

Mitigating Attacks on Open Functionality in SMS-Capable Cellular Networks

Patrick Traynor, William Enck, Patrick McDaniel, and Thomas La Porta
 Systems and Internet Infrastructure Security Laboratory
 Networking and Security Research Center
 Department of Computer Science and Engineering
 The Pennsylvania State University
 University Park, PA 16802
 {traynor, enck, mcdaniel, tlp}@cse.psu.edu

Abstract—¹ The transformation of telecommunications networks from homogeneous closed systems providing only voice services to Internet-connected open networks that provide voice and data services presents significant security challenges. For example, recent research illustrated that a carefully crafted DoS attack via text messaging could incapacitate all voice communications in a metropolitan area with little more than a cable modem. This attack highlights a growing threat to these systems; namely, cellular networks are increasingly exposed to adversaries both in and outside the network. In this paper, we use a combination of modeling and simulation to demonstrate the feasibility of targeted text messaging attacks. Under realistic network conditions, we show that adversaries can achieve blocking rates of more than 70% with only limited resources. We then develop and characterize five techniques from within two broad classes of countermeasures - queue management and resource provisioning. Our analysis demonstrates that these techniques can eliminate or extensively mitigate even the most intense targeted text messaging attacks. We conclude by considering the tradeoffs inherent to the application of these techniques in current and next generation telecommunications networks.

Keywords: telecommunications, sms, denial-of-service, open-functionality

I. INTRODUCTION

In addition to traditional voice communications, cellular systems offer a wide variety of data and text/short messaging services (SMS). Cellular providers have introduced SMS gateways between the phone networks and the Internet to increase the reach (and volume) of text messaging. These gateways are partially responsible for the soaring usage of text messaging. Some five billion text messages are sent each month in the United States alone [1]. Indeed, for significant numbers of users, text messaging has become the primary means of communication [2].

Such interconnectivity is not without serious security risks. The cellular infrastructure was designed to operate as a homogeneous and highly controlled system. Interconnectivity with the Internet invalidates many of the assumptions upon which the phone networks were designed. These failing assumptions lead to critical vulnerabilities. Enck et al. [3] showed that, under a simple model, small but carefully crafted volumes

of text messages could be used to incapacitate all cellular communications in large metropolitan areas. Particularly in times of crisis, the cost of the maliciously driven loss of cellular communications could be immeasurable.

In this paper, we consider both the reality and mitigation of previously postulated attacks on text messaging. Using analytical modeling and highly detailed simulation of the GSM air interface, we show that the simple model described in the original study underestimated the traffic volumes necessary to to effect a high-impact attack by approximately a factor of three. However, while the specific estimates made by Enck et al. were not perfect, we found their qualitative arguments of feasibility to hold under a range of realistic traffic models. In short, our findings demonstrate that cellular networks are in fact quite vulnerable to SMS-based attacks mounted by adversaries with even limited resources. Specifically, an attack capable of preventing the large majority of voice communications in a metropolitan area is indeed possible with the bandwidth available to a single cable modem.

In the presence of this reality, we have developed five techniques from within two broad classes of countermeasures, *queue management* and *resource provisioning*, to combat these attacks. Our goal is to maintain the high availability of voice telephony required by regulatory bodies while attempting to provide maximal throughput for high-value text messaging. We apply well-known queueing techniques including variants of Weighted Fair Queueing (WFQ) and Weighted Random Early Detection (WRED), which are well tested for addressing traffic overload in the Internet. These schemes attempt to provide differentiated service to voice and data, and hence alleviate resource contention.

The second class of solutions reapportion the wireless medium using the novel Strict Resource Provisioning (SRP), Dynamic Resource Provisioning (DRP) and Direct Channel Allocation (DCA) algorithms. Our modeling and simulation analyses of the countermeasures demonstrates their utility: the effect of the solutions ranged from partial attack mitigation for both flows to the prevention of attack-related voice blocking and the successful delivery of high priority text messages. A further exploration of the deployment of these solutions highlights a number of security, performance, and complexity tradeoffs. We discuss these tradeoffs throughout and make a

¹This is a preprint of the final TON article. A publication date has yet to be set.

number of recommendations to the community.

In this work, we make the following contributions:

- **Network/Attack Characterization:** We create a realistic characterization of system behavior under targeted SMS attacks. Such characterizations dually ascertain their existence and develop a profile of the effect of an attack under varying traffic intensity and arrival models.
- **Current Countermeasure Analysis:** We briefly consider popularly advertised solutions to targeted text messaging attacks. Principally, we find that the currently deployed “edge solutions” are largely ineffective against all but the most naïve attack.
- **Countermeasure Development and Evaluation:** We develop and characterize a number of countermeasures adapted from well-established queueing techniques and novel channel allocation strategies. Our analysis demonstrates that these attacks can be effectively mitigated by altering the traffic handling disciplines at the air interface. Hence, countering these attacks need not require a substantive change to internal structure or operation of cellular networks, but can be handled entirely by software changes at the base station.

The remainder of this paper is organized as follows: Section II discusses related work; Section III provides an overview of cellular signaling networks and characterizes targeted SMS attacks; Section IV offers a number of mitigation strategies and models their ability to mitigate these attacks; Section V details the attack and mitigation simulations; and Section VI offers concluding remarks and future work.

II. RELATED WORK

Physical disconnection from external networks has long been one of the most effective means of providing security for telecommunication systems. Accordingly, security in these networks has traditionally centered around the prevention of fraudulent access and billing. The changing needs of users, however, have forced the gradual erosion of well defined network borders. Whether due to new access patterns or the advent of new services (e.g., data networking via the Internet), systems that once relied upon isolation as a major portion of their defenses are no longer able to do so. Because fundamental assumptions about the underlying architecture of the critical communications infrastructure have changed, security measures addressing new classes of threats resulting from the interconnection of networks are essential. Similar observations and concerns have been expressed in the National Strategy to Secure Cyberspace [4].

Telecommunications networks are not the only systems to suffer from vulnerabilities related to expanded connectivity. Systems including Bank of America’s ATMs and 911 emergency services for Bellevue, Washington were both made inaccessible by the Slammer worm [5]. Although neither system was the target of this attack, simply being connected to the Internet made them experience significant collateral damage. Systems less directly connected to the Internet have also been subject to attack. Byers, et al. [6] demonstrated

one such attack using simple automated scripts and webforms. Immense volumes of junk postal mail could then be used to launch *Denial of Service* (DoS) attacks on individuals.

The typical targets of DoS attacks, however, are more traditional online resources. In 2000, for example, users were unable to reach Amazon, eBay and Yahoo! as their servers were bombarded with over a gigabit per second of traffic [7]. Since that time, sites ranging from software vendors [8] and news services [9] to online casinos [10] have all fallen victim to such attacks. While significant research has been dedicated to categorizing [11] and mitigating [12], [13], [14], [15], [16] these attacks, no solutions have seen widespread implementation. Because of the various transformations of data transiting between the Internet and telecommunications networks, the direct application of the above techniques would be ineffective.

Whether accidental or the result of malicious behavior, denial of service incidents have been studied and documented in telecommunications networks. The National Communications System published a study on the effects of text messages during emergency situations. Given realistic scenarios for usage, this technical bulletin argued that SMS resources needed to be increased 100-fold in order to operate under such conditions [17]. Operators have also reported problems with connectivity during holidays due to increased volumes of SMS traffic [18]. Enck, et al. [3] demonstrated that an adversary would be able to cause the same congestion in targeted metropolitan areas by injecting a relatively small amount of traffic. While a number of solutions were proposed in that work, none have yet been measured and compared.

III. SYSTEM/ATTACK CHARACTERIZATION

A. Message Delivery Overview

In the following subsection, we provide a high-level, simplified tutorial on text message delivery in cellular networks. Figures 1 and 2 illustrate this process.

1) *Message Insertion:* Messages may be submitted into the system from cell phones operating within the system or from external sources. We focus on delivery from external sources, but the message flow for all types of messages is similar.

An Internet-originated SMS message can be generated by any one of a number of *External Short Messaging Entities* (ESMEs). ESMEs include devices and interfaces ranging from email and web-based messaging portals to service provider websites and voicemail, services and can be attached to telecommunications networks either by dedicated connection or the Internet. When a message is injected into the network, it is delivered to the *Short Messaging Service Center* (SMSC). These servers are responsible for the execution of a “store-and-forward” protocol that eventually delivers text messages to their intended destination.

The contents and destination information from the message are examined by the SMSC and are then copied into a properly formatted packet. At this point, messages originating in the Internet and those created in the network itself become indistinguishable. Formatted text messages are then placed in an egress queue in the SMSC and await service.

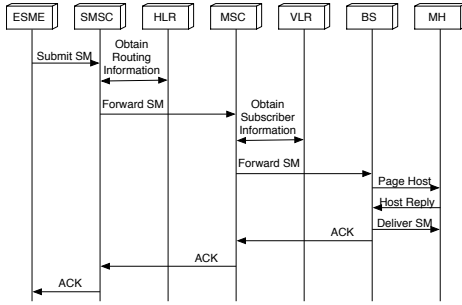


Fig. 1. A high level description of SMS delivery in an SS7 network.

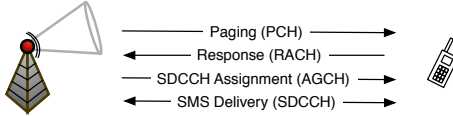


Fig. 2. An overview of SMS message delivery on the wireless or air interface. Incoming voice calls would follow a similar procedure except that they would receive a TCH after using the SDCCH.

2) *Message Routing*: Before an SMSC can forward a text message to a targeted mobile device, it must first determine the location of that device. To accomplish this, the SMSC queries a database known as the *Home Location Register* (HLR). The HLR is responsible for storing subscriber data including availability, billing information, available services and current location. With the help of other elements in the network, the HLR determines the routing information for the targeted device. If the desired phone is not available, the SMSC stores the message until a later time for subsequent retransmission. Otherwise, the SMSC receives the address of the *Mobile Switching Center* (MSC) currently providing service to the target device. The MSC delivers the text message over the wireless interface through its attached *Base Stations* (BS).

