Analysis of the Reliability of a Nationwide Short Message Service

Xiaoqiao Meng
NEC Laboratories America
Princeton, NJ 08540
Email: xqmeng@nec-labs.com

Petros Zerfos
Deutsche Telekom Laboratories
10587 Berlin, Germany
Email: petros.zerfos@telekom.de

Vidyut Samanta, Starsky H.Y. Wong, Songwu Lu
UCLA Computer Science Department
Los Angeles, CA 90095
Email: {vids, hywong1, slu}@cs.ucla.edu

***Abstract*—SMS has been arguably the most popular wireless data service for cellular networks. Due to its ubiquitous availability and universal support by mobile handsets and cellular carriers, it is also being considered for emergency notification and other mission-critical applications. Despite its increased popularity, the reliability of SMS service in real-world operational networks has received little study so far. In this work, we investigate the reliability of SMS by analyzing traces collected from a nationwide cellular network over a period of three weeks. Although the SMS service incorporates a number of reliability mechanisms such as delivery acknowledgement and multiple retries, our study shows that its reliability is not as good as we expected. For example the message delivery failure ratio is as high as 5.1% during normal operation conditions. We also analyze the performance of the service under stressful conditions, and in particular during a "flash-crowd" event that occurred in New Year's Eve of 2005. Two important factors that adversely affect reliability of SMS are also examined: bulk message delivery that may induce network-wide congestion, and the topological structure of the social network formed by SMS users, which may facilitate quick propagation of viruses or other malware.**

I. Introduction

The Short Message Service (SMS) has been arguably the most popular wide-area wireless data service worldwide. According to recent data [1], it accounts for more than 80% of data revenue generated for Western European mobile operators in 2005. The two large telecom operators in China (China Mobile and China Unicom) reported the SMS volume to be 304.14 billion messages in 2005. The SMS volume in the first quarter of 2006 recorded a 47% growth and reached 139.25 billion messages. The soaring popularity of SMS, along with its ubiquitous availability and support by handsets and cellular carriers, has stimulated numerous mobile data service providers to use it as their underlying data transport facility to maximize the reach to their potential customer base. As text messaging becomes an indispensable convenience of modern life in many parts of the world, cell-phone users place high expectations on it as an instant and reliable means for data communication. Even more importantly, SMS is being considered for mission-critical applications such as emergency alerts [2] and notification for natural disasters [3], for which reliable operation is of paramount importance.

Recent studies [4] [5] take a first look on the security issues related to the short message service and in particular its susceptibility to denial-of-service attacks and vulnerability to malware. They follow a "gray-box" approach for their analysis, either through probing from end-user handset devices or via simulations, due to the closed nature of cellular networks. Complementary to these efforts, in this work we offer an insider's view on aspects that affect the *reliable message delivery* of the service, through the analysis of traces that were obtained from a nation-wide cellular carrier.

This measurement-based study focuses on the reliability of the SMS service during both normal and overload operating conditions. Generally speaking, reliability denotes the ability of a system to consistently provide service with certain performance characteristics, even in the presence of stressful conditions that may threaten to disrupt the offered service. It may take several forms and includes notions such as high availability, resiliency, fault tolerance, and even security, depending on the context [2]. In this work, we first seek to establish a baseline characterization of the reliability for end-to-end short message transfer during normal traffic conditions. We then examine reliability during stress conditions of "flash-crowd" events [6] that are frequently observed during special occasions such as the New Year's Eve.

Two factors that may critically affect reliability of SMS in the near future are further investigated. The first is bulk message delivery [7], which has been increasingly employed by commercial entities to reach mass markets. It may incur heavy network congestion and reduce the message delivery quality. The second is the topological structure of the social network formed by the SMS users, which can be exploited to allow fast spreading of security attacks such as viruses and spams. To the best of our knowledge, this is the first study on SMS reliability that is based on measurements from a real, operational cellular carrier.

The traces used in our study were collected from a nationwide cellular carrier that serves 20 million cell phone users. The collected data records over 59 million messages transmitted within a three-week period in 2005. A summary of the results obtained from analyzing the logs is as follows:

- *Baseline of reliability*: We quantify end-to-end message delivery reliability by estimating message failure ratio and latency. The results are somewhat unexpected: both metrics are no better (and in some cases much worse) than those of other popular communication means such as email, VoIP and legacy telephony. For example, the over-

all SMS delivery failure ratio amounts to 5.1%, compared with 1.57% for end-to-end message loss for emails [8]. Another interesting finding is that the delivery latency for 91% of short messages is less than 5 minutes, but only 50% of the heavy users ever experience latency below the 5 minute mark. Overall, SMS seems to achieve a level of reliability that is no better than other communication means. However, it does so in an adverse operational environment, in which intermittent connectivity of end-user devices and network resource shortages are more frequently encountered than in the wired Internet.

- *Flash-crowd events and bulk messaging*: We witness a flash crowd event during the New Years' Eve of 2005. By characterizing user and system behavior during the event, we observe that, similar to flash-crowd events that occur in the World Wide Web [6], the sharp increase of message traffic volume in the New Year's Eve is caused by a much larger number of active subscribers, rather than increased traffic rate by each individual user. As for messages that are sent to groups of users in a bulk fashion by content providers, the main finding is that their traffic rate is highly dynamic. Moreover, bulk messages from different content providers appear to be synchronized to a certain degree, which poses a severe, potential risk for network-wide congestion. Solutions to addressing network congestion caused by both factors are also discussed.
- *Social network of SMS users*: As in most networks formed through human activity and interactions, the topological structure of the SMS user graph exhibits a "small-world" phenomenon: few contacts are needed to reach any user in the network regardless of the starting point. Using simulations, we show how this property can affect the resiliency of the network, by examining how fast a virus may propagate in the real-world SMS user population.

