Storms in Mobile Networks

Gokce Gorbil
g.gorbil@imperial.ac.uk

Omer H. Abdelrahman
o.abd06@imperial.ac.uk

Erol Gelenbe
e.gelenbe@imperial.ac.uk

Imperial College London
Department of Electrical and Electronic Engineering
Intelligent Systems and Networks Group
SW7 2AZ, London, United Kingdom

ABSTRACT

Mobile networks are vulnerable to signalling attacks and storms caused by traffic that overloads the control plane through excessive signalling, which can be introduced via malware and mobile botnets. With the advent of machine-to-machine (M2M) communications over mobile networks, the potential for signalling storms increases due to the normally periodic nature of M2M traffic and the sheer number of communicating nodes. Several mobile network operators have also experienced signalling storms due to poorly designed applications that result in service outage. The radio resource control (RRC) protocol is particularly susceptible to such attacks, motivating this work within the EU FP7 NEMESYS project which presents simulations that clarify the temporal dynamics of user behavior and signalling, allowing us to suggest how such attacks can be detected and mitigated.

Categories and Subject Descriptors

C.2.0 [**Computer-Communication Networks**]: General—*security and protection*; I.6.3 [**Simulation and Modeling**]: Applications; K.6.5 [**Management of Computing and Information Systems**]: Security and Protection

General Terms

Performance, Security

Keywords

Network attacks; malware; UMTS networks; 3G to 5G; signalling storms; network simulation; performance analysis

1. INTRODUCTION

Smart devices for mobile networks are becoming ubiquitous worldwide [1] and cover a large majority of mobile traffic [2]. As a consequence, cyber-criminals target mobile platforms [3–5], and mobile network operators (MNOs) now face increasing security issues [6], including signalling attacks and storms, which overload the control plane through excessive signalling [7–11] caused by smartphone malware. In signalling attacks, mobile services are disrupted through a distributed denial-of-service-like attack [12] that overloads the control plane rather than the data plane, and a large number of mobile devices may be compromised to form a botnet [13] as witnessed in 2012 [14]. Though such attacks can be mitigated by signalling-aware routing and control algorithms [15] in the core network, the standard UMTS environment currently does not offer such facilities, and various techniques can be used to render the attack more effective [16–20]. Since smart mobile devices are also used in emergencies [21–23], they can be also used to compromise the safety of emergency responders and evacuees [24].

On the other hand, signalling storms are caused by misbehaving mobile apps that frequently set-up and tear-down data connections, as exemplified in [25]. In another event, the excessive communications of in-game advertisements [26] have increased the need to create safe and network-friendly apps [27,28]. Such known events show that signalling attacks and storms do cause major outages, and motivate the MNOs to protect their customers from malware and to detect and mitigate signalling attacks [6,11]. The introduction of machine-to-machine (M2M) communications over mobile networks also poses a challenge since the signalling protocols in 3G and 4G mobile networks were not designed to handle frequent but small communications from a huge number of devices as is the case in M2M traffic.

Our initial research has shown that the *radio resource control* (RRC) protocol in UMTS networks [29] is susceptible to signalling attacks, as we discuss in more detail in the next section. The objective of this paper is to analyze the effect of RRC-based signalling attacks and storms in UMTS networks, in particular the manner in which such attacks cause the most load on the network. For this purpose, we develop a simulation model of a UMTS network, and present results from simulation experiments, which enable us to better understand the temporal dynamics of user behavior and signalling in the network, and to evaluate the impact of signalling storms on quality-of-experience which is not captured by the mathematical model we have developed earlier [30]. While similar work has focused on analyzing signalling behavior from an energy perspective [31–33], we hope to provide a greater understanding of the bottlenecks and vulnerabilities in the radio signalling system of mobile networks in order to pave the way for the detection and mitigation of signalling attacks and storms.

Q2SWinet'14, September 21–26, 2014, Montreal, QC, Canada.
Copyright is held by the owner/author(s). Publication rights licensed to ACM.
ACM 978-1-4503-3027-5/14/09 ...$15.00.
http://dx.doi.org/10.1145/2642687.2642688.

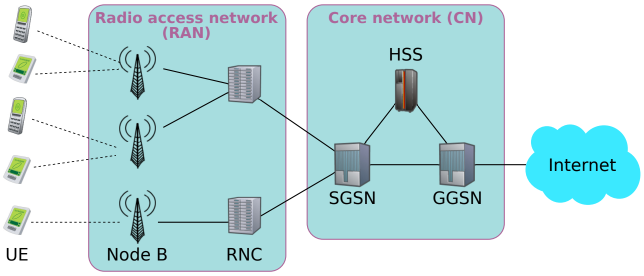


Figure 1: The basic architecture of a UMTS network. UEs are the mobile terminals, e.g. smartphones, connected to the mobile network via base stations (Node Bs). Node Bs maintain the radio channels with the UEs. The RNC controls the radio resources and the Node Bs in the RAN.