3) *Wireless Delivery*: An area of coverage in a wireless network is called a cell. Each cell is typically partitioned into multiple (usually three) sectors. We characterize the system on a per sector basis throughout the paper.

The air interface, or radio portion of the network, is traditionally divided into two classes of logical channels — the *Control Channels* (CCHs) and *Traffic Channels* (TCH). TCHs carry voice traffic after call setup has occurred. CCHs, which transport information about the network and assist in call setup/SMS delivery, are subclassified further. In order to alert a targeted device that a call or text message is available, a message is broadcast on the *Paging Channel* (PCH). Note that multiple base stations broadcast this page in an attempt to quickly determine the sector in which the targeted recipient is located. Upon hearing its temporary identifier on the PCH, available devices inform the network of their readiness to accept incoming communications using the slotted ALOHA-based *Random Access Channel* (RACH) uplink. A device is then assigned a *Standalone Dedicated Control Channel* (SDCCH) by listening to the *Access Grant Channel* (AGCH). If a text message is available, the base station authenticates the device, enables encryption, and then delivers the contents of the message over the assigned SDCCH. If instead a call is incoming for the device, the SDCCH is used to authenticate

TABLE I
COMMONLY USED VARIABLES

λ_{call}	Arrival rate of voice calls
λ_{SMS}	Arrival rate of text messages
$\mu_{SDCCH,call}^{-1}$	Service rate of voice calls at SDCCH
$\mu_{TCH,call}^{-1}$	Service rate of voice calls at TCH
$\mu_{SDCCH,SMS}^{-1}$	Service rate of text messages at SDCCH
ρ_{call}	Call traffic intensity
ρ_{SMS}	SMS traffic intensity

TABLE II
SYSTEM AND ATTACK PARAMETERS

μ_{TCH}^{-1}	120 sec [19]
$\mu_{SDCCH,call}^{-1}$	1.5 sec [19]
$\mu_{SDCCH,SMS}^{-1}$	4 sec [17]
λ_{call}	50,000 calls/city/hr .2525 calls/sector/sec
$\lambda_{SMS,attack}$	495 msgs/city/sec 9 msgs/sector/sec
$\lambda_{SMS,regular}$	138.6K/city/hr 0.7 msgs/sector/sec

the device and negotiate a TCH for voice communications.

B. System Vulnerability

All large scale attacks, whether targeting the digital or physical domain, evolve in the following phases: *recognition* (identification of a vulnerability), *reconnaissance* (characterization of the conditions necessary to attack the vulnerability), *exploit* (attacking the vulnerability) and *recovery* (cleanup and forensics). We approach targeted SMS attacks in the same fashion. Enck, et al. [3] provide a methodology for executing such an attack; we summarize it here.

The vulnerability in GSM cellular networks that allows for targeted text message attacks to occur is the result of bandwidth allocation on the air interface. Under normal operating conditions, the small ratio of bandwidth allocated to control versus traffic data is sufficient to deliver all messages with a low probability of blocking. However, because text messages use the same control channels as voice calls for delivery (SDCCHs), contention for resources occurs when SMS traffic is elevated. Given a sufficient number of SMS messages, each of which require on average four seconds for delivery [17], arriving voice calls will be blocked for lack of available resources.

Sending text messages to every possible phone number is not an effective means of attacking a network. The haphazard submission of messages is in fact more likely to overwhelm gateways between the Internet and telecommunications networks than to disrupt cellular service. An adversary must efficiently blanket only the targeted area with messages so as to reduce the probability of less effective collateral damage. The information to achieve such a goal, however, is readily available. Using tools including NPA-NXX Area Code Databases, Internet search engines and even feedback from service provider websites, an attacker can easily construct a “hit-list” of potential targets. Given this information, an adversary can then begin exploiting the bandwidth vulnerability.

The exploit itself involves saturating sectors to their SDCCH capacity for some period of time. In so doing, the majority of attempts to establish voice calls are blocked. For all of

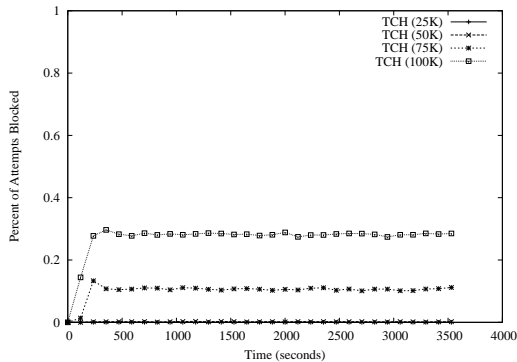


Fig. 3. Blocking characteristics of a network under a variety of traffic intensities.

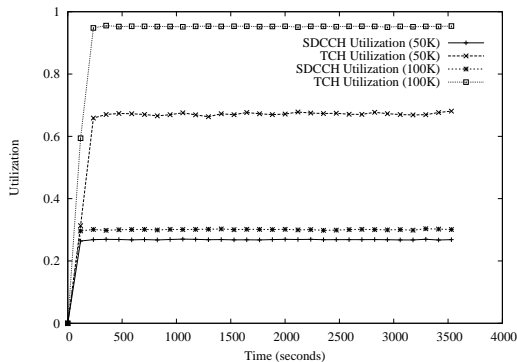


Fig. 4. Channel utilization characteristics of a network under a variety of traffic intensities.

Manhattan, which would typically be provisioned with 12 SDCCHs per sector, a perfectly executed attack would require the injection of only 165 messages per second, or approximately 3 messages/sector/second.

C. Network Characterization

We begin by developing characterizations of the GSM air interface under a range of standard operating conditions. To achieve these ends, we have developed a detailed GSM simulator. The design considerations and verification of its accuracy are discussed in our previous work [20]. A cellular deployment similar to that found in Manhattan [17] is used as our baseline scenario. Each of the 55 sectors in the city has 12 SDCCHs². We assume both call requests and text messages arrive with a Poisson distribution and that TCH and SDCCH holding times are exponentially distributed around the appropriate means (see Tables I and II) unless explicitly stated otherwise. Such values are well within standard operating conditions [21], [22], [1].

Figure 3 illustrates the blocking rates for traffic channels under four different voice traffic loads. Most relevant to the current discussion is the nonexistence of call blocking. The absence of such blocking reinforces the robustness of the design of GSM as a voice communication system. Specifically, the only points of congestion in the system are the traffic

²In reality, only the highest capacity sectors would be so over-provisioned [17], making this a conservative estimate for every sector in a city.

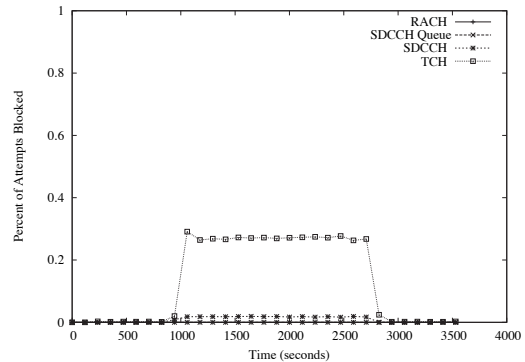


Fig. 5. Blocking characteristics of a network under emergency conditions (100K calls/hr, 276K msgs/hr).

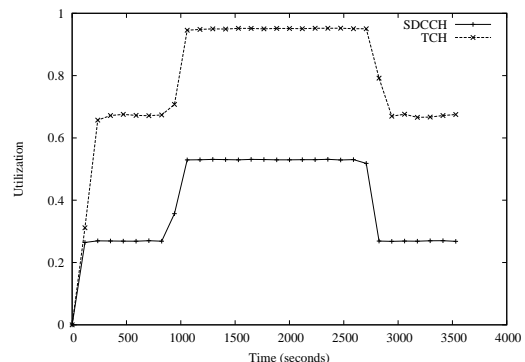


Fig. 6. Channel utilization characteristics of a network under emergency conditions (100K calls/hr, 276K msgs/hr).

channels themselves. Figure 4 further supports the blocking data by illustrating very low SDCCH utilization rates for offered loads of both 50 and 100K *calls/hour*.

Elevated loads may represent significant public gatherings (e.g., concerts, celebrations), holiday spikes or large-scale emergencies. Blocking on other channels begins to become observable only under such extreme circumstances. Figure 5 highlights an emergency situation in which the call and SMS rate spikes from 50K *calls/hour* to 100K *calls/hour* and 138K *msgs/hour* to 276K *msgs/hour*, respectively. Figure 6, which shows the channel utilization data for the “emergency” scenario, reinforces that it is only under extreme duress that other channels in the system begin to saturate.

D. Attack Characterization

In order to judge the efficacy of any countermeasure against targeted SMS attacks, it is necessary to fully characterize such an event. We seek to understand the observed conditions and the subtle interplay of network components given a wide range of inputs. We use the simulator described in the previous subsection and the parameters in Table II to understand such attacks.

To isolate the impact of blocking caused by SDCCH congestion, we do not include SDCCH queues; we examine the impact of such queues in Sections IV and V. If a call request or text message arrives when all SDCCHs are occupied, the request is blocked.

A sector is observed for a total of 60 minutes, in which the middle 30 minutes are exposed to a targeted SMS at-

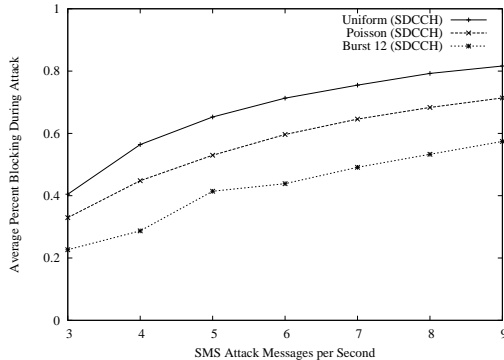


Fig. 7. The blocking probability for traffic exhibiting uniform, Poisson and bursty arrival patterns over varying attack strengths.

tack. The SMS attack intensity is varied between 4 and 13 times the normal SMS load, i.e., $\lambda_{SMS} = 165 \text{ msgs/sec}$ (3 messages/second/sector) to $\lambda_{SMS} = 495 \text{ msgs/sec}$ (9 messages/second/sector)³. All results are the average of 1000 runs, each using randomly generated traffic patterns consistent with the above parameters.