The rest of the paper is organized as follows. Section II provides background introduction to SMS, the underlying network elements, and the traces used in our study. In Section III we present results on the baseline reliability of SMS. In the followup three sections, we discuss three factors that may reduce the reliability of SMS. Specifically, in Section IV we discuss a flash-crowd event. In Section V we study bulk message delivery, and in Section VI we show a small-world phenomenon observed in the social network formed by SMS users. In Section VII we compare and contrast this paper to related work. Finally we conclude in Section VIII.

II. Background and Traces

A. SMS network architecture

SMS messages are transmitted over the Common Channel Signaling System 7 (SS7). SS7 is a global standard that defines the procedures and protocols for exchanging information among network elements of wireline and wireless telephone carriers. These network elements use the SS7 standard to exchange control information for call setup, routing, mobility management, etc. Figure 1 shows the typical network architecture for SMS communication.

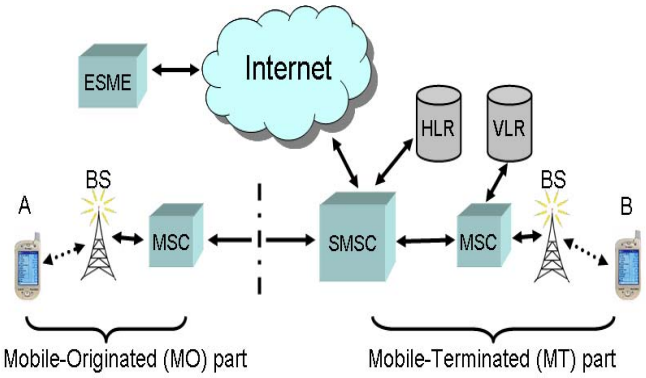


Fig. 1. Typical network architecture for SMS

Conceptually, the network architecture consists of two segments that are central to the SMS model of operation: the Mobile Originating (MO) part, which includes the mobile handset of the sender, a base station that provides the radio infrastructure for wireless communications, and the originating Mobile Switching Center (MSC) that routes and switches all traffic into and out of the cellular system on behalf of the sender. The other segment, the Mobile Terminating (MT) part, includes a base station and the terminating MSC for the receiver, as well as a centralized store-and-forward server known as SMS Center (SMSC). The SMSC is responsible for accepting and storing messages, retrieving account status, and forwarding messages to the intended recipients. It is assisted by two databases: the Home Location Registrar (HLR) and the Visitor Location Registrar (VLR). The two databases contain respectively permanent and temporary mobile subscriber information, e.g., the address of the MSC the device is associated with.

Though the Short Message Service has been popularized by the exchange of text messages among cell phone users, it has been increasingly used by businesses as a low-cost bearer to deliver various types of content such as ringtones, news, stock price, quizzes, and casting of votes. Such content providers, also known as External Short Message Entities (ESMEs), initiate or receive text messages through gateways which bridge the SMS interface to the Internet. We will analyze traffic of such messages in Section V.

For more details on the SMS network architecture and operations, we refer the interested readers to a tutorial [9] and specifications [10].

B. Traces

In 2004-2005, we obtained permission from a mobile operator to collect SMS traffic traces for a period of three-weeks. The operator is a nationwide cellular carrier with more than 20 million subscribers. The main set of traces were collected at the SMSC in the format of SMS Charging Data Records [11] (CDRs). Whenever the SMSC receives

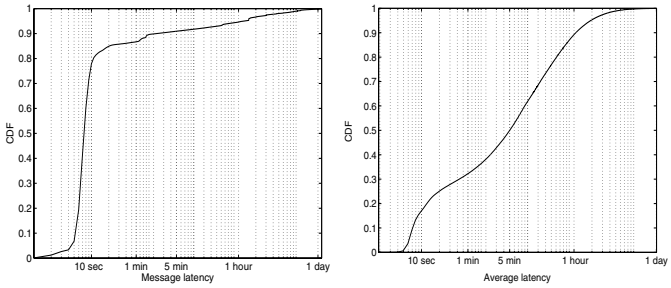
and forwards a message originated from a handset or ESME, two records are logged: one (MO-CDR) for the MO part of the communication (from the originator mobile handset to the SMSC), and another (MT-CDR) for the MT part (from the SMSC to the destination handset). Each record includes information such as: a timestamp at second-level granularity, the mobile identification and directory numbers for both the sender and receiver, number of retransmissions, address of the MSC that the sender (receiver) handset is associated with, message size, the result of the delivery attempt, and a reference value that can be used to associate the MO and MT records of one message.

The logs were stored in a proprietary, binary format used by the billing subsystem of the carrier, which we had to reverse-engineer in order to automate the retrieval of the ASCII values from the records. The automated tool that we developed for extracting values from the traces was developed iteratively in multiple steps, using a GUI-based utility from the network analyzer package of the carrier to guide the reverse-engineering effort. The steps taken involved the following: first, we identified the fields and their position in the records, by comparing a hexadecimal dump of a small set of records to the ASCII translation provided for that same set by the GUI tool. To make sure that we were able to identify all possible values that a field could have along with their meaning (its domain), we ran our converter/extractor software over the full set of more than 100 million records. For those values that were not in the records of the original set (and for the fields that took only a small finite number of distinct values), we used the GUI front-end to retrieve the remaining ASCII translations and complete our converter/extractor software.

The analysis presented in the rest of this paper is based on two sets of CDR logs collected at the carrier's premises. The main set used for the analysis (set-main) spans a period that starts at 15:01:02 on 04/04/2005 [1] and ends at 00:00:04 on 04/26/2005. Unfortunately, due to the collection process in the premises of the carrier, certain time periods are not logged, which amount in a total of 544,521 seconds ($\approx$ 6.3 days missing. After the binary-to-ASCII conversion we end up with 20Gb of data, containing 48,573,312 MO and 59,612,388 MT records. Based on the traces, 10,854,135 cellular users are observed to send 5.206G characters in text messages and receive 5.198G characters. A second set of CDR logs, which is shorter in duration, was collected around the New Year's Eve of 2005. The measurement duration started at 11:00:00 on 12/25/2004 and ended at 00:00:00 on 01/06/2005.

