Detection and Mitigation of Signaling Storms in Mobile Networks

Erol Gelenbe *Fellow*, Omer H. Abdelrahman, Gökçe Görbil
Intelligent Systems and Networks Group
Department of Electrical & Electronics Engineering
Imperial College, London SW72AZ, UK
{e.gelenbe,o.abd06,g.gorbil}@imperial.ac.uk

***Abstract*—Mobile Networks are subject to "signaling storms" launched by malware or apps, which overload the the bandwidth at the cell, the backbones signaling servers, and Cloud servers, and may also deplete the battery power of mobile devices. This paper reviews the subject and discusses a novel technique to detect and mitigate such signaling storms. Through a mathematical analysis we introduce a technique based on tracking time-out transitions in the signaling system that can substantially reduce both the number of attacked mobiles and the signaling overload in the backbone.**

***Index Terms*—Mobile Networks; Signaling Storms; Malware; Attack Mitigation**

I. Introduction

5G is challenged by broadband requirements such as video streaming and the Internet of Things that require low signaling overhead and quality of service (QoS) with higher traffic volume and bandwidths. However the mobile network control plane can be attacked by short and frequent communications that take advantage of vulnerabilities in signaling and simple applications such as paging [1], service requests [2] and radio resource control (RRC) [3], [4]. Such attacks can compromise a large number of mobile devices, or can target a list of mobiles by carefully timing the transmissions. They have been known to compromise connections for large sets of mobiles and there have been frequent industry reports about this matter [5]–[10]. Such attacks can also be the result of malfunctioning apps, and the outcome is to overload the individual mobiles, the wireless access networks, the signaling network and servers of the core network, and the Clouds which may be used in the process. Security is becoming ever more important, also due to the fact that critical applications [11]–[13] are transitioning to mobile devices supported by Clouds, including emergency management, Smart Electric Metering and Grid Control. signaling storms are mainly caused by misbehaving mobile apps that repeatedly establish and tear-down data connections [14] with a serious effect on the QoS of the network control plane [15], and many events have been reported to illustrate such attacks [5]–[8]. Similar events have also been observed for mobile devices that seek to connect to Cloud services [9], [16]. Thus significant efforts are required to be made to understand the security of mobile connections, making them resilient and reliable in the face of malicious apps [17].

II. Storms and signaling Overload

The Internet carries a lot of unwanted traffic [18], which includes remote DoS attacks, scanning worms, viruses [19] and spam. While mobile networks can be protected through middleboxes, a recent review [20], [21] found that 51% of carriers allow mobile devices to be probed from the Internet. Malware infections that target mobiles can generate excessive signaling, including premium SMS diallers, spammers, adware and bot-clients [10], and recent analysis of of mobile subscribers in China [22] indicates correlation between the frequency of resource-inefficient traffic and malicious activities. Thus signaling storms will continue to pose challenges with traffic growth [23] and IoT systems [24], [25].

A. The Mathematical Model

The analysis is conducted with a stochastic network model [26]–[30], and normal and attacked mobiles are represented by $s(t)$ at time $t \geq 0$:

$$s(t) = (b, B, C, A_1, a_1, \ \dots\ , A_i, a_i, \ ..; \ t) \tag{1}$$

where:

- b is the number of mobiles which are just starting their communication in low bandwidth mode,
- B is the number of unattacked mobiles which are in high bandwidth mode,
- C is the number of unattacked mobiles that have started to transfer or receive data or voice in high bandwidth mode,
- A_i is the number of attacked mobiles which are in high bandwidth mode and have undergone a time-out for $i-1$ times,
- a_i is the number of attacked mobiles which have entered low bandwidth mode from high bandwidth mode after i time-outs,

With Poisson arrivals at rate λ for new "calls" first admitted in state b, then requesting high bandwidth at rate r, so that with probability $1-\alpha$ the call is "normal" then entering state B. With probability α it will be an attacked call, requesting high bandwidth and enter stating A_1. Once a call enters state A_1, it will time-out after a time of average value τ^{-1}. Note that the time-out is a parameter that is set by the operator, and in practice it is of the order of a few seconds. After entering state A_1, if the mobile device or operator is very "clever", the call may be detected as being anomalous, and removed at rate β_1, which reflects the speed with which the is eliminated from further activity. Such a facility for blocking malicious calls is not currently available. However, typically the call will manage to request high bandwidth and enter state A_2 at rate r_1.