1.1 The Radio Resource Control Protocol

In UMTS networks, the radio resource control (RRC) protocol is used to manage resources in the radio access network (RAN) [29]. It operates between the UMTS terminals, i.e. the user equipment (UE), and the radio network controller (RNC). Figure 1 shows the basic architecture of a UMTS network, depicting the RAN and the core network (CN) elements comprising the packet-switched domain of the mobile network. The RNC is the switching and controlling network element in the RAN. It performs radio resource management (RRM) functions in order to guarantee the stability of the radio path and the QoS of radio connections by efficient sharing and management of radio resources. The RRC protocol is utilized for all RRM-related control functions such as the setup, configuration, maintenance and release of radio bearers between the UE and the RNC. The RRC protocol also carries all non-access stratum signalling between the UE and the CN.

In order to manage the radio resources, the RRC protocol associates a *state machine* to each UE, which is maintained synchronized at the UE and the RNC via RRC signalling messages. The RNC controls the transitions between the RRC states based on information it receives from the UEs and the Node Bs on available radio resources, conditions of the currently used radio bearers, and requests for communication activity. As shown in Fig. 2, there are typically four RRC states, given in order of increasing energy consumption and data rate: *idle, cell-PCH, cell-FACH* and *cell-DCH*[1]. In the rest of this paper, we refer to state *cell-X* simply as *X*. Whenever the UE is not in the idle state, it is in *connected mode* and has a signalling connection with the RNC. In connected mode, the location of the UE is known by the RNC at the level of a single cell, which is maintained by *cell updates* sent by the UE either periodically or when it changes cells. We describe the RRC states in more detail below.

Idle: This is the initial state when the UE is turned on. In this state, the UE does not have a signalling connection with the RNC, and therefore the RNC does not know the location of the UE. Its location is known by the CN at the accuracy of the location area or routing area, which is based on the latest mobility signalling the UE performed with the CN. Any downlink activity destined for a UE in idle mode will require *paging* in order to locate the UE at the cell level. Since the UE does not have an RNC connection, it cannot send any signalling or data until an RNC connection has been established.

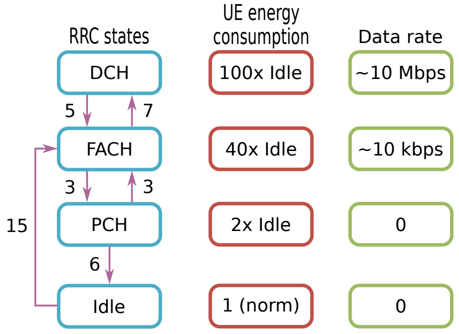


Figure 2: RRC states. The figure on the left shows the typical number of signalling messages exchanged within the RAN for each transition. The other figures show the approximate energy consumption and maximum data rate at the UE.

FACH: The UE is in connected mode, and the radio connection between the UE and the RNC uses only common channels which allow low-rate data transmission.

DCH: The UE is in connected mode, and the radio connection uses resources dedicated to the UE. While in DCH, the UE may use shared channels, dedicated channels or both. The data rate of the connection is significantly higher than the FACH state, but energy use is also higher.

PCH: This is a low-energy state that allows the UE to maintain its RNC connection and thus stay in connected mode, but it cannot send or receive any traffic while in this state. While in PCH, the UE listens to paging occasions on the paging channel. This state is optional and it can be enabled or disabled by the MNO according to their policies. Although the PCH state is a low-energy state, the UE still consumes more power than in the idle state. Therefore, some MNOs choose to disable the PCH state in order to allow the UE to return to idle mode quickly and thus reduce its energy consumption. We will investigate the effect of the PCH state on signalling load in Sec. 3.

State demotions from a higher to a lower state, e.g. DCH→FACH, occur based on radio bearer inactivity timers at the RNC. The exact order of state demotions is dependent on MNO policy, but a progression as shown in Fig. 2 is common, although some MNOs skip the FACH and/or PCH states. State promotions from the idle and PCH states occur depending on uplink and downlink activity. For example, when the UE has uplink data to send, it sends an "RNC connection request" if in idle, or a "cell update" if in PCH, to the RNC in order to move to a state where it can send and receive data. Whether the UE is promoted to the FACH or DCH state is dependent on MNO policy. A FACH→DCH transition is performed based on buffer occupancy of the uplink and downlink radio links as observed by the RNC.

Table 1 summarizes when RRC state transitions occur and the number of signalling messages exchanged to effect each transition. In our simulations, we assume the RRC state progression given in Fig. 2. The UE goes from idle to FACH initially, and then to DCH if the buffer threshold is reached. The UE goes from DCH to FACH upon demotion from DCH. Whether the UE goes from FACH to PCH, or to

[1] PCH: Paging channel; FACH: Forward access channel; DCH: Dedicated channel

Table 1: RRC state transitions, number of signalling messages exchanged, and service times at the RRC signalling server of the RNC

Transition	**Triggering event**	$\mathbf{r_{xy}}$	$\mathbf{c_{xy}}$	$\delta^{\mathbf{r}}_{\mathbf{xy}}$ (ms)
Idle→FACH	Uplink or downlink traffic	15	5	75
PCH→FACH	Uplink or downlink traffic	3	-	15
FACH→DCH	Radio link buffer threshold (Θ) reached, $\Theta = 1500$ B	7	-	35
DCH→FACH	Expiry of inactivity timer $T_1 = 6$s	5	-	25
FACH→Idle	Expiry of inactivity timer $T_2 = 12$s, PCH disabled	5	3	15
FACH→PCH	Expiry of inactivity timer $T_2 = 4$s, PCH enabled	3	-	10
PCH→Idle	Expiry of inactivity timer $T_3 = 20$min, PCH enabled	6	3	30

idle, depends on whether the PCH state is enabled. For an $x \rightarrow y$ transition, we use r_{xy} and c_{xy} to denote the number of signalling messages exchanged within the RAN and between the RAN and the CN, respectively.