Because delay variability is likely throughout the network, and because SDCCH holding times will not be deterministic due to varying processing times and errors on the wireless links, the perfect attack presented in our previous work [3] would be difficult to achieve in real networks. Accordingly, we investigate a number of flow arrival characteristics while considering exponentially distributed SDCCH holding times. Figure 7 illustrates the effectiveness of attacks when messages arrive with a Poisson, bursty (12 messages delivered back-to-back), or uniform distribution. Notice that, due to the addition of variability, bursty attacks are the least successful of the three. This is because it is unlikely that 12 text messages arriving back-to-back will all find unoccupied SDCCHs. Thus, blocking occurs on the attack messages, and legitimate traffic that arrives between bursts has a higher probability of finding an available SDCCH. The most effective attack is when messages arrive uniformly spaced; however, due to variable network delay, such an attack would also be difficult to realize.

Our remaining experiments therefore assume a Poisson distribution for the arrival of text messages. We use an attack intensity of 495 msgs/sec , which is equal to 9 messages/second/sector and yields a blocking probability of 71%. For this case we show the SDCCH and TCH utilization in Figure 8. This figure shows the effectiveness of the attack: during the attack, the SDCCH utilization is near 1.0, and the TCH utilization drops from close to 70% down to approximately 20%. This shows that although TCHs are available for voice calls, they cannot be allocated due to SDCCH congestion. Our experiments suggest that, at this rate, no other bottlenecks in the system exist, including other control channels or the SS7 signaling links.

³Because DoS attacks on the Internet frequently exhibit more than an entire year's volume of traffic [7], such an increase is relatively insignificant.

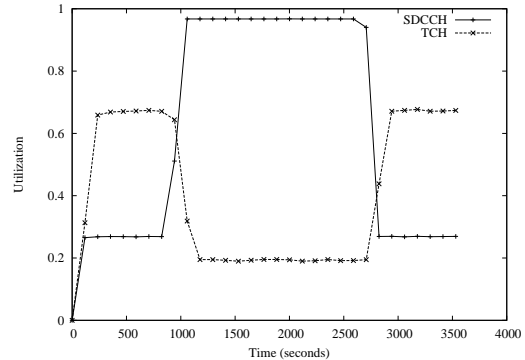


Fig. 8. The utilization of SDCCHs and TCHs for an attack exhibiting a Poisson interarrival at a rate of 495 messages/second.

IV. MITIGATION TECHNIQUES

Voice communications have traditionally received priority in telecommunications systems. Because voice has been the dominant means by which people interact via these networks, providers allow for the degradation of other services in order to achieve high availability for the voice services. There are, however, an increasing set of scenarios in which the priority of services begins to change.

On September 11th, 2001, service providers experienced significant surges in usage. Verizon Wireless reported the number of calls made increased by more than 100% above average levels. Cingular Wireless experienced an increase of over 1000% for calls bound for the greater Washington D.C area [17]. In spite of the increased call volume, SMS messages were still received in even the most inundated areas because the control channels used for their delivery remained uncongested. In both emergency and day-to-day situations, the utility of text messaging has increased to the same level as voice communications for significant portions of the population [2].

A simple analysis shows that repurposing all resources used in voice telephony for SMS delivery would greatly increase overall network throughput. Specifically, more users would be able to communicate concurrently under such a scheme. However, regulations constraining the availability of voice services during times of emergency prevent such an approach from being implemented. Given requirements for reliable voice and an ever-increasing demand for and reliance upon SMS, mitigation strategies must not only maintain the availability of voice services, but also maximize the throughput of legitimate text messages.

There are three traditional approaches to combating congestion. The first is rate limitation of the traffic source — in this case the interfaces on which messages are submitted. The network can also be protected by shedding traffic or using scheduling mechanisms. Finally, resources may be reallocated to alleviate the network bottleneck. We examine these solutions below. While other solutions including processor overload controls [23], [24] and more distributed mechanisms [25] are potentially applicable, we examine the above solutions because of their generality and ease of deployment.

A. Current Solutions

Cellular providers have introduced a number of mitigation solutions into phone networks to combat the SMS-based DoS attacks. These solutions focus on *rate limiting* the source of the messages and are ineffective against all but the least sophisticated adversary. To illustrate, the primary countermeasure discovered by the authors of the original study was a per-source volume restriction at the SMS gateway [3]. Such restrictions would, for example, allow only 50 messages from a single IP address. The ability to spoof IP addresses and the existence of zombie networks render this solution impotent. Another popular deployed solution filters SMS traffic based on the textual content. Similar to SPAM filtering, this approach is effective in eliminating undesirable traffic only if the content is predictable. However, an adversary can bypass this countermeasure by generating legitimate looking SMS traffic from randomly generated simple texts, e.g., “Remember to buy milk on your way home from the office. -Alice”

Note that these and the overwhelming majority of other solutions deployed in response to the SMS vulnerability can be classified as *edge solutions*. Ineffective by construction because of their lack of context, such solutions try to regulate the traffic flowing from the Internet into the provider network at its edge. Limiting the total volume of traffic coming across all interfaces results simply in reduced income under normal operating conditions. For example, a total of 1,000 email-generated text messages per second distributed across a nation cause no ill effects to the network and generates significant revenue; however, Section III shows that such a volume targeted to one region is more than sufficient to paralyze the network.

Rate limitation is largely unattractive even within the core network. The distributed nature of Short Messaging Service Centers (SMSCs), through which all text messages flow, makes it difficult to coordinate real-time filtering in response to targeted attacks. Moreover, because provider networks cover huge geographic areas and consist of hundreds of thousands of network elements, any compromised element can be a conduit for malicious traffic. If left unregulated, the connections between provider networks can also be exploited to inject SMS traffic.

Therefore, for the purposes of this discussion, we assume that an adversary is able to successfully submit a large number of text messages into a cellular network. The defenses below are dedicated to protecting the resource that is being exploited in the SMS attack – the bandwidth constrained SDCCHs. Note that the Internet faces a similar conundrum: once dominant perimeter defenses are failing in the face of dissolving network borders, e.g., as caused by wireless connectivity and larger and more geographically distributed networks [26]. As is true in the Internet, we must look to techniques providing “defense in depth” to protect telecommunications networks.

In the following sections we discuss mitigation techniques based on *queue management* and *resource provisioning*. For each solution we provide some basic analysis to provide insight; the motivation for parameter selection is covered in more detail in Section V.

B. Cellular Data

A number of providers have begun using cellular data services such as GPRS to assist in the delivery of text messages. While such services provide higher bandwidth, they do not necessarily mitigate targeted text messaging attacks. While the coverage of cellular data networks is expanding, service is not universal. Moreover, because data channels can be reclaimed for use by voice calls during periods of elevated traffic, SMS is often delivered over SDCCHs in areas with access to cellular data services. Finally, as discussed in a follow on to this work [27], such services are subject to even lower bandwidth attacks. Accordingly, targeted text messaging attacks have not been made obsolete by cellular data networks and must therefore be addressed.

C. Queue Management Techniques

1) *Weighted Fair Queueing*: Because we cannot rely on rate limitation at the source of messages, we now explore network-based solutions. Fair Queueing [28] is a scheduling algorithm that separates flows into individual queues and then apportions bandwidth equally between them. Designed to emulate bit-wise interleaving, Fair Queueing services queues in a round-robin fashion. Packets are transmitted when their calculated interleaved finishing time is the shortest. Building priority into such a system is a simple task of assigning weights to flows. Known as *Weighted Fair Queueing* (WFQ) [29], this technique can be used to give incoming voice calls priority over SMS.

We provide a simplified analysis to characterize the performance of WFQ in this scenario. We apply WFQ to the service queues of the SDCCH. We create two waiting queues, one for voice requests and one for SMS requests, respectively. The size of the call queue is 6 and the size of the SMS queue is 12 (discussed in Section V). To determine the relative blocking probability and utilization of the voice and SMS flows, we begin by assuming the conditions set forth in Tables I and II.

WFQ can be approximated as a general processor sharing system (GPS) [30]. The average service rate of such systems is the weighted average of the service rates of all classes of service requests. In our case we have two types of requests: voice calls with $\lambda_{call} = 0.2525 \text{ calls/sector/sec}$ and an average service time on the SDCCH of $\mu_{call}^{-1} = 1.5$ seconds, and SMS requests with $\lambda_{SMS} = 9.7 \text{ msgs/sector/sec}$ (attack traffic + regular traffic) and $\mu_{SMS}^{-1} = 4$ seconds. Therefore, for our system, $\mu^{-1} = 3.94$ seconds.

Although our system has multiple servers (SDCCHs), and is thus an M/M/n system, because it is operating at high loads during an attack, it may be approximated by an M/M/1 system with its $\mu = n\mu'$, where μ' is the service rate calculated above. Using these values, and accounting for the weighting of 2:1 for servicing call requests, the call traffic intensity $\lambda_{call}/\mu_{call} = \rho_{call} = 0.04$, and the expected call queue occupancy is about 1%. Because the ρ_{SMS} is much greater than 1, its SMS queue occupancy is approximately 100%. When combined, the total queue occupancy is approximately 67%.

These numbers indicate that the WFQ-based approach would sufficiently protect voice calls from targeted SMS attacks. Section V offers additional insight through simulation.

2) *Weighted Random Early Detection*: Active queue management has received a great deal of attention as a congestion avoidance mechanism in the Internet. *Random Early Detection* (RED) [31], [32], one of the better known techniques from this field, is a particularly effective means of coping with potentially damaging quantities of text messages. While traditionally used to address TCP congestion, RED helps to prevent queue lockout and RED drops packets arriving to a queue with a probability that is a function of the weighted queue occupancy average, Q_{avg} . Packets arriving to a queue capacity below a threshold, t_{min} , are never dropped. Packets arriving to a queue capacity above some value t_{max} are always dropped. Between t_{min} and t_{max} , packets are dropped with a linearly increasing probability up to $P_{drop,max}$. This probability, P_{drop} , is calculated as follows⁴:

$$P_{drop} = P_{drop,max} \cdot (Q_{avg} - t_{min}) / (t_{max} - t_{min}) \quad (1)$$

The advantages to this approach are twofold: first, lockout becomes more difficult as packets are purposefully dropped with greater frequency; secondly, because the capacity of busy queues stays closer to a moving average and not capacity, space typically exists to accommodate sudden bursts of traffic. However, one of the chief difficulties with traditional RED is that it eliminates the ability of a provider to offer quality of service (QoS) guarantees because all traffic entering a queue is dropped with equal probability. *Weighted Random Early Detection* (WRED) solves this problem by basing the probability a given incoming message is dropped on an attribute such as its contents, source or destination. Arriving messages not meeting some priority are therefore subject to increased probability of drop. The dropping probability for each class of message is tuned by setting $t_{priority,min}$ and $t_{priority,max}$ for each class.