Proceeding in the same manner, in state A_i the anomalous call will again not start a normal communication, so it will eventually time-out after an average time τ_i^{-1} and enter state a_i:

$$\begin{aligned} \Lambda_{A_1} &= \alpha\Lambda_b, \\ \Lambda_{a_i} &= \Lambda_{A_i}, \\ \Lambda_{A_{i+1}} &= \Lambda_{a_i}\frac{r_i}{r_i+\beta_i}, \\ &= \alpha\Lambda_b\prod_{l=1}^{i} f_l, \text{ where} \\ f_l &= \frac{r_l}{r_l+\beta_l}, \end{aligned} \tag{2}$$

and Λ_b is the rate at which calls enter state b, which will be determined below from a more detailed analysis. Different calls will interfere each other via (a) the access to limited wireless bandwidth, and (b) possible congestion due to signaling and traffic in the backbone. However if we neglect these points, calls act independently of each other so that the average number of calls in each of the "attacked" states, that are denoted by a_i and A_i, is the average arrival rate of calls into the state, multiplied by the average time spent by a call in that state, so that we have:

$$\begin{aligned} N_{A_1} &= \frac{\alpha\Lambda_b}{\tau_1}, \\ N_{A_i} &= \frac{\alpha\Lambda_b}{\tau_i}\prod_{l=1}^{i-1} f_l, \; i>1, \\ N_{a_i} &= \frac{\alpha\Lambda_b}{r_i+\beta_i}\prod_{l=1}^{i-1} f_l, \; i>1. \end{aligned} \tag{3}$$

As a consequence, the total average number of attacked calls becomes:

$$\begin{aligned} N_a &= \sum_{i=1}^{\infty}[N_{a_i}+N_{A_i}], \\ &= \alpha\Lambda_b\sum_{i=1}^{\infty}\{[\prod_{l=1}^{i-1} f_l][\frac{1}{\tau_i}+\frac{1}{r_i+\beta_i}]\}. \end{aligned} \tag{4}$$

The rate parameters r_i are actually congestion dependent since a mobile can only access bandwidth if enough bandwidth for a reasonable level of QoS is available, and if interference will not be excessive. Let W denote the bandwidth that is available in a given cell through the effect of one or more base stations. If we call N_i the *average number* of mobiles that are in state $i \in \{b, B, C, a, A\}$, then the bandwidth availability will depend essentially on N_B, N_C, N_{A_i} because for a total amount of bandwidth in the system at a base station level of say W, the total amount of *available bandwidth* may be expressed as some value $W^* = W - w_1(N_B + N_C + \sum_i N_{A_i}) - w_2(N_b + \sum_i N_{a_i})$ where w_1 and w_2 denote the bandwidth allocated per high and low bandwidth request, respectively. Thus in reality the rates r_i will be "slowed down" as W^* becomes smaller since requests will be delayed or will even remain unsatisfied. The matter is of course more complex, because not only the bandwidth allocation itself but the error probabilities in the channel will be affected by the number of mobiles that are actually communicating via the base stations.

Now with regard to normal or un-attacked calls, once a call requests high bandwidth and enters state B, it will start communicating and this will be expressed as a transition rate κ which takes the call into "communication state" C. From C the call's activity may be interrupted, as when a mobile device stops sending or receiving data to/from a web site, or when a voice call has a silent period, in which case the call will return to state B at rate μ. Similarly, the call may end at rate δ, leaving the system.

From B it may either return to C at rate κ signifying that transmission or reception has started once again, or it may time-out at rate τ and return to state b. Once it returns to state b after a time-out, the call can try again to enter state B or state A as a normal or attacked call, since we have to include the fact that a normal call may become an attacked call after acquiring malware during its "normal" communication with a web site or with another mobile. As a consequence, we can calculate the rates at which the calls enter these normal operating states become:

$$\begin{aligned} \Lambda_b &= \lambda + \frac{\tau}{\tau+\kappa}\Lambda_B, \\ \Lambda_B &= (1-\alpha)\Lambda_b + \frac{\mu}{\mu+\delta}\Lambda_C, \\ \Lambda_C &= \frac{\kappa}{\kappa+\tau}\Lambda_B, \end{aligned} \tag{5}$$

which yields:

$$\begin{aligned} \Lambda_B &= \gamma\Lambda_b, \text{ where} \\ \gamma &= \frac{1-\alpha}{1-\frac{\mu\kappa}{(\mu+\delta)(\kappa+\tau)}}, \text{ and} \\ \frac{\tau}{\tau+\kappa}\gamma &= \frac{\tau(1-\alpha)}{\tau+\kappa-\frac{\mu\kappa}{\mu+\delta}}, \\ \Lambda_b &= \frac{\lambda}{1-\frac{\tau}{\tau+\kappa}\gamma}, \text{ so} \\ &= \frac{\lambda}{1-\frac{\tau(1-\alpha)}{\tau+\kappa-\frac{\mu\kappa}{\mu+\delta}}}, \\ \Lambda_B &= \frac{\lambda\gamma}{1-\frac{\tau}{\tau+\kappa}\gamma}, \\ \Lambda_C &= \frac{\kappa\lambda\gamma}{\kappa+\tau(1-\gamma)}. \end{aligned} \tag{6}$$