The RRC protocol was designed to manage the limited radio resources among multiple UEs and to decrease energy use at the UE. It is therefore biased towards demoting the UE to a lower state as soon as possible, especially if the UE is in the DCH or FACH state. Indeed, as the number of smartphones accessing UMTS networks has increased, the industry has introduced improvements and changes in order to get more data rate out of limited radio resources, such as HSDPA and HSUPA, and to improve the energy use of smartphones. For example, fast dormancy enables the UE to indicate to the RNC when it has no more uplink data to send for a speedier demotion to the PCH or idle state. In addition, some MNOs choose to disable the PCH state in order to allow the UE to return to idle mode quickly and thus reduce its energy consumption. As we will discuss in Sec. 3, this tendency to perform hasty RRC demotions result in excessive signalling load in the mobile network, especially in the case of deliberate attacks or signalling storms that result from poorly designed applications.

The RNC will customarily release radio resources for a UE soon after activity ceases in its channel, making those resources available for other UEs. Thus it uses short inactivity timers, which are in the order of 2–10 seconds (see Table 1). These short timers make the RRC protocol susceptible to signalling attacks, as an attacker that approximately determines the values of the T_1 and T_2 timers can then launch a devastating attack from a relatively small number of compromised UEs, as we discuss in Sec. 3. In addition, when combined with the "chatty" nature of many mobile applications, the tendency to deallocate radio channels quickly necessarily leads to increased RRC signalling in order to reconfigure or setup channels that were released a short time ago, rendering the mobile network vulnerable to RRC based signalling storms.

We thus focus on the RRC protocol in order to better understand its signalling behavior, and investigate under which conditions signalling load becomes excessive. Section 2 describes our simulation model of UMTS networks. In Sec. 3, we describe our experimental setup and discuss our findings on the effect of signalling attacks targeting the RRC protocol. Analytical models [34] are a useful way to gain insight into signalling attacks and storms [30], but in this paper we focus on simulation methods which can provide insight into the dynamics of signalling storms.

2. SIMULATION OF UMTS NETWORKS AND SIGNALLING ANOMALIES

The mathematical model we developed previously [30] can provide insight into UE signalling behavior. However we need to simulate large scale mobile networks, and therefore we have designed a simulation tool that is distributed across multiple concurrent processes, and covers the simulated mobile network with combined *packet* and *call* representations of the traffic. Packet communications include SMS, email, web browsing, and instant messaging, while call-level communications include voice, voice-over-IP, and multimedia.

In the control plane, the UE model consists of the session management (SM), GPRS mobility management (GMM) and RRC layers. In the data plane, it contains the application layer, which has circuit-switched and IP applications representing all user activity, the transport layer (TCP and UDP) and a simplified IP layer that is adapted for mobile networks. We have a simplified model of the radio link control (RLC) layer, but we do not explicitly model the MAC and PHY layers; effects of changes in radio conditions are modeled as random variations in the data rate of the radio channels and bearers, which are given in Table 2. Uplink and downlink radio transmissions over a radio bearer (RB) are modeled by two single server, single FIFO queue pairs, one for each direction. The service time at the transmission server is calculated based on the length of the currently transmitted RLC packet and the variable data rates for the RBs. The RLC buffer threshold for triggering a FACH → DCH transition is 1500 bytes[2]. Each UE has one signalling RB and one data RB. We assume that the signalling and data RBs are always in the same RRC state; i.e. if the UE is in the DCH state, then both RBs operate over one or more dedicated channels. In addition to the transmission delays for the RBs, propagation and processing delays are also modeled. We also model the usual communication delays (i.e. transmission, propagation and processing delays) over wired links connecting the different network elements, e.g. between the RNC and the SGSN.

Our RNC model has the RRC, RANAP, NBAP and GTP protocols[3]. The RRC model in the RNC consists of a single signalling server and a single FIFO queue, used to model the processing time δ^r_{xy} for RRC signalling messages. The server handles two classes of signalling messages, where one class consists of signalling messages that effect a state tran-

[2] Our RLC buffer threshold is higher than what is normally used in UMTS networks since we do not model segmentation at the RLC layer.

[3] RANAP: Radio access network application part; NBAP: Node B application part; GTP: GPRS tunneling protocol

Table 2: Radio bearer model configurations. Notice that variable data rates are used in order to reflect the dynamic conditions in the radio interface between the UE and the Node B.