We consider the use of authentication as a means of creating messaging priority classes. For example, during a crisis, messages injected to a network from the Internet by an authenticated municipality or from emergency personnel could receive priority over all other text messages. A number of municipalities already use such systems for emergency [33] and traffic updates [34]. Messages from authenticated users within the network itself receive secondary priority. Unauthenticated messages originating from the Internet are delivered with the lowest priority. Such a system would allow the informative messages (i.e., evacuation plans, additional warnings) to be quickly distributed amongst the population. The remaining messages would then be delivered at ratios corresponding to their priority level. We assume that packet priority marking occurs at the SMSCs such that additional computational burden is not placed on base stations.

Here, we illustrate how WRED can provide differentiated service to different classes of SMS traffic using the attack scenario described in Tables I and II. We maintain separate queues, which are served in a round robin fashion, for voice requests and SMS requests. We apply WRED to the SMS queue. We assume legitimate text messages arrive at a sector

with an average rate of 0.7 *msgs/sec* with the following distribution: 10% high priority, 80% medium priority, and 10% low priority. The attack generates an additional 9 *msgs/sec*.

To accommodate sudden bursts of high priority SMS traffic, we choose an SMS queue size of 12. Because we desire low latency delivery of high priority messages, we target an average queue occupancy $Q_{avg} = 3$.

To meet this objective, we must set $t_{low,min}$ and $t_{low,max}$. For M/M/n systems with a finite queue of size m , the number of messages in the queue, N_Q , is:

$$N_Q = P_Q \frac{\rho}{1 - \rho} \quad (2)$$

where:

$$P_Q = \frac{p_0(m\rho)^m}{m!(1 - \rho)} \quad (3)$$

where:

$$p_0 = \left[\sum_{n=0}^{m-1} \frac{(m\rho)^n}{n!} + \frac{(m\rho)^m}{m!(1 - \rho)} \right]^{-1} \quad (4)$$

Setting $N_Q = 3$, we derive a target load $\rho_{target} = 0.855$. ρ_{target} is the utilization desired at the SDCCHs. Thus, the packet dropping caused by WRED must reduce the actual utilization, ρ_{actual} or $\lambda_{SMS}/(\mu_{SMS} \cdot n)$, caused by the heavy offered load during an attack, to ρ_{target} . Therefore:

$$\rho_{target} = \rho_{actual}(1 - P_{drop}) \quad (5)$$

where P_{drop} is the overall dropping probability of WRED. For traffic with average arrival rate of $\lambda_{SMS} = 9.7$ *msgs/sec*, $\rho_{actual} = 3.23$. Solving for P_{drop} ,

$$P_{drop} = 1 - \frac{\rho_{target}}{\rho_{actual}} = 0.736 \quad (6)$$

P_{drop} can be calculated from the dropping probabilities of the individual classes of messages by ($\lambda_{low} = 9.07$):

$$P_{drop} = \frac{P_{drop,high} \cdot \lambda_{high} + P_{drop,med} \cdot \lambda_{med} + P_{drop,low} \cdot \lambda_{low}}{\lambda_{SMS}} \quad (7)$$

Because we desire to deliver all messages of high and medium priority, we set $P_{drop,high} = P_{drop,med} = 0$. Using Equation 7, we find $P_{drop,low} = 0.787$. This value is then used in conjunction with Equation 1 to determine $t_{low,min}$ and $t_{low,max}$.

The desired average queue occupancy, Q_{avg} , is 3. From equation 1, $t_{low,min}$ must be an integer less than the average queue occupancy. This leaves three possible values for $t_{low,min}$: 0, 1, and 2. The best fit is found when $t_{low,min} = 0$ and $t_{low,max} = 4$, resulting in 75% dropping of low priority traffic.

Using this method it is possible to set thresholds to meet delivery targets. Of course, depending on the intensity of an attack, it may not be possible to meet desired targets according to Equation 7, i.e., it may not be possible to limit blocking to only low priority traffic. While the method outlined here provides just an approximate solution, given the quantization error in setting $t_{low,min}$ and $t_{low,max}$ (they must be integers), we believe the method is sufficient. We provide more insight into the performance of WRED in Section V.

⁴Some variants of RED additionally incorporate a *count* variable. Equation 1 is the simplest version of RED defined by RFC 2309 [32].

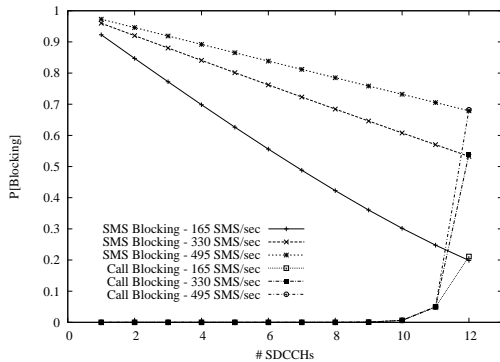


Fig. 9. The probability that incoming calls and SMS messages are blocked in a system implementing SRP. We vary the number of SDCCHs that will accept SMS requests from 1 to 12(all).

D. Resource Provisioning

None of the above methods deal with the system bottleneck directly; rather, they strive to affect traffic before it reaches the air interface. An alternative strategy of addressing targeted SMS attacks instead focuses on the reallocation of the available messaging bandwidth. We therefore investigate a variety of techniques that modify the way in which the air interface is used.

To analyze these techniques we resort to simple Erlang-B queueing analysis. We present a brief background here. For more details see Schwartz [30]. In a system with n servers, and an offered load in Erlangs of A , the probability that an arriving request is blocked because all servers are occupied is given by:

$$P_B = \frac{\frac{A^n}{n!}}{\sum_{l=0}^{n-1} \frac{A^l}{l!}} \quad (8)$$

The load in Erlangs is the same as the utilization, ρ , in a queueing system; it is simply the offered load multiplied by the service time of the resource. The expected occupancy of the servers is given by:

$$E(n) = \rho(1 - P_B) \quad (9)$$

1) *Strict Resource Provisioning*: Under normal conditions, the resources for service setup and delivery are over-provisioned. At a rate of 50,000 *calls/city/hour* in our baseline scenario, for example, the calculated average utilization of SDCCHs per sector is approximately 2%. Given this observation, if a subset of the total SDCCHs can be used only by voice calls, blocking due to targeted SMS attacks can be significantly mitigated. Our first air interface provisioning technique, *Strict Resource Provisioning* (SRP), attempts to address this contention by allowing text messages to occupy only a subset of the total number of SDCCHs in a sector. Requests for incoming voice calls can compete for the entire set of SDCCHs, including the subset used for SMS. In order to determine appropriate parameters for systems using SRP, we apply Equations 8 and 9.

To isolate the effectiveness of SRP, we consider a system with no queue. Figure 9 shows the blocking probabilities for

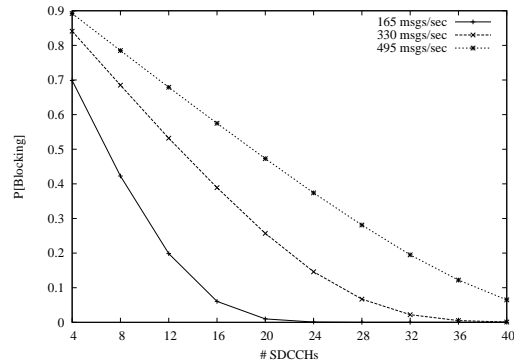


Fig. 10. The probability of an incoming call/message blocking in a sector for a varying number of SDCCHs.

a system using SRP when we vary the number of SDCCHs that will accept SMS requests from 1 to 12 (all). Because incoming text messages only compete with voice calls for a subset of the resources, any resulting call blocking is strictly a function of the size of the subset of voice-only SDCCHs. The attacks of intensity 165, 330, and 495 *msgs/city/sec* (3, 6, and 9 messages/second/sector) have virtually no impact on voice calls until the full complement of SDCCHs are made available to all traffic. In fact, it is not until 10 SDCCHs are made available to SMS traffic that the blocking probability for incoming voice calls reaches 1%.

By limiting the number of SDCCHs that will serve SMS requests, the blocking for SMS is increased. When only six SDCCHs are available to text messages, blocking probabilities for SMS are as high as 84%. Because significant numbers of people rely upon text messaging as their primary means of communication, such parameters should be carefully tuned. We will discuss the impact of additional factors after examining the results of simulation in Section V.

2) *Dynamic Resource Provisioning*: While SRP re-provisions capacity on existing SDCCHs, other over-provisioned resources in the sector could be manipulated to alleviate SDCCH congestion. For example, at a rate of 50,000 *calls/hour*, each sector uses an average of 67% of its TCHs. If a small number of unused TCHs could be repurposed as SDCCHs, additional bandwidth could be provided to mitigate such attacks.

Our second air interface technique, *Dynamic Resource Provisioning*, attempts to mitigate targeted text messaging attacks by temporarily reclaiming a number of TCHs (up to some limit) for use as SDCCHs. This approach is highly practical for a number of reasons. First, increasing the bandwidth (762 bits/second) of individual SDCCHs is difficult without making significant changes to either the radio encoding or the architecture of the air interface itself. Because major changes to the network are extremely expensive and typically occur over the course of many years, such fixes are not appropriate in the short term. Secondly, dynamically reclaiming channels allows the network to adjust itself to current conditions. During busy hours such as morning and evening commutes, for example, channels temporarily used as SDCCHs can be returned to the pool of TCHs to accommodate elevated voice traffic needs. Lastly, because SDCCHs are assigned via the AGCH, allocating incoming requests to seemingly random

timeslots requires almost no changes to handset software.

