lambda_H}{\mu_H}].$$

When PCH is disabled, we have:

$$\hat{\alpha}_H = \sqrt{\frac{\tau_L[1 + \frac{\lambda_H}{\mu_H}]}{\sigma_{LH}^{-1} + [1 + \frac{\lambda_H}{\mu_H}]\tau_H^{-1} + \sigma_{HL}^{-1}}} - \lambda_H$$

and the resulting load on the SGSN becomes:

$$\hat{\gamma}_c = \frac{m_{DH} + m_{LD}}{\sigma_{DH}^{-1} + \sigma_{HL}^{-1} + \sigma_{LD}^{-1} + (1 + \frac{\lambda_H}{\mu_H})(\frac{1}{\tau_H} + \frac{1}{\tau_L} + \frac{2}{\hat{\alpha}_H + \lambda_H})}.$$

Fig. 2 shows that $\hat{\alpha}_H$ provides a good estimate of the optimum value α_H^* even when $\lambda_L > \lambda_H$.

Fig. 4 illustrates the manner in which the frequency of malicious traffic bursts affects signalling overhead as well as the *tail* which is the time the UE spends in FACH or DCH waiting for a time-out to expire. During these inactive periods, the mobile wastes considerable radio resources in the network as well as its own limited battery energy. As the attack rate increases, the proportion of time the UE remains inactive in either FACH or DCH also increases, while its average data volume is almost constant. This observation could be used by anomaly detection techniques to distinguish between normal "heavy" users and attackers: the former can be recognised by their low inactive times, while the latter can be detected by frequent connection attempts and low data volume.

Finally, we examine in Fig. 5 the effect of a signalling storm on the RNC and SGSN when the total number of UEs is 10,000 and the percentage of misbehaving ones is increased from 0 to 20%. Comparing the maximum load on the targeted network component and the corresponding load on the other, we see that PCH state prevents a situation where both the RNC and SGSN are simultaneously exposed to worst case loads, which happens when IDLE→ FACH is the bottleneck transition in the radio network (cf. Section III-A). In general, the radio network is less sensitive to the choice of the malicious bursts, as long as they are frequent, and thus it is more vulnerable to signalling storms. On the other hand, the load on the core network changes dramatically when the storm is optimised, which may not happen often, making signalling overloads in the SGSN a less likely event. This does not, however, include the effect of complex pricing and business models used by the operator which may exacerbate signalling load in the core network.

IV. Conclusions

This paper has focused on the behaviour of a mobile network user with a view to determining network overload

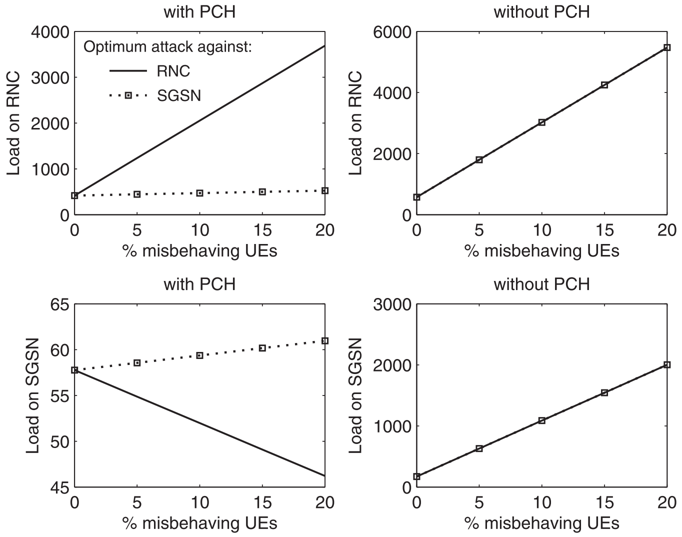

Fig. 5. Load on RNC and SGSN versus percentage of mobile devices participating in a storm out of 10,000 users, when $\lambda_L^{-1} = 600, \mu_L^{-1} = 5, \lambda_H^{-1} = 600, \mu_H^{-1} = 180, \tau_H^{-1} = \tau_L^{-1} = 5$s and $\tau_P^{-1} = 10$ min. When PCH is enabled a storm can cause maximum load on *either* the radio or core network, but without PCH both of them could be targeted simultaneously.