State	Direction	RB configuration
FACH	Uplink	truncated normal distribution, $\mu = 10$ Kbps, $\sigma = 3$ Kbps, min = 4 Kbps, max = 16 Kbps
DCH	Uplink	truncated normal distribution, $\mu = 5$ Mbps, $\sigma = 1.5$ Mbps, min = 1 Mbps, max = 11.5 Mbps
FACH	Downlink	truncated normal distribution, $\mu = 20$ Kbps, $\sigma = 5$ Kbps, min = 8 Kbps, max = 32 Kbps
DCH	Downlink	truncated normal distribution, $\mu = 10$ Mbps, $\sigma = 3$ Mbps, min = 2 Mbps, max = 21.1 Mbps

sition $x \rightarrow y$ (e.g. the RB setup message), and the second class includes all other signalling messages. The service time assigned to the first class reflects the time taken to allocate and deallocate radio resources by the RNC, whereas a default and smaller service time is used for the second class, which is assumed to be 1 ms in our experiments. The service times used in our experiments for signalling messages belonging to the first class are given in Table 1. As the handler of RRC state transitions, this server will be one of the main points of interest in our simulations, and as we discuss in Sec. 3 it will become overloaded as the severity of the signalling attacks increases. It is important to note that the signalling server is not subject to a prior artificial load, and therefore all the load it handles is due to the UEs that are present in the simulation scenario.

3. EXPERIMENTS AND RESULTS

In order to understand the effect of RRC based signalling attacks in UMTS networks, we implemented our simulation model in the OMNeT++ simulation framework [35], and present results from our simulation experiments. The UMTS network topology used in the simulations closely resembles the architecture shown in Fig. 1. In our simulations we have 1000 UEs in an area of 2x2 km^2, which is covered by 7 Node Bs connected to a single RNC. The CN consists of the SGSN and the GGSN, which is connected to 10 Internet hosts acting as web servers. All UEs attach to the mobile network at the start of the simulation, and remain attached. We simulate high user activity in a 2.5 hour period, during which users are actively browsing the web. Our web browsing model is based on industry recommendations [36], and is described below.

3.1 Web Browsing Model of the User

We model interactive web browsing behavior using a self-similar traffic model as shown in Fig. 3. The parameters of the web traffic model are random variables from probability distributions, which are based on web metrics released by Google [37]. The activity period represents the time that the UE is active during a 24 hour period, i.e. the hours during the day that it is generating web traffic. The idle period between two activity periods is the remaining hours within the 24 hours. The first activity period starts after an activation delay $\mathbf{d_a}$, and consists of one or more browsing sessions. The first session within an activity period starts after an initial session delay $\mathbf{d_s}$, and the time between the last and the first main request in one session and the next respectively, is the inter-session interval $\mathbf{i_s}$.

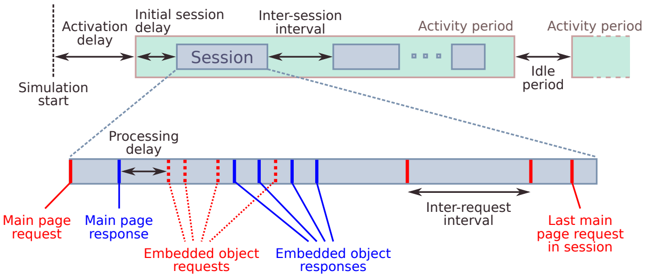


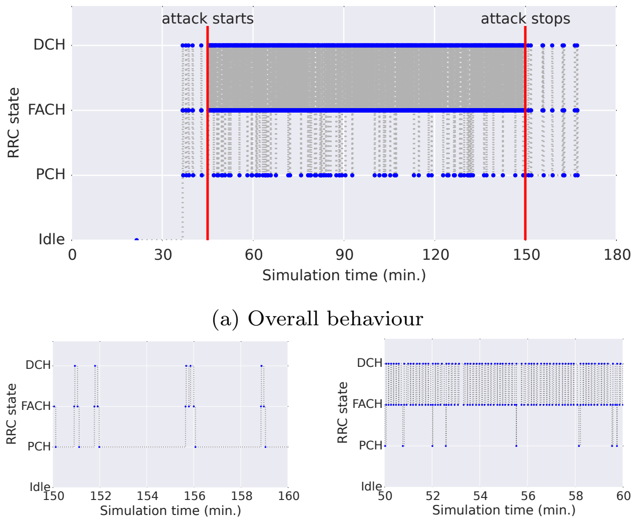
Figure 3: Web traffic model representing interactive user browsing. Note that time is not drawn to scale.

Within a session, the user generates requests for web pages, which are called main page requests, and the first request is scheduled at the start of the session. The request results in a page response from the web server, which is subject to a processing delay $\mathbf{d_{p_c}}$ at the client, representing the time it takes for the web client at the UE to process the received response. A web page contains zero or more embedded objects, and the client generates an embedded object request for each one. We assume that HTTP version 1.1 is used and that each embedded object request is pipelined over a single TCP connection. The length of a request is denoted by $\mathbf{l_r}$. The inter-request interval $\mathbf{i_r}$ is the time between the generation of two consecutive main page requests, and it is independent of the reception of the responses. The session length is controlled by the number of main page requests $\mathbf{n_s}$ in the session.

The web server generates a response for each request it receives after a processing delay $\mathbf{d_{p_s}}$. The length of a main page response is $\mathbf{l_m}$, and it excludes the sizes of any embedded objects and TCP/IP headers. The number of embedded objects per page is $\mathbf{n_e}$ and we model two types of objects: image and text (e.g. CSS documents, scripts). The size of an embedded object is $\mathbf{l_{img}}$ and $\mathbf{l_{txt}}$ for image and text objects, respectively. $\mathbf{R_{img}}$ gives the ratio of image objects to all embedded objects in a page. In the simulations, a client selects a web server uniformly at random for each main page request.