As a supplementary measurement set, we collected SS7 logs in the same time period as the CDR collection. The SS7 logs record the SS7 protocol level operations involved in text message delivery. Thus they expose more information than the CDR, e.g., message priority, root causes for delivery failure. Unfortunately, our collected SS7 logs are not complete as they only account for 10%-20% of all the SS7-level operations. Therefore, we limit the usage of SS7 logs to very few studies such as the analysis of message failure.

[1] All clock times used in this paper are in the local time zone of the country where the carrier provides the SMS service.



2.1: CDF for operator-perceived average message latency

2.2: CDF for user-perceived average message latency

Fig. 2. Message latency

III. A BASELINE FOR THE RELIABILITY OF SMS

SMS is commonly perceived as an instant, no-failure communication medium. Message failure ratio, defined as the percentage of messages that fail to reach their receivers, is expected to be close to zero, and message latency, measuring how long it takes a message to be received by its intended recipient, is anticipated to be low as well, usually in the order of seconds. However, our measurement results show that the above understanding is not completely accurate in reality. It turns out that the failure ratio and latency for short messages can be quite high and on par with other communication means such as legacy telephone service, email and VoIP.

A. Message delivery failure

As described in Section II-A, every message delivery consists of two phases, the Mobile Originating (MO) and the Mobile Terminating (MT) phase, each one of which accounts for one record in the CDR trace. A failed message delivery can be caused by a failure in either of these two phases. However, an unsuccessful delivery attempt in the MO phase is not shown in our traces because only those messages successfully received by the SMSC are recorded. Therefore, in this paper we only consider message delivery failures occurring in the MT phase, during which the SMSC tries to forward stored messages to their destination. Accordingly, the reported failure ratio is an underestimate of the actual value.

In the main set of CDR traces that were logged during an operating period with normal SMS traffic load, we observe a significant 5.1% of messages fail to reach their destination. In typical scenarios, a message delivery fails when the SMSC attempts multiple times to deliver the message until either the maximum allowable retry number is reached or the message expires. The break down is 3.5% for the former case and 1.6% for the latter for all the messages in our traces. Each time a SMSC attempts to deliver a message to the recipient, the attempt may fail due to a specific reason that can be further read from an error code field in the SS7 trace. These reasons are categorized in Table I. Note that the percentages reported

Failure reason	Percentage	Average retries	Average delivery latency (seconds)	Explanation of failure reason
SMS delivery postponed	86.4%	2.5	35.0	Caused by no page response, destination busy, memory full, destination out of service, etc.
Destination no longer at this address	12.8%	1.1	4.6	The mobile receiver is no longer at the temporary SMS routing address. The message sender should should not re-use the temporary SMS routing address.
Network resource shortage	0.4%	2.6	55.5	A required terminal resource (e.g., memory, etc.) is not available to process this message.
Other causes	Each accounts for less than 0.2% of failure instances			

TABLE I
SUMMARY OF REPORTED REASONS FOR MESSAGE DELIVERY FAILURE

in the table are relative to the number of messages logged in the SS7 trace and not those recorded in the CDR traces.

The failure ratio of the SMS service is worse than other conventional communication means. For example, the call-failure ratio for legacy telephone service is about 0.01% [12], while the same metric for VoIP stands at 0.9% [13]. It is difficult to come by measurement data on message loss of the email service. In a recent study on email dependability, Moors and Lang [8] measured the reliability levels of 16 popular, free email providers and found that end-to-end email loss depends on the provider: while some of them have nearly zero loss ratio for delivering emails, a few others exhibit a loss ratio as high as 10%. By averaging over all providers, they estimated the loss ratio to be 1.57%. Considering that our reported SMS failure ratio (5.1%) is a conservative estimate of the actual value, we conclude that SMS is no better than other conventional communication means in terms of delivery failure ratio.

B. Latency

We examine message latency as perceived both from the cellular operator and the users. While the "operator-perceived" latency is based on the whole set of collected messages, the "user-perceived" latency is on a per-user basis. Message latency is defined as the time that elapses between the reception of a short message by the SMSC from the originator user (found in the MO record) till the moment when it is actually delivered to the destination user (logged in the MT record). The latency in our context does not include the time needed for a message to reach the SMSC from the moment it is transmitted by the originator's device. This latency component, which also includes the air interface time of the sender, is relatively short and in the order of a few seconds according to [5].

Figure 2.1 plots the CDF for operator-perceived message latency by considering all successfully delivered messages. We observe that 91% of delivered messages have latency less than 5 minutes. Moors and Lang [8] reported that 92% of the emails sent in their experiment were delayed for less than 30 seconds, however the setup of their testing environment presented more favorable conditions than the typical operational reality for SMS: recipients' email clients were connected to the Internet with abundant bandwidth and without any interruption. Finally, although the majority of short messages have latency within several minutes, we do observe that a few messages experience extremely large delay with the maximum recorded value approaching 4 days and 5 hours!

User-perceived message latency is calculated only for those users that frequently use the service and send more than 15 text messages during the nearly 15 days logged in our trace, or one message per day on average. The CDF for such user-perceived latency is plotted in Figure 2.2, which shows that most users experience a relatively large average latency. More specifically, less than 50% of users have average latency less than 5 minutes. When we compare this result with the average latency as perceived by the operator, one realizes the difficulties in matching users' expectations with operational requirements when dealing with large user bases.

In summary, the SMS service exhibits a message failure ratio no better than other communication media such as telephony, VoIP and email. The average latency typically ranges from several seconds to one hour, which explains the characterization of the SMS service as an "almost-instant" communication medium. Having explored the basic level of reliability one might expect from the SMS service, we shift our focus to three specific factors that can significantly affect its reliability, namely flash-crowd events, bulk messaging, and topological structure of the social network formed by SMS users.