Figure 10 demonstrates the blocking probability for incoming calls and text messages in a sector using DRP to add a variable number of SDCCHs. Again, no queue was used. The ability of an attacker to block all channels is significantly reduced as the number of SDCCHs increases. Attackers are therefore forced to increase the intensity of their attack in order to maintain its potency. For attacks at a rate of 165 msgs/sec , doubling the number of available SDCCHs reduces the calculated blocking caused by an attack by two orders of magnitude. The blocking probability caused by attacks at higher rates, in which the number of Erlangs is greater than the number of SDCCHs, decreases in roughly a linear relationship to the number of SDCCHs added.

One potential drawback with DRP is that by subtracting TCHs from the system, it is possible to increase call blocking because of TCH exhaustion. In fact, the reclamation of TCHs for use as SDCCHs increases the blocking probability for voice calls from 0.2% in the base scenario (45 TCHs, 12 SDCCHs) to 1.5% where 40 SDCCHs are available (a reduction to 40 TCHs). Section V offers additional insight into the tradeoffs inherent to this scheme.

3) *Direct Channel Allocation*: From the security perspective, the ideal means of eliminating the competition for resources between call setup and SMS delivery would be through the separation of shared mechanisms. Specifically, delivering text messages and incoming call requests over mutually exclusive sets of channels would prevent these flows from interfering with each other. The challenge of implementing such a mechanism is to do so without requiring significant restructuring of the network architecture or incurring tremendous expense. As previously mentioned, such fundamental changes in network operation are typically too expensive and time consuming to be considered in the short term. While the SRP technique provides a rudimentary separation, it is possible to further isolate these two types of traffic.

As mentioned in the previous section, DRP is easily implementable because the AGCH specifies the location of the SDCCH allocated for a specific session. After call requests finish using their assigned SDCCH, they are instructed to listen to a specific TCH. Because the use of a TCH is the eventual goal of incoming voice calls, it is therefore possible to shortcut the use of SDCCHs for call setup. Incoming calls could therefore be directed to a TCH, leaving SDCCHs exclusively for the delivery of SMS messages. This technique, which we refer to as *Direct Channel Allocation (DCA)*, removes the shared SDCCH channels as the system bottleneck.

Calculating blocking probabilities for a system implementing DCA is a matter of analyzing SDCCH and TCH blocking for the two independent flows. For 165 msgs/sec , text messages have a calculated blocking probability of approximately 20%. This value increases to 68% as the attack intensity increases to 495 msgs/sec . Voice calls, at an average rate of $50,000 \text{ calls/hour}$, have a blocking probability of 0.2%. Note that because the shared bottleneck has been removed, it becomes extremely difficult for targeted text messaging attacks to have any effect on voice communications. In Section V, we will highlight these new potential points of contention.

V. RESULTS AND DISCUSSION

In order to characterize each of our proposed mitigation techniques, we simulate attacks against networks with the parameters of Tables I and II unless otherwise noted. The RACH parameters used optimal settings [35]. We also recorded no blocking on the RACH or SS7 signaling links during any simulation. Because of small deviations from the mean (95% confidence interval of ± 0.006), we omit error bars for clarity. Additional information regarding the verification of the simulator and statistical significance of our results are available in the Appendix.

As previously mentioned, the techniques provided in the following subsections are designed to preserve the availability of voice telephony and then maximize the throughput of legitimate text messaging. The performance and tradeoffs associated with each approach vary significantly. We summarize these results at the end of each subsection.

A. Queue Management Strategies

In the following two subsections we present our simulation results for WFQ and WRED. For both solutions, we maintain queues of size 6 and 12 for call requests and SMS messages, respectively. We experimented with several values of queue size and found these to provide a good balance between additional latency and robustness. Note that because the arrival rate of SMS messages is greater than the processing rate ($\rho > 1$), no finite queue can prevent dropping.

1) *Weighted Fair Queueing*: Buffering alone is not sufficient to protect against congestion [28], [36]; rather, mechanisms designed to mediate tail drop blocking are necessary. We apply WFQ with a weight of 2 for calls and 1 for SMS messages to ensure voice calls receive a suitable amount of SDCCH bandwidth.

Figure 11 illustrates the resulting blocking for a sector implementing WFQ. The preferential treatment of voice traffic eliminates the blocking previously seen in an unprotected system. Incoming text messages, however, continue to experience roughly the same blocking (72%) observed by all traffic in the base attack scenario. As is shown in Figure 12, the queue itself does nothing to prevent congestion. Total queue utilization is 65%. As two-thirds of the queue space is available to text messaging, this represents a near total average occupancy of the SMS queue and a virtually unused voice traffic queue. Such an observation conforms to our analytical results. This figure also demonstrates the ability to protect voice services, as TCH utilization is not lowered during the attack.

The advantage to implementing the WFQ mechanism is not only its relative simplicity, but also its effectiveness in preventing degradation of voice services during targeted SMS attacks. Unfortunately, the granularity for prioritizing text messages is insufficient to provide adequate service to those users relying upon text messaging as their dominant means of communication. Accordingly, if users believe that their traffic is unlikely to be delivered, their faith in text messaging as a reliable service will decrease. While finer granularity can be provided by adding one queue per SMS class, this solution will result in inefficient memory use and complexity. We discuss means of adding such granularity through the use of WRED.

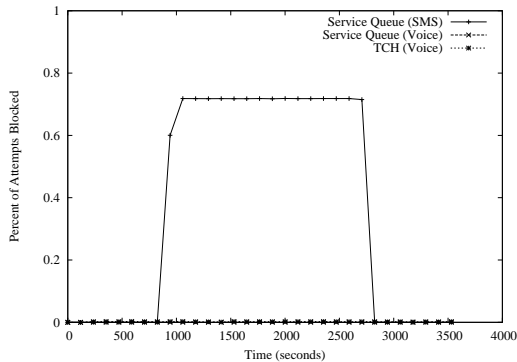


Fig. 11. The simulated blocking probability for a sector implementing WFQ. Notice that voice calls are unaffected by the attack, whereas the majority of text messages are dropped.

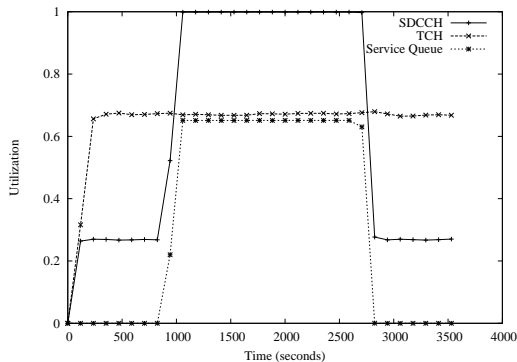


Fig. 12. Utilization for a sector implementing WFQ. Note that TCH utilization is constant throughout the attack.

2) *Weighted Random Early Detection*: The use of a prioritized dropping policy allows a system to offer similar prioritization to WFQ while maintaining only a single queue. In our implementation of WRED, we maintain one queue for voice requests (size of 6) and one queue for SMS messages (size 12) and apply WRED to the SMS queue. We differentiate the SMS traffic by setting different thresholds for each class. We assume that SMS traffic is marked upstream as having high, medium, or low priority. We assign the thresholds as ($t_{high,max} = t_{high,min} = 12$), medium ($t_{med,max} = 10, t_{med,min} = 6$) and low ($t_{low,max} = 4, t_{low,min} = 0$) priority. These priorities correspond directly to emergency priority users, network customers and Internet-originated messages, respectively. Q_{avg} is maintained as a simple weighted average with a weight of 0.8 on the most recent sample.

Figure 13 gives the blocking for each of the three priorities of text messages. Because voice calls never block in these simulations, we omit them from this graph. Both high and medium priority flows also do not experience blocking throughout the simulations. The blocking of Internet-originated messages averages 77%, approximately the same blocking probability experienced by all incoming messages in the base attack scenarios. Service queue utilization, shown in Figure 14, is 20%. With a total queue capacity of 18, this corresponds to an average occupancy of 3.88 messages. Also notice that the TCH occupancy is maintained throughout the attack.

The parameters used in this simulation are the same as those in Section IV. We set the medium priority thresholds to

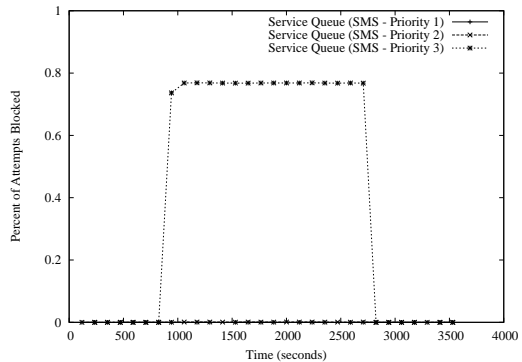


Fig. 13. The simulated blocking probability for a sector implementing WRED. Unlike WFQ, only Internet-originated text messages are dropped at an elevated frequency.

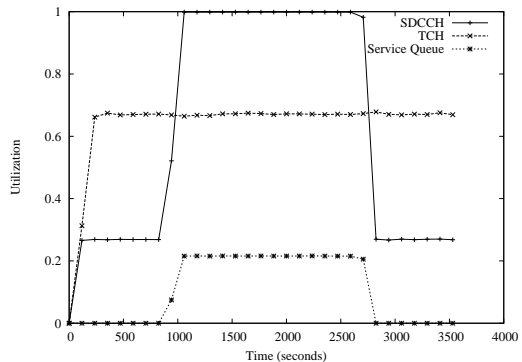


Fig. 14. Utilization for a sector implementing WRED. Note that the queue occupancy is low due to the decreased priority of Internet-originated messages.

allow some loss at very high loads to protect the high priority traffic under extreme circumstances, but because our average queue occupancy is about 3.9, no dropping of medium priority messages occurs. This matches well with our analytical results.

Systems implementing WRED not only match the elimination of voice call blocking seen through the use of WFQ, but also offer significantly improved performance in terms of message delivery. Implementing this solution, however, faces its own challenges. The authentication of high priority messages, for example, would require the use of additional infrastructure. High priority messages originating outside the network, such as emergency messages distributed by a city, may require the use of a dedicated line and/or the use of a public key infrastructure (PKI) for authentication. Because of historical difficulties effectively achieving the latter [37], implementing such a system may prove difficult. Even with such protections, this mechanism fails to protect the system against insider attacks. If the machine responsible for sending high priority messages into the network or user phones are compromised by malware, systems implementing WRED lose their messaging performance improvements over the WFQ solution. Note that networks not bounding priority to specific geographic regions can potentially be attacked through any compromised high priority device.