3.2 Attack Model

In our evaluation we mainly consider *DCH attacks*, where the attacker aims to overload the control plane by causing superfluous promotions to the DCH state, and therefore needs to know when a demotion from DCH occurs in the UE. A similar *FACH attack* can be launched where the demotion of interest is from the FACH state, but it is generally more difficult to launch because it requires knowledge of the RRC buffer thresholds and measurement of user traffic volume. The error between the actual transition time and the estimated one is denoted by τ_L and τ_H in the FACH and DCH attack scenarios, respectively. In FACH attacks, the attacker sends a small data packet to a random Internet server in order to cause a promotion to FACH. Higher rate data traffic is generated in DCH attacks in order to cause the buffer threshold to be reached and therefore result in a promotion to DCH.



(a) Overall behaviour

(b) Zoom on normal behaviour

(c) Zoom on attack behaviour

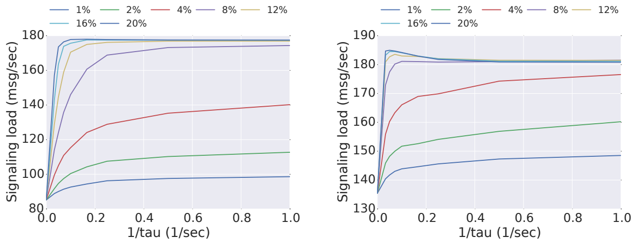
Figure 4: RRC state transitions at a misbehaving UE. DCH attack scenario, PCH enabled, $\tau_H = 2$s

For simulation purposes, our RRC model at the UE informs all registered malicious applications when an RRC state transition occurs. Before launching the next attack, the attacker waits for a period of τ_L or τ_H after a suitable demotion is detected, e.g. from DCH to FACH in the DCH attack case, where τ_L, τ_H are random variables. In our experiments, we assume that τ_L, τ_H are exponentially distributed with mean $= \{0, 1, 2, 4, 6, 10, 14, 20, 30\}s$ to simulate varying degrees of error on behalf of the attacker. For signalling storms, the τ's represent the "synchronization" between the RRC state machine of the UE and the misbehaving application, while the attack scenario represents whether the misbehaving application generates low-rate or high-rate traffic. We present results from the DCH attack scenario only since the FACH attack scenario produces similar behaviour in most cases.

Figure 4 shows the RRC state transitions of a misbehaving UE as captured during a simulation run. As shown in detail in Fig. 4b, when there is no attack, the number of state transitions in a given period are small and due to normal (e.g. web browsing) traffic generated and received by the UE. The UE does not spend long periods in "active" states, i.e. FACH and DCH, quickly transitioning down to PCH when the PCH state is enabled. However, when the UE is misbehaving, either due to malware or a misconfigured app, the profile of the state transitions notably changes as shown in Fig. 4c. The number of state transitions significantly increases, with most transitions occurring between the FACH and DCH states in this DCH attack example. It is this back-and-forth transitioning behaviour that causes excessive signalling load in the mobile network, while the load on the data plane is mostly unaffected.

3.3 Simulation Results

We performed simulation experiments in order to investigate the effect of signalling attacks and storms due to the RRC protocol on the RAN and the CN. We vary the number of compromised or misbehaving UEs from 1% to 20% of all UEs. Both normal and misbehaving UEs generate *normal traffic* based on the web browsing model described above. The misbehaving applications are activated gradually between 20 and 30 minutes from the start of the simulation in order to prevent artifacts such as a huge spike of signalling load due to many malicious applications coming online at the same time. We collect simulation data only from the period when all misbehaving UEs are active. Each data point in the presented results is an average of five simulation runs with different random seeds. The relevant RRC protocol parameters are as given in Table 1.



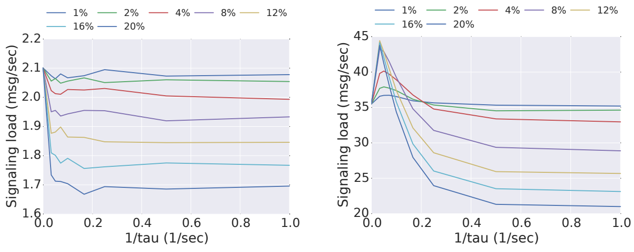
(a) PCH enabled

(b) PCH disabled

Figure 5: Signalling load in the radio access network as a result of DCH attacks, for different number of attackers. The $1/\tau_H = 0$ case corresponds to a "no attack" scenario.