IV. FLASH CROWD EVENTS

In this section we report a flash crowd event observed from the CDR logs collected around the period of New Year's Eve (NYE) of 2005.

A. Characteristics of a flash crowd event

Figure 4 plots the message arrival rate (per min) for the period around the New Year's Eve of 2005. The figure shows a very sharp increase in SMS traffic, which occurs right before the turn of the year. The figure also shows several periods during which message volume is zero. As mentioned in Section II, this is artificial and due to the intermittent data collection process at the premises of the carrier, although the gap that appears immediately after 00:00:00 on 01/01/2005 is due to the measurement equipment been taken offline to reduce load in the cellular network. We further zoom in and

	Duration	Senders	Receivers	Message arrivals	Message type		
					Mobile-to-mobile	Mobile-to-ESME	ESME-to-mobile
Normal period	[12/28/13:40:00, 12/31/16:40:00]	1,477,331	1,707,895	16,953,324	86%	2%	12%
	3 days and 3 hours						
NYE	[12/31/16:40:00, 01/01/00:00:00]	911,472	2,757,274	5,486,572	95%	1%	4%
	7 hours and 20 minutes						

TABLE II
BASIC STATISTICS FOR A NORMAL TIME PERIOD AND FOR NYE

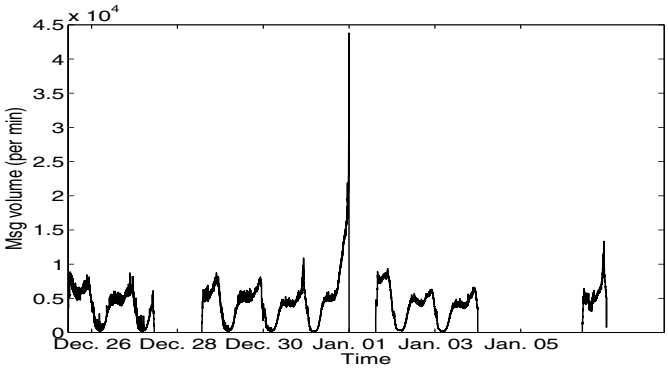
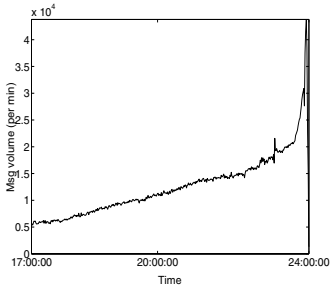
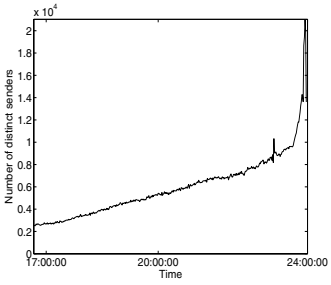


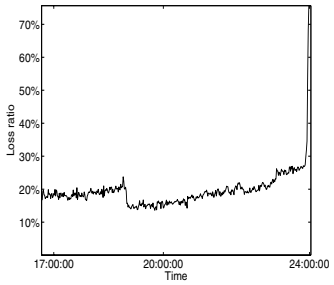
Fig. 4. Message arrival rate around New Years' Eve of 2005



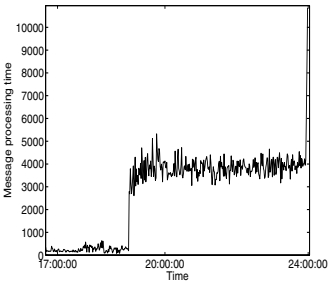
5.1: Message arrival rate



5.2: Arrival rate for distinct users



5.3: Message failure ratio



5.4: Message latency

Fig. 5. Statistics for traffic in NYE

plot in Figure 5.1 the message arrival rate for the last seven hours before the year turns to 2005. The figure clearly shows that the arrival rate is steadily increasing and reaches a peak of almost eightfold the normal traffic level as it gets close to mid-night.

This eightfold increase in traffic can introduce congestion, which in turn adversarily affects the reliability of the service as it has been experienced by ordinary users over the past years [14]. We seek to understand how user and traffic behavior are different between the NYE and other normal time periods. For this purpose, in Table II we compare basic statistics between NYE and the three days prior to NYE. Besides confirming a much higher message arrival rate in NYE, Table II shows that the number of senders and receivers in NYE are significantly higher than in a normal period with equal length. Furthermore, the last column of Table II indicates that the majority of the traffic (95%) in NYE are for mobile-to-mobile communications, which indicates that such a scenario mostly reflects cellphone users' behavior instead of content providers'. Therefore, the activity during the NYE period could be described as: the SMS human user population has a sharp increase, and most of these human users send text messages almost simultaneously. Such an activity is usually labeled as a *flash-crowd event*.

We further examine the behavior of senders during the NYE period. Figure 5.2 shows the arrival rate for distinct senders. By considering Figures 5.2 and 5.1 together, we observe that individual senders generate similar numbers of messages, i.e., 2-3 messages per sender on average. In other words, the distribution of messages among users is quite flat. Following this reasoning, it is clear that the exceedingly large message volume in NYE is caused by the sharp increase of the number of active users of the SMS service, instead of an increase in individual user's traffic. This observation in similar to what has been reported for flash-crowd events that stress Web servers in the Internet [6]. Another interesting finding is that 60% of the users appearing in NYE did not send any messages in the three days prior to NYE. Our conjecture is that these users are normally infrequent text message users and yet suddenly become active because of the special occasion.

Message failure ratio and latency in NYE are shown in Figures 5.3 and 5.4 respectively. Figure 5.3 shows that the message failure ratio is high but suddenly drops at 19:00:00, then it keeps steadily increasing. Accordingly, message latency grows from several minutes to about an hour around 19:00:00. We speculate that this phenomenon is relevant to the queuing policies employed by the cellular carrier, but this conjecture needs further verification from the operator.