3) *Summary*: As demonstrated in this subsection, queue management techniques are a valuable tool for protecting voice telephony from targeted SMS attacks. The benefits of these schemes, however, are more limited than in traditional data

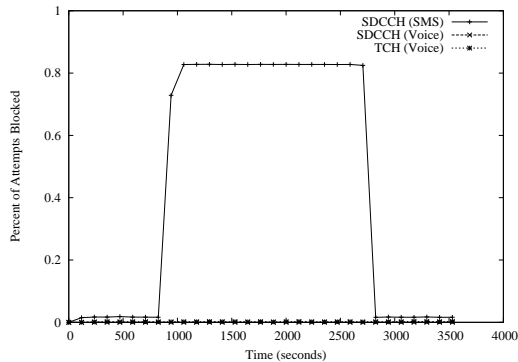


Fig. 15. Blocking for a sector implementing SRP

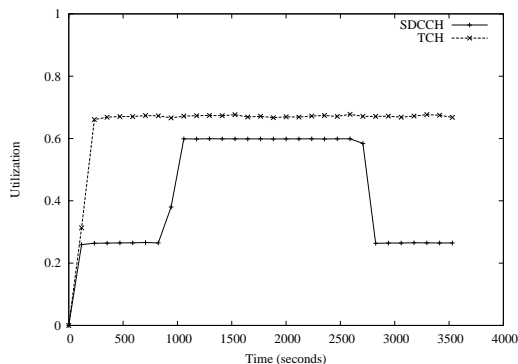


Fig. 16. Channel utilization under SRP

networks. For instance, the shedding of TCP packets in an IP network should cause flow control mechanisms to reduce transmission rates. Because the same does not happen here, queue management techniques essentially “bail water” until an attack subsides. Accordingly, such mechanisms should be relied upon as a last line of defense. We therefore look to techniques that reapportion the resources on the air interface for a more flexible response.

B. Air Interface Strategies

1) *Strict Resource Provisioning*: Before characterizing the SRP technique, careful consideration was given to the selection of operating parameters. Because many MSCs are capable of processing up to 500K calls/hour, we engineer our solution to be robust to large spikes in traffic. We therefore allow SMS traffic to use 6 of the 12 total SDCCHs, which yields a blocking probability of 1% of voice calls by the SDCCH when voice traffic requests reach 250,000 calls/hour. (Note that calls would experience an average blocking probability of 71% due to a lack of TCHs with requests at this intensity.) Because these networks are designed to operate dependably during elevated traffic conditions, we believe that the above settings are realistic.

The blocking probabilities for SMS and voice flows in a sector implementing SRP are shown in Figure 15. Because SRP prevents text messages from competing for all possible SDCCHs, voice calls experience no blocking on the SDCCHs throughout the duration of the attack. Text messages, however, are blocked at a rate of 83%. Channel utilization, illustrated in Figure 16, gives additional insight into network conditions. Because calling behavior remains the same during the attack,

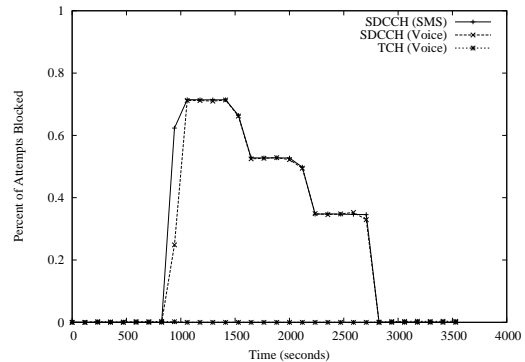


Fig. 17. Blocking for a sector implementing DRP

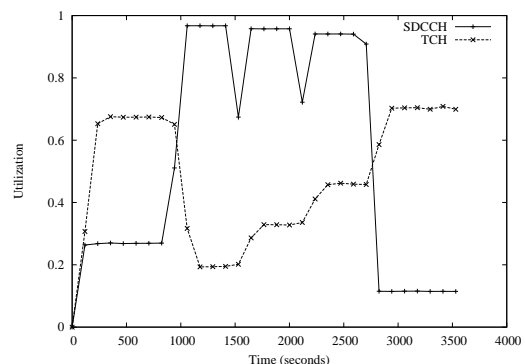


Fig. 18. Channel utilization under DRP

the resources allocated by the network are more than sufficient to provide voice service to users. By design, SDCCH utilization plateaus well below full capacity. While the SDCCHs used by text messages have an average utilization of 97%, the SDCCHs used by incoming voice calls average a utilization of 6.3%. This under-use of resources represents a potential loss of utility as the majority of text messages (legitimate or otherwise) go undelivered.

The difficulty with this solution is correct parameter setting. While theoretical results indicate that allocating 10 SDCCHs only increases call blocking to 1%, voice traffic volumes fluctuate throughout the day. Provisioning resources in a static fashion must account for worst-case scenarios and therefore leads to conservative settings. While protecting the network from an attack, such a mechanism may actually hinder the efficiency of normal operation. When traffic channels are naturally saturated, as may be common during an emergency or elevated traffic scenario discussed in Section III-C, such hard limits actually prevent users from communicating. Furthermore, as unsustained bursts of text messages are generally innocuous, such a limitation may directly impact the provider’s ability to generate revenue as user perception of SMS as a real-time service erodes. Determining the correct balance between insulation from attacks and resource utilization becomes non-trivial. Accordingly, we look to our other techniques for more complete solutions.

2) *Dynamic Resource Provisioning*: Although it is possible to reclaim any number of TCHs for use as SDCCHs under the DRP mechanism, we limited the candidate number of channels for this conversion to two. In these experiments, a single TCH

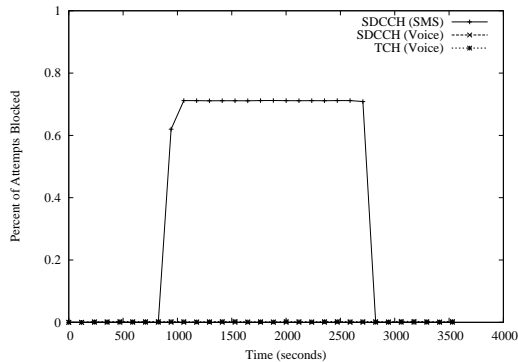


Fig. 19. Blocking for a sector implementing DCA

was repurposed into 8 SDCCHs every 10 minutes during the attack. This separation was designed to allow the network to return to steady state between channel allocations. While converting only two channels is not enough to completely eliminate attacks at high intensities, our goal is to understand the behavior of this mechanism.

The blocking probabilities for SMS and voice flows in a sector implementing the DRP technique are illustrated in Figure 17. As TCHs are converted for use as SDCCHs, the blocking probabilities for both incoming SMS and voice requests fall from 72% to 53% and eventually 35%. This represents a total reduction of the blocking probability by approximately half. Call blocking due to TCH exhaustion was not observed despite the reduced number of available TCHs. Figure 18 illustrates a gradual return towards pre-attack TCH utilization levels as additional SDCCHs are allocated. The effects of the reprovisioning are also obvious for SDCCH utilization. The downward spikes represent the sudden influx of additional, temporarily unused channels. While SDCCH utilization quickly returns to nearly identical levels after each reallocation, more voice calls can be completed due to a decrease probability of the attack holding all SDCCHs at any given time.

As was a problem for SRP, determining the correct parameters for DRP is a difficult undertaking. The selection of two TCHs for conversion to SDCCHs illustrates the utility of this mechanism, but is not sufficient for real settings. To reduce the blocking probability on SDCCHs below the values observed for TCHs, a total of 48 SDCCHs would have to be made available. This leaves 39 TCHs, which results in a call blocking probability of 2.1% due to TCH exhaustion. Elevations in the volume of voice calls would likely require the release of some number of reclaimed TCHs to be repurposed to their original use.

The decision to convert channels is also non-trivial. Whereas the decision to reallocate channels at specific times was decided statically in our simulation, dynamically determining these parameters would prove significantly more challenging. Basing reclamation decisions on small observation windows, while offering greater responsiveness, may result in decreased resource use due to thrashing. If the observation window becomes too large, an attack may end before appropriate action can be taken. As was observed for SRP, the static allocation of

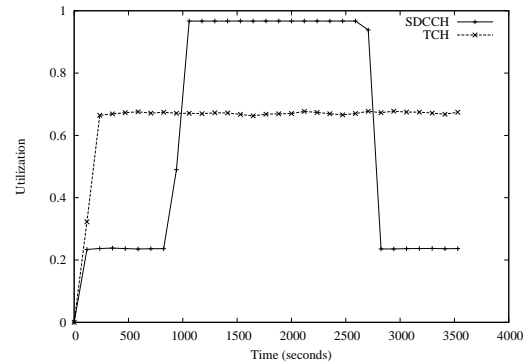


Fig. 20. Channel utilization under DCA

additional SDCCHs faces similar inflexibility problems. Low resource utilization under normal operating conditions again represents a potential loss of opportunity and revenue.

3) *Direct Channel Allocation*: To simulate the DCA mechanism, incoming voice calls skip directly from the RACH to the next available TCH. An average of 1.5 additional seconds was added to each incoming call duration to account for the processing formerly occurring on an SDCCH. As is shown in Figure 19, voice calls arriving in a sector implementing the DCA scheme experience no additional blocking during a targeted SMS attack. Figure 20 confirms the results in the previous figure by showing the constant TCH utilization throughout the duration of the attack. No additional assistance is provided for the delivery of text messages under DCA.

While removing the bottleneck on the shared path of SMS delivery and voice call setup, DCA potentially introduces new vulnerabilities into the network. One advantage of using SDCCHs to perform call establishment is that users are authenticated before they are assigned TCHs. Under the DCA model, however, valuable traffic channels can be occupied before users are ever authenticated. Using a single phone planted in a targeted area, an attacker could simply respond to all paging messages and then ignore all future communications from the network. Because there are legitimate reasons to wait tens of seconds for a phone to reply to a page, an attacker could force the network to open and maintain state for multiple connections that would eventually go unused. Note that because paging for individual phones occurs over multiple sectors, a single rogue phone could quickly create a black-hole effect. Such an attack is very similar to the classic SYN attack observed throughout the Internet. While seemingly the most complete, the potential for additional damage made possible because of the DCA approach should be carefully considered.