Figure 5 shows the signalling load in the RAN under DCH attacks, with PCH enabled and disabled. As τ_H decreases or the number of attackers increase, the number of signalling messages sent and received by the RNC towards the RAN increases as expected. The rate of increase is dependent on $1/\tau_H$ and higher when the number of attackers is high. We can see that whether the PCH state is enabled does not affect the behaviour of the signalling load in the RAN significantly, but it still decreases the signalling load. An interesting observation is that when PCH is disabled, there is a maximum load when the percentage of attackers is $\geq$ 8% that is attained with a high τ_H. This is worrying since it shows that a maximum signalling load can be induced in the RAN by signalling storms when a sufficient number of UEs misbehave without requiring a high level of synchronization between the misbehaving application and the RRC state machine. Enabling the PCH state addresses this issue. Another useful observation is that given a fixed number of attackers, RRC attacks are *self-limiting*: as signalling load on the RNC increases, this prevents attackers from being able to attack the network at a high rate since they are themselves subject to longer waits for channel allocations. We will re-visit this issue when we discuss congestion at the RNC signalling server.

Figure 6 shows the signalling load in the CN under DCH attacks, with PCH enabled and disabled, and demonstrates the advantage of enabling the optional PCH state. We observe that whether the PCH state is enabled has a significant effect on the signalling load in the CN. This is because most RRC induced signalling with the CN occurs when the UE enters and exits the *idle* state. Enabling the PCH state, which normally has a very long inactivity timer (T_3), prevents the UE from entering the idle state prematurely, significantly decreasing the signalling load in the CN at the cost of slightly more energy consumption at the UE. Therefore, our recommendation would be to enable PCH as a first step in the mitigation of RRC based signalling attacks and storms. Enabling the PCH state also eliminates the problem



(a) PCH enabled

(b) PCH disabled

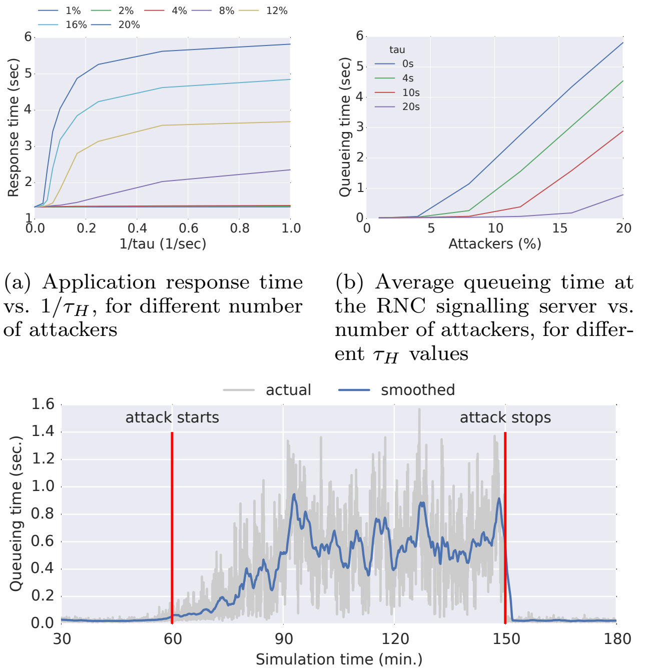
Figure 6: Signalling load in the core network as a result of DCH attacks, for different number of attackers. The $1/\tau_H = 0$ case corresponds to a "no attack" scenario.

of the maximum signalling load observed in Fig. 6b for high values of τ_H, which is due to the interaction between τ_H and the RRC inactivity timers T_1 and T_2. When $\tau_H > T_1 + T_2$, the UE enters the *idle* state as a result of inactivity, and then the misbehaving application causes the UE to go into FACH or DCH in order to send data, resulting in excessive signalling with the CN. The long T_3 timer of the PCH state solves this issue.

Our results so far demonstrate how the mobile network infrastructure is seriously affected by RRC based signalling anomalies. These anomalies also have an appreciable impact on the quality-of-experience (QoE) of the mobile user. Figure 7a shows the application response time at a normal UE, which is defined as the time between when the user requests a web page and when all of the web page is received. The response time is not greatly affected when there are very few misbehaving UEs and when τ_H is high. But delay increases by up to 400% as the severity of the attack increases with increasing number of attackers and $1/\tau_H$. User normally tolerate a wait of 2–10 seconds for a web page to download [38,39], and therefore the observed response times are significant from a QoE view. The affected mobile users are highly likely to attribute the bad QoE to the MNO, so the MNO has one more incentive to detect and mitigate signalling problems in its network.

The main reason for the increase in application response time is the time it takes for the UE to acquire a radio channel in order to send and receive data, which includes, in addition to communication delays between the UE and the RNC, the service and queueing times experienced by the RRC signalling messages effecting the channel acquisition. Figure 7b shows that queueing time at the RRC signalling server component of the RNC greatly increases as the number of attackers increase. We observe that effects of congestion at the server become significant when the percentage of attackers is > 8%, affecting application response time for normal users, and also placing a limit on the impact of signalling attacks on the network since the attackers themselves are subject to longer delays for channel acquisition.

Figure 7c shows the queueing times of RRC signalling messages at the RNC during a single simulation run, and provides a deeper view of the effect of signalling storms in the RAN. We observe that before the storm begins, everything is normal and queueing times are expectedly low given the service times (see Table 1). However, as some of the UEs start misbehaving, queueing times significantly increase. What is noticeable is that the effect of the storm is not immediate and signalling load gradually increases after the storm starts. This is because misbehaving UEs do not start generating attack traffic until a suitable state transition occurs, which happens only after the UE sends or receives normal traffic. However, the effect of the end of the storm is immediate, and queueing times very quickly drop to normal levels as the RNC serves the queued signalling messages.