B. Mitigation of the impact of flash-crowd events

The observed flash-crowd event can severely disrupt system operation by congesting the underlying signaling channels. This is because short messages are delivered over the signaling channels, as opposed to traffic channels that are dedicated to carrying voice traffic. The operator needs to ensure that text message traffic does not consume all the bandwidth of the signaling channels as this would prevent the setup of voice, data, and fax calls. The main approach that has been proposed to achieve this goal is dynamic channel allocation [15], which adopts two strategies to handle the problem. The first strategy temporarily converts a traffic channel into multiple signaling channels when the load due to short messages is high, thus increasing the signaling channel capacity on demand. The second strategy is called "immediate assignment to traffic channel". During peaks of SMS traffic load, whenever a voice call is initiated, the call setup phase is immediately diverted and handled over a traffic channel instead of a signaling channel. As a result, the precious bandwidth of the signaling channels is preserved for short message delivery.

V. Bulk messaging

In addition to mobile-to-mobile communication, SMS supports one-to-many message delivery, according to which External Short Message Entities (ESMEs) such as Internet content providers can send messages en masse to multiple mobile handsets. Since mobile network operators usually sell messages wholesale at a low price, Internet content providers and retailers use bulk messaging as a cost-effective way to reach a large number of their potential customers.

As it has been recently shown [5], bulk messaging can significantly affect the reliability of the cellular systems because of two reasons: first, while mobile-to-mobile messages are originally injected into the cellular network via the air interface, bulk messages are usually injected via wired links that connect the mobile carriers to the Internet. As a result of the high bandwidth of the wired links, the incoming rate for bulk messages can be very large compared with the limited bandwidth of the wireless interface. Second, certain messages in bulk are destined to many individual users simultaneously and may consume the bandwidth of the control channel, consequently affecting the operation of voice call services [7].

In our collected CDR traces, we identify more than one thousand ESMEs through their distinguishing 4-digit phone numbers. They account for about 20% of the overall traffic, showing that mobile-to-mobile communications are still the leading source of short messages. Nine of them are labeled as major providers considering that each of them contributes to more than 2% of the traffic from all ESMEs. We will focus on these nine providers whose services include ringtone downloading, media delivery, TV updates, headline news, etc.

An interesting observation is that traffic generated from these providers is more dynamic than the overall traffic. This can be seen by comparing the overall traffic from Figure 4 with Figure 6, which depicts the message volume for four major content providers. For the nine main content providers studied, the ratio between their peak traffic rate and their respective average rate ranges from 9.4 to 53.7. It is substantially higher than the overall traffic, for which the ratio is 3.7. Another observation on bulk messaging is that traffic generated by multiple ESMEs appears to be temporally synchronized to a fairly high degree. For example, as noted in Figure 6, the peak rate for the four providers occurs at the same time-of-day. To further quantify this observation, we represent each provider's message rate as a time series and calculate the correlation coefficient for every pair of providers. In total there are 36 pairs formed by these nine providers. One third of those have a correlation coefficient higher than 0.8, while 55.5% of them have a coefficent higher than 0.6. The above two observations, high dynamics of bulk message traffic and temporal synchronization of the sources, can cause congestion incidents in the signaling channels of GSM and CDMA networks, especially as SMS-based media and notification services are becoming increasingly popular among mobile users.

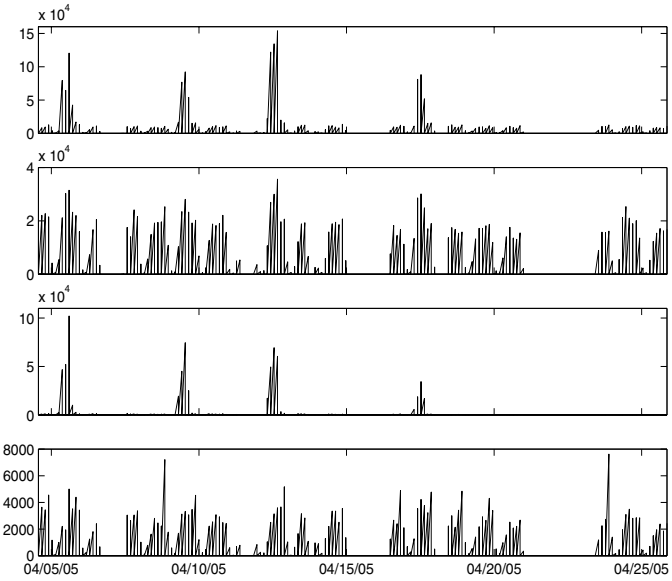


Fig. 6. Message volume for four content providers (msg/hr)

A. Strategies to prevent congestion caused by bulk messaging

A general principle for preventing congestion is to throttle traffic as close to the source as possible. Accordingly, rate regulation mechanisms close to (or on) the gateways to which those ESMEs are connected can be effective in preventing congestion incidents. In addition, we believe it is advisable to disrupt the synchronized traffic patterns of different content providers by enhancing cooperation among ESMEs, in order to have them schedule their peak rate hours at different times-of-day. Cellular operators can appropriately adjust their pricing policy to provide incentives to ESMEs for enabling such cooperation.