4) *Summary*: The resource management techniques presented above offer a number of valuable countermeasures against targeted SMS attacks. At a high level, SRP provides functionality to the weighted fair queueing approach under high load by ensuring that channels are always available to voice traffic. SRP, however, experiences even high rates of SMS blocking than weighted fair queueing (83% vs 72%). DRP allows the network to accommodate spikes in traffic by reapportioning unused TCHs, thereby making the network more flexible to a wider range of operating conditions. As

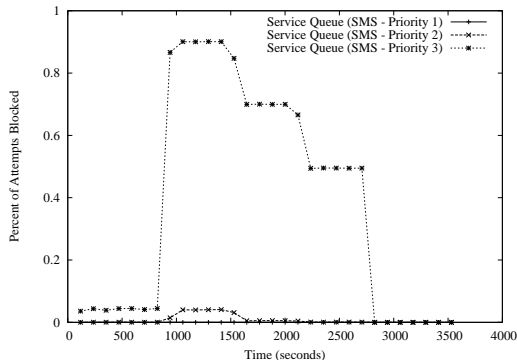


Fig. 21. Blocking for a sector implementing both WRED and DRP.

we discovered in the DCA case, however, the repurposing of resources must be carefully executed so as to not introduce new vulnerabilities into the system.

C. Combining Mechanisms

There is no “silver-bullet” for maintaining a high quality of service for both text messaging and voice calls during a targeted SMS attack. As the above techniques demonstrate, each potential solution has its own weaknesses. The combination of such solutions, however, offers techniques robust to a wider array of threats. We examine two examples in which the fusion of mechanisms provides additional protections.

While directly addressing the bandwidth issue that makes targeted SMS attacks possible, the DRP technique lacks granularity to separate incoming voice and SMS requests. WRED, on the other hand, provides such traffic classification but is unable to react to attacks originating from trusted sources. To illustrate the benefits of layering these techniques, we increase the volume of legitimate traffic to $2 \text{ msgs/sector/sec}$, with 90% of that traffic being medium priority and the remaining 10% split equally between high and low priority flows. Such an increase would be representative of the elevated volumes of messages sent from crowded events such as concerts or public celebrations such as New Year’s Eve gatherings. Figure 21 shows the result of the combination of the two techniques during an attack. Because of the naturally increased volume of legitimate traffic, subscriber-to-subscriber traffic experiences approximately 5% blocking in a sector only implementing WRED. As DRP activates and adds additional SDCCHs, only the attack traffic is dropped. Such a technique may be especially valuable during an emergency, as additional bandwidth can be provisioned to clients less likely to be malicious.

Another potentially beneficial combination is SRP and DRP. Given high volumes of voice traffic, a provider may not be able to repurpose enough SDCCHs to eliminate the effects of a targeted text messaging attack. Instead, a subset of the total channels could be reserved for voice requests. In so doing, voice blocking due to targeted text messaging attacks could be eliminated. All additional channels could be added to reduce blocking for text messages. Figure 22 illustrates an attack scenario in which two TCHs are reclaimed for use as SDCCHs, with 18 of 24 total SDCCHs made available to SMS. Note both the elimination of call blocking and the gradual reduction of blocking rates for text messages.

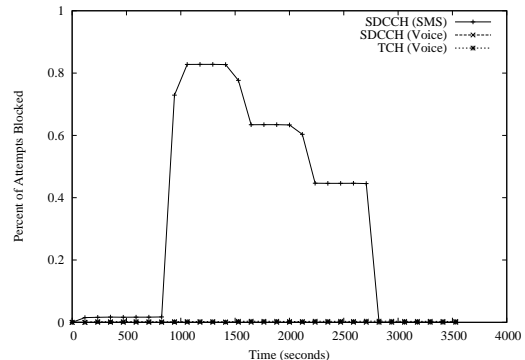


Fig. 22. Blocking for a sector combining DRP and SRP.

Other combinations are less useful. Integrating WRED and SRP, for example, would simply reduce the bandwidth made available for even high priority mechanisms. Accordingly, the network may experience decreased throughput for legitimate messages than under either scheme alone. The use of DCA with any other mechanism fails to prevent the vulnerability introduced in the previous subsection, and therefore does not warrant further investigation.

While no susceptible examples were uncovered during the course of this research, the combination of any of the above mitigation techniques should also be carefully considered. This fusion may lead to the creation of new or magnification of previously mentioned vulnerabilities. Accordingly, additional testing on development networks should be conducted before such integration could occur.

VI. CONCLUSION

In this paper, we characterize the impact of targeted SMS attacks on GSM cellular networks. Our simulations and analyses provide additional attestation that the network bottleneck resides in the SDCCHs and that an adversary can cause blocking probabilities as high as 70% across Manhattan with a cable modem. We then present a number of techniques from queue management and resource provisioning with the goal of maintaining the availability of voice telephony and providing high throughput for text messaging with varying results. WFQ and WRED offer a last line of defense by separating traffic, but fail to help the network absorb additional traffic. SRP functions similarly. The remaining two methods, DRP and DCA, allow the network to dedicate unused resources to the problem; however, DCA creates a serious new vulnerability and is therefore not a viable solution. In spite of these shortcomings, all of the above techniques except for DCA offers an effective means of mitigating the impact of targeted SMS attacks on voice calls.

The attacks discussed throughout are representative of growing and increasingly problematic class of vulnerabilities. The connectivity between the Internet and traditional voice networks introduces new avenues for exploit: once confined to exploiting only inert hosts, remote adversaries can debilitate the services we depend on to carry on our daily lives. In a broader sense, the ability to control the physical world via the Internet is inherently dangerous, and more so when the affected components are part of critical infrastructure. This

work provides some preliminary solutions and analysis for these vulnerabilities. Essential future work will seek more general solutions that address these vulnerabilities in current and next generation networks.

REFERENCES

- [1] K. Maney, "Surge in text messaging makes cell operators :-), July 27 2005.
- [2] "Young prefer texting to calls," <http://news.bbc.co.uk/2/hi/business/2985072.stm>, June 2003.
- [3] W. Enck, P. Traynor, P. McDaniel, and T. F. La Porta, "Exploiting Open Functionality in SMS-Capable Cellular Networks," in *Proceedings of the ACM Conference on Computer and Communication Security (CCS)*, November 2005.
- [4] "The National Strategy to Secure Cyberspace," http://www.us-cert.gov/reading-room/cyberspace_strategy.pdf, February 2003.
- [5] D. Moore, V. Paxson, S. Savage, C. Shannon, S. Staniford, and N. Weaver, "Inside the Slammer Worm," *IEEE Security and Privacy*, vol. 1, no. 4, July 2003.
- [6] S. Byers, A. Rubin, and D. Kormann, "Defending Against an Internet-based Attack on the Physical World," *ACM Transactions on Internet Technology (TOIT)*, vol. 4, no. 3, pp. 239–254, August 2004.
- [7] M. Richtel, "Yahoo Attributes a Lengthy Service Failure to an Attack," *The New York Times*, February 8 2000.
- [8] C. Haney, "NAI is latest DoS victim," http://security.itworld.com/4339/NWW116617_02-05-2001/page.1.html, February 5 2001.
- [9] P. Roberts, "Al-Jazeera Sites Hit With Denial-of-Service Attacks," *PCWorld Magazine*, March 26 2003.
- [10] S. Berinato, "Online Extortion – How a Bookmaker and a Whiz Kid Took On an Extortionist and Won," *CSO Online*, May 2005.
- [11] J. Mirkovic and P. Reiher, "A Taxonomy of DDoS Attacks and DDoS Defense Mechanisms," *ACM SIGCOMM Computer Communication Review*, vol. 34, no. 2, pp. 39–53, 2004.
- [12] A. Juels and J. G. Brainard, "Client Puzzles: A Cryptographic Countermeasure Against Connection Depletion Attacks," in *Proceedings of Network and Distributed System Security Symposium (NDSS)*, 1999.
- [13] S. Savage, D. Wetherall, A. Karlin, and T. Anderson, "Practical network support for IP traceback," in *Proceedings of ACM SIGCOMM*, October 2000, pp. 295–306.
- [14] J. Ioannidis and S. Bellovin, "Implementing Pushback: Router-Based Defense Against DDoS Attacks," in *Proceedings of Network and Distributed System Security Symposium (NDSS)*, February 2002.
- [15] A. Keromytis, V. Misra, and D. Rubenstein, "SOS: Secure Overlay Services," in *Proceedings of ACM SIGCOMM*, 2002.
- [16] B. Waters, A. Juels, J. Halderman, and E. Felten, "New client puzzle outsourcing techniques for DoS resistance," in *Proceedings of ACM CCS'04*, 2004, pp. 246–256.
- [17] National Communications System, "SMS over SS7," http://www.ncs.gov/library/tech_bulletins/2003/tib_03-2.pdf, Tech. Rep. Technical Information Bulletin 03-2 (NCS TIB 03-2), December 2003.
- [18] Mike Grenville, "Operators: Celebration Messages Overload SMS Network," <http://www.160characters.org/news.php?action=view&nid=819>, November 2003.
- [19] Nyquetek, Inc. "Wireless Priority Service for National Security," <http://wireless.fcc.gov/releases/da051650PublicUse.pdf>, 2002.
- [20] P. Traynor, W. Enck, P. McDaniel, and T. La Porta, "Mitigating Attacks on Open Functionality in SMS-Capable Cellular Networks," in *Proceedings of the Twelfth Annual ACM International Conference on Mobile Computing and Networking (MobiCom)*, September 2006.
- [21] Lucent Technologies, "5ESS(R) 2000 - Switch Mobile Switching Centre (MSC) for Service Providers," <http://www.lucent.com/products/solution/0,,CTID+2019-STID+10048-SOID+824-LOCL+1,00.html>, 2006.
- [22] Motorola Corporation, "Motorola GSM Solutions," www.motorola.com/networkoperators/pdfs/GSM-Solutions.pdf, 2006.
- [23] R. Isukapalli, T. Alexiou, and K. Murakami, "Global Roaming and Personal Mobility with COPS Architecture in SuperDHLR," *Bell Labs Technical Journal*, vol. 7, no. 2, pp. 3–18, 2002.
- [24] M. Whitehead, "GOCAP – one standardised overload control for next generation networks," *BT Technology Journal*, vol. 23, no. 1, January 2005.
- [25] The Internet Engineering Task Force, "Congestion and Pre-Congestion Notification (PCN)," <http://www.ietf.org/html.charters/pcn-charter.html>, 2007.
- [26] G. Kunene, "Perimeter Security Ain't What It Used to Be, Experts Say," *DevX.com*, 2004.
- [27] P. Traynor, P. McDaniel, and T. La Porta, "On Attack Causality in Internet-Connected Cellular Networks," in *Proceedings of the USENIX Security Symposium*, 2007.
- [28] J. B. Nagle, "On Packet Switches with Infinite Storage," *IEEE Transactions on Communications*, vol. COM-35, no. 4, April 1987.
- [29] A. Demers, S. Keshav, and S. Shenker, "Analysis and Simulation of a Fair Queueing Algorithm," in *Proceedings of ACM SIGCOMM*, 1989, pp. 3–12.
- [30] M. Schwartz, "," in *Telecommunication Networks - Protocols, Modeling and Analysis*. Addison-Wesley Publishing Company, 1987.
- [31] S. Floyd and V. Jacobson, "Random Early Detection Gateways for Congestion Avoidance," *IEEE/ACM Transactions on Networking*, vol. 1, no. 4, pp. 397–413, August 1993.
- [32] B. Branden, D. Clark, J. Crowcroft, B. Davie, S. Deering, D. Estrin, S. Floyd, V. Jacobson, G. Minshall, C. Partridge, L. Peterson, K. Ramakrishnan, S. Shenker, J. Wroclawski, and L. Zhang, "RFC 2309 - Recommendations on Queue Management and congestion Avoidance in the Internet," rfc2309.txt, 1998.
- [33] Roam Secure, "17 Counties & Cities in Washington, DC Region deploy Roam Secure Alert Network," http://www.roamsecure.net/story.php?news_id=52, September 2005.
- [34] Tamara Neale, "VDOT LAUNCHES NEW 511 EMAIL ALERT SERVICE," <http://www.virginiadot.org/infoservice/news/newsrelease.asp?ID=CO-511-06>, February 2006.
- [35] C. Luders and R. Haferbeck, "The Performance of the GSM Random Access Procedure," in *Vehicular Technology Conference*, June 1994, pp. 1165–1169.
- [36] R. Jain, "Myths about congestion management in high speed networks," *Interworking: Research and Experience*, vol. 3, pp. 101–113, 1992.
- [37] C. M. Ellison and B. Schneier, "Ten Risks of PKI: What You're Not Being Told About Public-Key Infrastructure," *Computer Security Journal*, vol. 16, no. 1, pp. 1–7, 1999.
- [38] J. Hedden, "Math::Random::MT::Auto - Auto-seeded Mersenne Twister PRNGs," <http://search.cpan.org/~jhedden/Math-Random-MT-Auto-5.01/lib/Math/Random/MT/Auto.pm>, version 5.01.
- [39] 3rd Generation Partnership Project, "Physical layer on the radio path; General description, Tech. Rep. 3GPP TS 04.18 v8.26.0.
- [40] —, "Physical layer on the radio path; General description, Tech. Rep. 3GPP TS 05.01 v8.9.0.