(a) Application response time vs. $1/\tau_H$, for different number of attackers

(b) Average queueing time at the RNC signalling server vs. number of attackers, for different τ_H values

(c) Queueing time at the RNC signalling server vs. simulation time, $\tau_H = 2$s, 8% attackers

Figure 7: Effect of DCH attacks on application response time and queueing time at the RNC signalling server, PCH enabled

4. DISCUSSION AND FUTURE WORK

In this paper, we have presented a simulation based study of signalling storms in mobile networks, looking at how storms may occur and their effects on the performance of the mobile network and the quality-of-experience (QoE) of the users. Our results have shown that even when a small fraction of the mobile users misbehaves, the signalling components in the mobile network, e.g. the RNC, are overloaded, resulting in at best a degradation in the QoE and at worst a denial-of-service for all mobile users connected to the same network element. Considering that network elements such as the RNC and the SGSN normally handle all users in a small city, the effects of storms can be far-reaching.

Our results indicate that mobile network operators should enable the PCH state in their RRC protocol configuration since it significantly reduces the signalling load on the core network and thus protecting it from the effects of RRC-induced storms. Another recommendation would be to increase the RRC inactivity timers T_1 and T_2 in order to reduce the number of RRC state transitions and thus the signalling load. However, this would affect all users in the mobile network and have a negative impact on the energy consumption of mobile devices, which is a major consideration of operators due to the advent of smartphones. Thus, storm detection and mitigation methods are needed that can distinguish between normal and signalling-heavy users so as to reduce the impact of storms while not needlessly punishing non-misbehaving users.

In future work we hope to use smart traffic management techniques [40,41] to identify the sources and locations of signalling attacks and to reduce their effect on other users of the mobile network. Using representative traffic features of signalling attacks and storms, we will use detection methods that will reduce false positives in order not to penalize data-heavy users, and introduce additional delays in control state transitions in order to optimally mitigate against attacks.

Acknowledgments

This work is part of the EU FP7 project NEMESYS (Enhanced Network Security for Seamless Service Provisioning in the Smart Mobile Ecosystem), under grant agreement no. 317888 within the FP7-ICT-2011.1.4 Trustworthy ICT domain.

5. REFERENCES

[1] (2013, Feb.) Mobile device market to reach 2.6 billion units by 2016. Canalys. [Online]. Available: http://www.canalys.com/newsroom/mobile-device-market-reach-26-billion-units-2016

[2] (2013, Feb.) Cisco visual networking index: Global mobile data traffic forecast update, 2012–2017. White Paper. [Online]. Available: http://www.cisco.com/en/US/solutions/collateral/ns341/ns525/ns537/ns705/ns827/white_paper_c11-520862.pdf

[3] A. P. Felt, M. Finifter, E. Chin, S. Hanna, and D. Wagner, "A survey of mobile malware in the wild," in *Proc. 1st ACM W'shop on Security and Privacy in Smartphones and Mobile Devices (SPSM'11)*, 2011, pp. 3–14.

[4] M. Chandramohan and H. B. K. Tan, "Detection of mobile malware in the wild," *IEEE Computer*, vol. 45, no. 9, pp. 65–71, Sep. 2012.

[5] Y. Zhou and X. Jiang, "Dissecting Android malware: Characterization and evolution," in *Proc. 2012 IEEE Symposium on Security and Privacy (SP'12)*, May 2012, pp. 95–109.

[6] E. Gelenbe, G. Gorbil, D. Tzovaras, S. Liebergeld, D. Garcia, M. Baltatu, and G. L. Lyberopoulos, "NEMESYS: Enhanced network security for seamless service provisioning in the smart mobile ecosystem," in *Information Sciences and Systems 2013*, ser. LNEE, vol. 264. Springer, Oct. 2013, pp. 369–378.

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

[8] 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 on Wireless Security (WiSe'06)*, Sep. 2006, pp. 75–84.

[9] P. P. Lee, T. Bu, and T. Woo, "On the detection of signaling DoS attacks on 3G wireless networks," in *Proc. 26th IEEE Int. Conf. on Computer Communications (INFOCOM'07)*, May 2007, pp. 1289–1297.

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

[11] O. H. Abdelrahman, E. Gelenbe, G. Gorbil, and B. Oklander, "Mobile network anomaly detection and mitigation: The NEMESYS approach," in *Information Sciences and Systems 2013*, ser. LNEE. Springer, Oct. 2013, vol. 264, pp. 429–438.

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

[13] C. Mulliner and J.-P. Seifert, "Rise of the iBots: Owning a telco network," in *Proc. 5th Int. Conf. on Malicious and Unwanted Software (MALWARE'10)*, Oct. 2010, pp. 71–80.

[14] D. Maslennikov and Y. Namestnikov. (2012, Dec.) Kaspersky security bulletin 2012: The overall statistics for 2012. Kaspersky Lab. [Online]. Available: http://www.securelist.com/en/analysis/204792255/Kaspersky_Security_Bulletin_2012_The_overall_statistics_for_2012

[15] E. Gelenbe, "Sensible decisions based on qos," *Computational Management Science*, vol. 1, no. 1, pp. 1–14, dec 2003.