One existing approach for limiting congestion induced by bulk messages is *filtering* [16]: the SMSC intercepts all

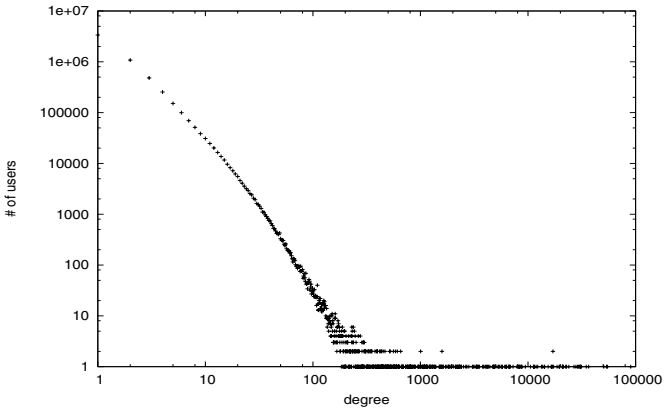


Fig. 7. Distribution of node degree in the SMS social network

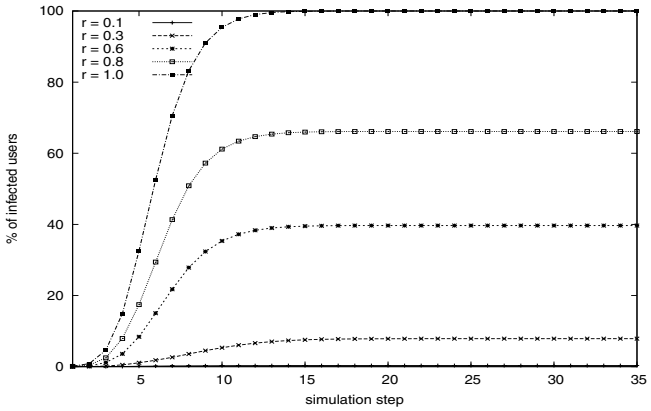


Fig. 8. Infection speed for different infection probabilities (r)

incoming messages and filters out bulk messages based on rules that specify quota per provider. Cell broadcasting [17] is another mechanism that could alleviate congestion. It has been widely deployed, though rarely enabled in cellular networks. By using cell broadcasting to deliver bulk messages, a few time slots in the broadcast channel are reserved in each cell. When bulk messages are destined to multiple recipients in the same cell, the messages will be sent in those reserved time slots instead of in their dedicated signaling channel.

VI. Social network of SMS users

As the SMS service (and its successor Multimedia Messaging Service –MMS) is becoming a very popular type of computer-mediated communication, viruses and malware that use SMS as transport for propagation have already started to appear [18] [19]. With the increased communication and computation capabilities of mobile handsets, one can only predict that in the near future malware that propagates over SMS/MMS will be even more commonly encountered. For this reason, it is important to study the structure of the social network formed by SMS users, and assess its resiliency to the propagation of virus, worms and other malicious software. Similar studies have been recently performed for other computer-mediated social networks such as those formed through the use of email (e.g. [20] [21]).

From the CDR traces we extract the social network formed by over 7.4 million SMS users. We represent its topology with an undirected graph: every node corresponds to a SMS user and two nodes are connected with an edge if one or more short messages have ever been exchanged between the users that correspond to the two nodes. It is worth noting that due to the construction methodology the graph is not a multigraph, neither directed. The degree of a node denotes the number of contacts the user has.

From this graph, we identify the largest connected subgraph, which consists of 5,795,574 nodes, or 78.3% of the total number of cell-phone users in the social network. Figure 7 shows the distribution of node degrees for all the SMS users in the subgraph. As expected, the vast majority of them have only 1 or 2 contacts.

We also observe very few users with an unusually high degree (more than 1000 contacts). Considering the traces only span three weeks, it seems that these nodes correspond to ESMEs, although in the construction of the graph we already removed all nodes whose phone number does not start with the digit that has been assigned to cellular users in the country where the cellular network operates. We then calculate the characteristic path length L_{actual} and the clustering coefficient C_{actual} [2] of this connected subgraph and compare them with the respective values L_{random} and C_{random} of a random graph with the same number of nodes and average node degree. It turns out that $L_{actual} = 3.61 \gtrsim L_{random} = 3.15$ and $C_{actual} = 0.62 \gg C_{random} = 0.0007$. The above inequalities and the heavy-tailed distribution of node degrees shown in Figure 7 provide strong indication that the social network formed by SMS users follows a "small-world" structure [22]. It is well known that a small-world graph possesses the commonly termed "six-degrees of separation" property, which implies that infectious viruses can spread very rapidly.

To investigate the effect of the specific topology to the propagation speed of a virus that uses SMS/MMS for spreading to other handsets, we run simulations on this connected subgraph using the simplified virus model of [22]. According to this model, all nodes are initially considered healthy, and at time $t = 0$ a randomly chosen node is marked as infected. Infected nodes are removed permanently after a period that lasts one unit of simulated time. During this time, each infected node tries to infect each of its healthy neighbors with infection probability r and the virus spreads in that fashion until it either infects the whole population of SMS users, or it dies out.

Figure 8 plots the fraction of infected users as a function of simulated time for different infection probabilities r. Results are averaged over 100 simulation runs. As we can see from the graph, the fraction of users that eventually obtain the virus

[2]Intuitively, the characteristic path length measures the typical separation between two vertices in the graph, whereas the clustering coefficient measures the cliquishness of a typical neighbourhood. The detailed definitions can be found in [22].

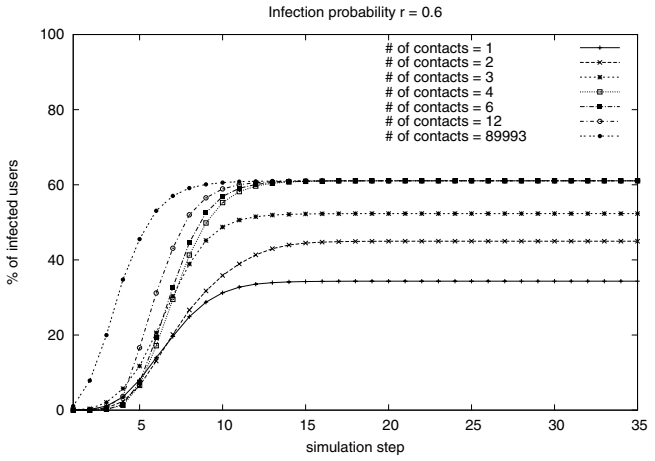


Fig. 9. Infection speed for multiple initially infected users

depends on the infection probability: the higher its value the more users will be reached by using the simplified virus model. One interesting observation though is that, regardless of the actual value or r, the number of users that eventually get infected reaches 90% of its maximum value in only 10 units of simulated time. When $r = 1.0$, which means that all neighbors of a node get infected and essentially the graph is traversed in a breadth-first fashion, this means that it takes propagation along a path of 10 hops to reach 90% of the whole population.