Patrick Traynor is a Ph.D. candidate in the department of Computer Science and Engineering at the Pennsylvania State University, where he is the lead graduate student for the Systems and Internet Infrastructure Security Laboratory. Patrick received his B.S. in Computer Science from the University of Richmond in 2002 and his M.S. in Computer Science and Engineering from the Pennsylvania State University in 2004. His research efforts are focused primarily on the security of interconnected IP and telecommunications networks.

William Enck is a PhD candidate in the department of Computer Science and Engineering at the Pennsylvania State University. William received a B.S. and M.S. in Computer Science and Engineering from the Pennsylvania State University in 2004 and 2006, respectively. He is currently interested in security for telecommunications and other areas of network and operating systems.

Patrick McDaniel is the Hartz Family Career Development Assistant Professor in the Computer Science and Engineering Department at the Pennsylvania State University, and co-director of the Systems and Internet Infrastructure Security Laboratory. He received his Ph.D. from the University of Michigan in 2001 where he studied the form, algorithmic limits, and enforcement of security policy. Prior to joining Penn State, Patrick was a senior technical staff Member of the Secure Systems Group at AT&T Labs-Research and Adjunct Professor of the Stern School of Business at New York University.

Patrick's recent research efforts have focused on telecommunications security, distributed systems security, network security, language-based security, and public policy and technical issues in digital media. Patrick is a past recipient of the NASA Kennedy Space Center fellowship, a frequent contributor to the IETF security standards, and has authored many papers and book chapters in various areas of systems security. He is the co-chair of the 2007 and 2008 IEEE Symposium on Security and Privacy, and served as the Program Chair of the 2005 USENIX Security Symposium, the Vice Chair for Security and Privacy for WWW 2005, and is the Chair of the Industry and Government Track at the 2005 and 2007 ACM Computer and Communications Security conference. Patrick is also an associate editor of the journal ACM Transactions on Internet Technologies and a guest editor of the IEEE Transactions on Software Engineering. Prior to pursuing his Ph.D. in 1996, Patrick was a software architect and program manager in the telecommunications industry.

Thomas La Porta received his B.S.E.E. and M.S.E.E. degrees from The Cooper Union, New York, NY, and his Ph.D. degree in Electrical Engineering from Columbia University, New York, NY. He joined the Computer Science and Engineering Department at Penn State in 2002 as a Full Professor. He is the Director of the Networking and Security Research Center at Penn State. Prior to joining Penn State, Dr. La Porta was with Bell Laboratories since 1986. He was the Director of the Mobile Networking Research Department in Bell Laboratories, Lucent Technologies where he led various projects in wireless and mobile networking. He is a Bell Labs Fellow.

Dr. La Porta was the founding Editor-in-Chief of the IEEE Transactions on Mobile Computing and served as Editor-in-Chief of IEEE Personal Communications Magazine. He is currently the Director of Magazines for the IEEE Communications Society and

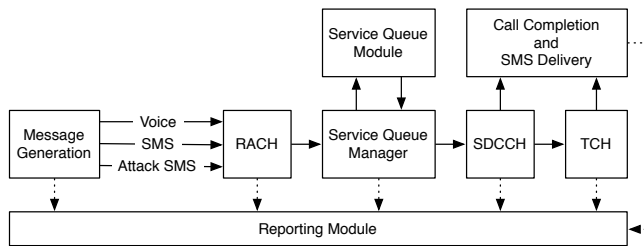


Fig. 23. Simulator Architecture

is a member of the Communications Society Board of Governors. He has published over 50 technical papers and holds 28 patents. His research interests include mobility management, signaling and control for wireless networks, mobile data systems, and protocol design.

VII. APPENDIX

While analytics easily characterize simplistic network loads, it is difficult to analyze multi-faceted input flows. To more fully characterize targeted SMS attacks and mitigation techniques, we developed an extensible, detailed GSM simulator. A number of simulation environments, including OPNET and NS2, were considered. The GSM air interface modules for both OPNET and NS2 focus on mobility management and did not contain support for SMS. More importantly, neither design was amenable to modeling potential countermeasures. Commercial GSM simulators were also considered, but were highly complex, specialized, and cost prohibitive.

The core simulator is over 5,000 lines of C with supplement scripts written in various languages. In total, the project contains over 8,000 lines of code. All code was written with flexibility of configuration and extensibility in mind. Figure 23 depicts our simulator architecture. Solid lines indicate voice and SMS message flow; dashed lines indicate reporting of events, including creation, stage entrance, blocking, and completion. Simulation begins in the *Message Generation stage*, where messages are randomly created using a Mersenne Twister Pseudo Random Number Generator [38]. Message generation can follow Poisson, uniform, or bursty arrival patterns. After creation, messages proceed to the *RACH stage*, which strictly follows 3GPP TS 04.18 [39] and is tunable using the *max_retrans* and *tx_integer* variables. The *Service Queue Manager stage* assigns messages to an SDCCH. If desired, a pluggable *Service Queue Module* can be defined using standard interface callback functions. If possible, the Service Queue Manager assigns a message to an SDCCH. Rather than simulating exact communication and compensating for retransmission, messages are held in the *SDCCCH stage* for an exponential mean service time corresponding to message type. For accuracy, each SDCCH services messages by decreasing counters only during frames defined in 3GPP TS 05.01 [40]. When counters reach zero, SMS messages complete and voice messages attempt to acquire a TCH. Like the SDCCH stage, the *TCH stage* uses an exponential mean hold time to simulate channel occupancy. TCHs service messages during every frame, and when the hold time counter reaches zero, the call is complete, and the TCH is released.

The accuracy of the simulator was verified by running simple base scenarios and comparing blocking and utilization

to values obtained using equations 8 and 9. Base scenarios similar to those in the paper were created: 12 SDCCHs, 45 TCHs, voice and SMS service times with exponential means of 1.5 and 4.0 seconds respectively, voice holding time with exponential mean of 120 seconds, voice call loads representing 25K, 50K, 75K, and 100K calls per hour to Manhattan, and SMS messages loads of 1, 2, and 3 messages per second to a sector. Voice and SMS loads were simulated separately and the average blocking and utilization of 1,000 runs was compiled. Both blocking and utilization are measured on a scale from 0 to 1.0. Statistical analysis showed that 95% of all samples fell within 0.006 of the mean, with many close to 0.001. For example, SDCCH utilization for the SMS load of 3 *msgs/sec* was 0.809609 ± 0.00133944 . Compared to calculated blocking and utilization, all simulated means deviated less than 20%, with all utilization means less than 2%. All but one simulated blocking mean was within 10% calculated values. The outlier was resulted from simulating a very small blocking value and was off by -5×10^{-4} , which therefore is acceptable.