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

[17] P. H. Perala, A. Barbuzzi, G. Boggia, and K. Pentikousis, "Theory and practice of RRC state transitions in UMTS networks," in *Proc. 2009 IEEE Global Communications Conf. W'shops (Globecom'09 Wshops)*, Nov. 2009, pp. 1–6.

[18] F. Qian, Z. Wang, A. Gerber, Z. M. Mao, S. Sen, and O. Spatscheck, "Characterizing radio resource allocation for 3G networks," in *Proc. 10th ACM SIGCOMM Conf. on Internet Measurement (IMC'10)*, Nov. 2010, pp. 137–150.

[19] Z. Qian, Z. Wang, Q. Xu, Z. M. Mao, M. Zhang, and Y.-M. Wang, "You can run, but you can't hide: Exposing network location for targeted DoS attacks in cellular networks," in *Proc. 19th Annual Network and Distributed System Security Symposium (NDSS'12))*, Feb. 2012.

[20] Z. Wang, Z. Qian, Q. Xu, Z. Mao, and M. Zhang, "An untold story of middleboxes in cellular networks," *ACM SIGCOMM Computer Communication Review - SIGCOMM'11*, vol. 41, no. 4, pp. 374–385, Aug. 2011.

[21] E. Gelenbe and F.-J. Wu, "Large scale simulation for human evacuation and rescue," *Computers and Mathematics with Applications*, vol. 64, no. 12, pp. 3869–3880, Dec. 2012.

[22] E. Gelenbe, G. Gorbil, and F.-J. Wu, "Emergency cyber-physical-human systems," in *Proc. 21st Int. Conf. on Computer Communications and Networks (ICCCN'12)*. IEEE Computer Society, August 2012, pp. 1–7.

[23] A. Filippoupolitis, G. Gorbil, and E. Gelenbe, "Spatial computers for emergency support," *The Computer Journal*, vol. 56, no. 12, pp. 1399–1416, Dec. 2013.

[24] G. Gorbil and E. Gelenbe, "Resilience and security of opportunistic communications for emergency evacuation," in *Proc. 7th ACM W'shop on Performance Monitoring and Measurement of Heterogeneous Wireless and Wired Networks (PM2HW2N'12).* ACM, October 2012, pp. 115–124.

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

[26] S. Corner. (2011, Jun.) Angry Birds + Android + ads = network overload. IT Wire. [Online]. Available: http://www.itwire.com/business-it-news/networking/47823

[27] (2012, Feb.) Smarter apps for smarter phones! GSMA. [Online]. Available: http://www.gsma.com/technicalprojects/wp-content/uploads/2012/04/gsmasmarterappsforsmarterphones0112v.0.14.pdf

[28] S. Jianto, "Analyzing the network friendliness of mobile applications," Huawei, Tech. Rep., Jul. 2012. [Online]. Available: www.huawei.com/ilink/en/download/HW_146595

[29] "3GPP TS 25.331: Universal mobile telecommunications system (UMTS) radio resource control (RRC) protocol specification," 3GPP, technical specification. [Online]. Available: http://www.3gpp.org/DynaReport/25331.htm

[30] O. H. Abdelrahman and E. Gelenbe, "Signalling storms in 3G mobile networks," in *Proc. 2014 IEEE Int. Conf. on Communications (ICC'14)*, Sydney, Australia, Jun. 2014, accepted for publication.

[31] 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)*, Apr. 2007, pp. 964–968.

[32] 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.

[33] 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.

[34] E. Gelenbe, "Probabilistic models of computer systems part II: Diffusion approximations: waiting times and batch arrivals," *Acta Informatica*, vol. 12, no. 4, pp. 285–303, 1979.

[35] A. Varga and R. Hornig, "An overview of the OMNeT++ simulation environment," in *Proc. 1st Int. Conf. on Simulation Tools and Techniques for Communications, Networks and Systems W'shops (Simutools'08)*, Mar. 2008, pp. 60:1–60:10.

[36] "cdma2000 evaluation methodology - revision A," Technical document, 3GPP2, May 2009, 3GPP2 C.R1002-A, version 1.0. [Online]. Available: http://www.3gpp2.org/public_html/specs/C.R1002-A_v1.0_Evaluation_Methodology.pdf

[37] S. Ramachandran. (2010, May) Web metrics: Size and number of resources. Google. [Online]. Available: https://developers.google.com/speed/articles/web-metrics

[38] J. Nielsen, *Usability Engineering.* Cambridge, MA, USA: Morgan Kaufmann, Sep. 1993, ch. 5.

[39] F. F.-H. Nah, "A study on tolerable waiting time: How long are Web users willing to wait?" *Behaviour & Information Technology*, vol. 23, no. 3, pp. 153–163, 2004.

[40] E. Gelenbe, N. Schmajuk, J. Staddon, and J. Reif, "Autonomous search by robots and animals: A survey," *Robotics and Autonomous Systems*, vol. 22, no. 1, pp. 23–34, 1997.

[41] E. Gelenbe and Z. Kazhmaganbetova, "Cognitive packet network for bilateral asymmetric connections," *IEEE Trans. Industrial Informatics, accepted for publication*, 2014.