Figure 9 shows how quickly a virus propagates in the SMS network when the initially infected node has different degree, and the infection probability r is 0.6. Due to the "small-world" structure of the SMS social network, the choice of the node that first gets infected does not affect the speed of propagation considerably: even when infection starts from a node with a single contact, the time needed to reach the maximum number of users that get affected under the given infection probability is very close to the time needed when the propagation is initiated from the best connected user (having 89,993 contacts!). It's worth noting though that the choice of the initial node (in terms of the number of contacts it has) does affect the fraction of nodes that eventually will get infected: the total number of nodes the virus reaches is considerably less when the virus starts propagation from nodes with 1–3 contacts than from users with a higher node degree, as shown in Figure 9.

VII. Related Work

The role of SMS service as a popular communication medium used by hundreds of millions of people worldwide has sparked recent research interests in exploring the security and vulnerability aspects of the service. Enck *et al.* in [5] analyze the feasibility of launching a denial-of-service attack against SMS-capable cellular networks via Internet SMS-gateways. They follow a "gray-box" approach by probing the cellular network from end-user phone devices, in an attempt to infer the behavior of SMS delivery. In [23], Traynor *et al.* further analyze the feasibility of such denial-of-service attacks. They also discuss several techniques to mitigate the attacks through analysis and simulations. The author of [7] comments on the potential danger of bulk messaging, which is identified by us as a factor potentially affecting SMS reliability. Nevertheless, no quantitative analysis on the problem is provided in [7]. Simulated topologies are generated for the study in [4], which analyzes malware that affects mobile phones and proposes a simulation framework for studying virus propagation models in such environments. In [2], the authors perform a capacity analysis of the cellular network for supporting SMS delivery in critical scenarios. Their study is based on parameters specified in technical specifications of GSM networks.

In this paper instead, we follow a "white-box" approach by analyzing real SMS traces collected from a nationwide cellular carrier. Our presented results are based on large-scale real measurement and data. In contrast, most of the previously related work are based on idealized assumptions, or small-scale experiments from a handful of mobile devices. In a recent work [24], we presented preliminary measurement results on the same trace used in this paper. While the focus of [24] is statistics for SMS traffic, the current work is to identify factors affect the reliability of SMS. In addition, we used SS7 traces to analyze message failure ratio while [24] does not.

Several researchers [25] [26] [27] [28] [29] [30] study the increased load imposed on the signaling channel due to delivery of SMS/MMS messages along with buffer management techniques to alleviate the problem. Due to the closed nature of cellular networks and unavailability of real data, the authors in all of these studies are forced to make assumptions about traffic loads, service delivery times, message expiration times and other parameters that are used in their models. The measurement results presented in this paper can help evaluate their proposed solutions through more realistic simulations.

Several equipment vendors have started offering solutions for mitigating the increased load caused by large volume of short messages and improving the reliability of the service. Cisco [31] proposes a new network architecture that deviates from the traditional centralized model of the SMSC and enables edge nodes (such as MSCs) to make more intelligent routing decisions on message delivery. Intel [32] advocates the use of modular components to increase capacity of SMSC and other network elements. Huawei [33] offers a multi-level cache for their SMSC to handle excessive SMS traffic rate. While each of these solutions has its own merit, a combined approach would most likely bear the best results, given the multitude of factors that affect the reliability and QoS of SMS delivery, as we showed in the previous sections.

The analysis in Section VI on the effect of the topological structure of the social network of SMS users was inspired by several previous studies that were conducted for other networks (e.g. [22] [21] [20]). To the best of our knowledge, this is the first study of such virus propagation on a real graph of SMS users.

VIII. Conclusion

The Short Message Service has been the most popular and profitable wide-area wireless data service in recent years, registering hundreds of millions of subscribers worldwide. As it is being considered as a primary communication medium in emergency situations and other mission-critical application scenarios, an evaluation of the level of reliable and resilient delivery service that it can offer is critical for assessing its suitability for such purposes.

We conducted a trace-based study of the SMS service in a real, operational cellular network, and characterized a baseline for the reliability of SMS in terms of message failure ratio and latency. Overall, the baseline reliability of SMS service is no better (and in some cases worse) than that of other communication media such as email, traditional telephony and VoIP. However, one should interpret this result considering the fact that the SMS service is offered in a challenging operational environment, where disconnections due to end-user mobility, non-ubiquitous network coverage and resource limitations of devices are frequently encountered.

We further investigated three specific factors that could adversely affect SMS reliability. The first factor, flash-crowd events, can significantly disrupt the normal operations of the underlying cellular network as a large number of subscribers attempt to access the service simultaneously. The second factor is bulk messages that are delivered mostly from content providers to large groups of mobile users. Traffic originated from such sources is highly dynamic, and message arrivals from different providers appear to be synchronized. As applications based on bulk messaging become increasingly popular, cellular operators should take measures to protect their networks from congestion that might occur due to bulk messages. Last, the structure of the social network formed by SMS users is found to exhibit the "small-world" property. Through simulations we show that viruses and other malware can spread rapidly in such a network.

As a final note, we would like to stress that this study is based on measurements obtained from a single cellular operator. It would be useful to collect traces from other mobile carriers, to further test and validate the generality of the results presented in this work.

References

[1] Netsize S.A., "The Netsize guide 2005 Edition - The mobile is open for business," February 2005.