in signalling servers and base stations that can result from signalling misbehaviours such as signalling storms. Such misbehaviours can be caused by poorly designed mobile apps, outages in cloud services, large scale malware infections, or malicious network attacks. In the course of this work we have derived a Markov model of user behaviour that can also be exploited in other studies concerning mobile networks as a whole. The Markov model has been solved analytically, and used to derive conditions and parameters for which the signalling misbehaviours can cause the largest damage and which therefore need to be avoided. The analytical results have been illustrated with several numerical examples, and we expect that this work will lead to ideas relating to control algorithms that can adaptively react to network measurements so as to eliminate or mitigate the effect of signalling storms and DoS attacks.

Acknowledgment

The authors acknowledge the support of the EU FP7 project NEMESYS, Grant Agreement no. 317888.

References

[1] J. Serror, H. Zang, and J. C. Bolot, "Impact of paging channel overloads or attacks on a cellular network," in *Proc. 5th ACM W'shop Wireless Security (WiSe'06)*, LA, CA, Sep 2006, pp. 75–84.

[2] W. Enck, P. Traynor, P. McDaniel, and T. La Porta, "Exploiting open functionality in SMS-capable cellular networks," in *Proc. 12th ACM Conf. Computer and Communications security (CCS'05)*, Alexandria, VA, Nov 2005, pp. 393–404.

[3] P. P. C. Lee, T. Bu, and T. Woo, "On the detection of signaling DoS attacks on 3G/WiMax wireless networks," *Comput. Netw.*, vol. 53, no. 15, pp. 2601–2616, Oct 2009.

[4] F. Ricciato, A. Coluccia, and A. D'Alconzo, "A review of DoS attack models for 3G cellular networks from a system-design perspective," *Comput. Commun.*, vol. 33, no. 5, pp. 551–558, Mar 2010.

[5] E. Gelenbe and G. Loukas, "A self-aware approach to denial of service defence," *Comput. Netw.*, vol. 51, no. 5, pp. 1299–1314, Apr 2007.

[6] E. Gelenbe, "Steps towards self-aware networks," *Commun. ACM*, vol. 52, no. 7, pp. 66–75, Jul 2009.

[7] ——, "Dealing with software viruses: A biological paradigm," *Inf. Secur. Tech. Rep.*, vol. 12, no. 4, pp. 242–250, Sep 2007.

[8] O. H. Abdelrahman and E. Gelenbe, "Packet delay and energy consumption in non-homogeneous networks," *Comput. J.*, vol. 55, no. 8, pp. 950–964, Aug 2012.

[9] A. Barbuzzi, F. Ricciato, and G. Boggia, "Discovering parameter setting in 3G networks via active measurements," *IEEE Commun. Lett.*, vol. 12, no. 10, pp. 730–732, Oct 2008.

[10] F. Qian *et al.*, "Characterizing radio resource allocation for 3G networks," in *Proc. 10th Internet Measurement Conf. (IMC'10)*, Melbourne, Australia, Nov 2010, pp. 137–150.

[11] Z. Qian *et al.*, "You can run, but you can't hide: Exposing network location for targeted DoS attacks in cellular networks," in *Proc. Network and Distributed System Security Symp. (NDSS'12)*, San Diego, CA, Feb 2012, pp. 1–16.

[12] Z. Wang *et al.*, "An untold story of middleboxes in cellular networks," in *Proc. ACM SIGCOMM*, Toronto, Canada, Aug 2011, pp. 374–385.