[2] Office of the Manager, "SMS over SS7," Tech. Rep. NCS TID 03-2, National communications system, December 2003.

[3] H. Rheingold, "SMS disaster warning system," December 2004.

[4] A. Bose and K. G. Shin, "On mobile viruses exploiting messaging and bluetooth services," in *Proc. of the 2nd International Conference on Security and Privacy in Communication Networks*, (Baltimore), 2006.

[5] W. Enck, P. Traynor, P. McDaniel, and T. L. Porta, "Exploiting open functionality in SMS-capable cellular networks," in *Proc. of the 12th ACM Conference on Computer and Communications Security (CCS)*, (New York, NY, USA), pp. 393–404, ACM Press, 2005.

[6] J. Jung, B. Krishnamurthy, and M. Rabinovich, "Flash crowds and denial of service attacks: Characterization and implications for CDNs and web sites," in *Proc. of World Wide Web (WWW)*, (Hawaii, USA), May 2002.

[7] M. Wood, "SMS bulk messaging, the problem and the solution," March 2004. http://www.ceasa-int.org/library/7_sms_mass_messaging_problems_V1-2.doc.

[8] T. Moors, "Email dependability, presentation at Email Management World conference," 2004. http://uluru.ee.unsw.edu.au/∼tim/dependable/email/emw_moors.pdf.

[9] G. Peersman, S. Cvetkovic, P. Griffiths, and H. Spear, "The global system for mobile communications short message service," *IEEE Personal Communications*, June 2000.

[10] Telecommunications Industry Association, "TIA/EIA-637-A, Short Message Service." December 1999.

[11] 3GPP, "TS 32.205, Charging Data Description for the Circuit Switched (CS) domain; release 5." v.5.6.0, March 2003.

[12] Olga Kharif, "VoIP providers: Heeding the call," November 2005. http://http://www.businessweek.com/technology/content/nov2005/tc20051128_964764.htm.

[13] Keynote Systems Inc., "VoIP Competitive Intelligence Study," 2006. http://www.keynote.com/company/press_room/releases_2006/01.25.06.html.

[14] The Times of India, "Mobile network chokes on New Year's Eve," January 2nd 2005.

[15] M. Schwartz, *Mobile Wireless Communications*. Cambridge University Press, ISBN 0-521-84347-2, December 2004.

[16] Cisco Systems, "SMS spam and fraud prevention," 2005. http://www.cisco.com/application/pdf/en/us/guest/netsol/ns278/c654/cdccont_0900aecd80250cb6.pdf.

[17] ETSI GTS GSM, "Short message service cell broadcast (SMSCB) support on the mobile radio interface ." 0.4.12-v5.0.0, July 1996.

[18] F-secure, "F-Secure virus information pages: Commwarrior." http://www.f-secure.com/v-descs/commwarrior.shtml.

[19] Kaspersky, "Timofonica virus: questions and answers." http://www.kaspersky.com/news?id=68.

[20] Y. Wang and C. Wang, "Modeling the effects of timing parameters on virus propagation," in *Proc. of the 2003 ACM workshop on Rapid malcode (WORM)*, (New York, NY, USA), pp. 61–66, ACM Press, 2003.

[21] C. Zou, D. Towsley, and W. Gong, "Email virus propagation modeling and analysis," in *TR-CSE-03-04*, Uni. of Massachusetts, 2004.

[22] D. Watts and S. Strogatz, "Collective dynamics in small-world networks," *Nature*, vol. 393, pp. 440–442, Jun 1998.

[23] P. Traynor, W. Enck, P. McDaniel, and T. L. Porta, "Mitigating attacks on open functionality in SMS-capable cellular networks," in *Proc. of the Twelfth Annual ACM International Conference on Mobile Computing and Networking (MOBICOM)*, (Los Angeles), September 2006.

[24] P. Zerfos, X. Meng, V. Samanta, and S. Lu, "A Study of the Short Message Service of a Nationwide Cellular Carrier," in *Proc. of ACM SIGCOMM Internet Measurement Conference (IMC)*, (Rio de Janeiro, Brazil), October 2006.

[25] Z. Naor, "An efficient short messages transmission in cellular networks," in *Proc. of 23th Annual IEEE Conference on Computer Communications (INFOCOM)*, (Hong Kong), March 2004.

[26] Y.-R. Haung and J.-M. Ho, "Overload control for short message transfer in GPRS/UMTS networks," *Inf. Sci. Inf. Comput. Sci.*, vol. 170, no. 2-4, pp. 235–249, 2005.

[27] M. Ghaderi and S. Keshav, "Multimedia messaging service: system description and performance analysis," in *Proc. of IEEE/ACM Wireless Internet Conference (WICON)*, (Budapest, Hungry), July 2005.

[28] W. Lin, Z. Liu, H. Stavropoulos, and C. H. Xia, "Hard deadline queueing system with application to unified messaging service," *ACM SIGMETRICS Performance Evaluation Review*, vol. 31, no. 2, pp. 31–33, 2003.

[29] Y.-R. Haung, "Determining the optimal buffer size for short message transfer in a heterogeneous GPRS/UMTS network," *IEEE Transactions on Vehicular Technology*, vol. 52, pp. 216–225, January 2003.

[30] M. Markou and C. Panayiotou, "Dynamic control and optimization of buffer size for short message transfer in GPRS/UMTS networks," in *Proc. of IEEE International Conference on Information & Communication Technologies: From Theory to Applications*, 2004.

[31] Cisco Systems, "A study in mobile messaging," Spring 2004. http://www.cisco.com/warp/public/cc/so/neso/mbwlso/mbmsg_wp.pdf.

[32] Intel, "SMS messaging in SS7 networks: optimizing revenue with modular components," 2003. ftp://download.intel.com/network/csp/pdf/8706wp.pdf.

[33] Huawei, "A prosperous future for mobile data service," 2005. http://www.huawei.com/publications/view.do?id=276&cid=94&pid=61.