[13] E. Gelenbe and F.-J. Wu, "Large scale simulation for human evacuation and rescue," *Comput. Math. Appl.*, vol. 64, no. 12, pp. 3869–3880, 2012.

[14] Nokia Siemens Networks Smart Labs, "Understanding smartphone behavior in the network," White paper, Jan 2011. [Online]. Available: http://www.nokiasiemensnetworks.com/sites/default/files/document/Smart_Lab_WhitePaper_27012011_low-res.pdf

[15] GSMA, "Smarter apps for smarter phones!" Feb 2012. [Online]. Available: http://www.gsma.com/technicalprojects/wp-content/uploads/2012/04/gsmasmarterappsforsmarterphones0112v.0.14.pdf

[16] S. Jiantao, "Analyzing the network friendliness of mobile applications," Huawei, Tech. Rep., Jul 2012.

[17] G. Reddig, "OTT service blackouts trigger signaling overload in mobile networks," Sep 2013. [Online]. Available: http://blogs.nsn.com/mobile-networks/2013/09/16/ott-service-blackouts-trigger-signaling-overload-in-mobile-networks/

[18] Rethink Wireless, "DoCoMo demands Google's help with signalling storm," Jan 2012. [Online]. Available: http://www.rethink-wireless.com/2012/01/30/docomo-demands-googles-signalling-storm.htm

[19] S. Corner, "Angry birds + android + ads = network overload," Jun 2011. [Online]. Available: http://www.itwire.com/business-it-news/networking/47823

[20] Arbor Networks, "Worldwide infrastructure security report," 2012. [Online]. Available: http://www.arbornetworks.com/research/infrastructure-security-report

[21] F. Ricciato, P. Svoboda, E. Hasenleithner, and W. Fleischer, "On the impact of unwanted traffic onto a 3G network," in *Proc. 2nd Int. W'shop Security, Privacy and Trust in Pervasive and Ubiquitous Computing (SecPerU'06)*, Lyon, France, Jun 2006, pp. 49–56.

[22] D. Maslennikov, "Mobile malware evolution: Part 6," Kaspersky Lab, Tech. Rep., Feb 2013. [Online]. Available: http://www.securelist.com/en/analysis/204792283/Mobile_Malware_Evolution_Part_6

[23] H. Haverinen, J. Siren, and P. Eronen, "Energy consumption of always-on applications in WCDMA networks," in *Proc. 65th IEEE Vehicular Technology Conf. (VTC'07-Spring)*, Dublin, Ireland, Apr 2007, pp. 964–968.

[24] J.-H. Yeh, J.-C. Chen, and C.-C. Lee, "Comparative analysis of energy-saving techniques in 3GPP and 3GPP2 systems," *IEEE Trans. Veh. Technol.*, vol. 58, no. 1, pp. 432–448, Jan 2009.

[25] C. Schwartz *et al.*, "Smart-phone energy consumption vs. 3G signaling load: The influence of application traffic patterns," in *Proc. 24th Tyrrhenian Int. W'shop Digital Communications (TIWDC'13)*, Genoa, Italy, Sep 2013, pp. 1–6.

[26] O. H. Abdelrahman, E. Gelenbe, G. Görbil, and B. Oklander, "Mobile network anomaly detection and mitigation: The NEMESYS approach," in *Proc. 28th Int. Symp. Computer and Information Sciences (ISCIS'13)*, ser. LNEE, vol. 264. Paris, France: Springer, 2013, pp. 429–438.

[27] E. Gelenbe *et al.*, "NEMESYS: Enhanced network security for seamless service provisioning in the smart mobile ecosystem," in *Proc. ISCIS'13*, ser. LNEE, vol. 264. Springer, 2013, pp. 369–378.

[28] GSMA, "Fast dormancy best practises," White paper, Jul 2011. [Online]. Available: http://www.gsma.com/newsroom/ts18-v10-tsg-prd-fast-dormancy-best-practices