III. The Time-Out

The expression for N_a in (4) provides us with insight into how the time-out may be used to mitigate attacks. Indeed, combining the expression for N_a with Λ_b given in (6) when

τ_i, r_i and β_i do not depend on i, we have:

$$\begin{aligned} N_a &= \alpha\lambda \frac{r+\beta+\tau}{\beta\tau[1-\frac{\tau(1-\alpha)}{\tau+\kappa-\frac{\mu\kappa}{\mu+\delta}}]}, \\ &= \alpha\lambda \frac{r+\beta+\tau}{\beta\tau[1-\frac{\tau(1-\alpha)}{\tau+\frac{\delta\kappa}{\mu+\delta}}]}, \ or \\ &= \alpha\lambda \frac{(r+\beta+\tau)(\tau+\frac{\delta\kappa}{\mu+\delta})}{\beta\tau(\alpha\tau+\frac{\delta\kappa}{\mu+\delta})}. \end{aligned} \tag{7}$$

so that we may study how N_a varies with τ. In particular, we easily see that:

- As $\tau \rightarrow 0$, the effect the time-out is removed since it is infinite, and we have $N_a \rightarrow +\infty$ which indicates that the number of attacked mobiles will grow indefinitely because a finite time-out helps to identify and eliminate the attacked mobile devices.
- If $\tau \rightarrow +\infty$ the time-out is very fast and $N_a \rightarrow \frac{\lambda}{\beta}$. Note that λ which is the rate of incoming calls may be quite high in the thousands of calls per minute, while β^{-1} is the average time it takes to decide that a given mobile has been attacked, and may take minutes. As a result, their product $\lambda\beta^{-1}$ may also be quite high.

Thus it will be better to choose an optimum value of τ between these two extremes, which helps to minimise the total number of attacked mobiles N_a. When we take the derivative of (8) we remain with a second degree equation in τ, the solution of which yields:

$$1/\tau^*_{N_a} = \begin{cases} \frac{\sqrt{(1-\alpha)[\frac{\kappa\delta}{(\mu+\delta)(\beta+r)}-\alpha]}-\alpha}{\frac{\kappa\delta}{\mu+\delta}}, & \text{if } \frac{\kappa\delta}{(\mu+\delta)(\beta+r)} > \frac{\alpha}{1-\alpha}, \\ 0, & \text{otherwise.} \end{cases} \tag{8}$$

A simple order of magnitude estimate will tell us that $\delta << \mu$ since a complete call will typically be much longer than the time between successive accesses to a web site, or "silent" periods within an interaction from a mobile device can be numerous but short in comparison with the duration of the call as a whole. Similarly, we can assume that $r >> \beta$ since the time it will take to identify and eliminate an attacked call will be much longer than the time needed to request high bandwidth once the call is initiated. Finally, κ may be of the same order of magnitude to r or much smaller, because the transmission times that are represented by κ^{-1} are very short if the device is downloading or uploading bursts of data, but may be much longer (i.e. κ much smaller) if the mobile device is downloading video streams. Thus we can expect that in practice the condition

$$\frac{\kappa\delta}{(\mu+\delta)(\beta+r)} > \frac{\alpha}{1-\alpha},$$

that guarantees the existence of a non-zero minimum value is only satisfied for quite a small value of the attack probability α.

IV. Mitigation

Choosing a relatively small value of the time-out of the order of a few seconds is useful, but an additional mechanism is needed to mitigate the effect of storms. Therefore we suggest that a counter value n be selected so that as long as the number of *successive times* that the mobile uses the time-out is *less than* n, then the mobile remains attached to the network. However as soon as this number reaches n, then the mobile is detached after a time of average value β^{-1}. Thus β^{-1} can be viewed as the decision time plus the physical detachment time that is needed. Based on this principle, and with reference to our earlier definition of β_i, we have:

$$\beta_i = \begin{cases} 0, & 1 \leq i < n, \\ \beta, & i \geq n \end{cases}$$

so mitigation is activated when high bandwidth is requested n *successive* times, each followed by a time-out. Using the previous analysis, the number of average number of attacked calls becomes:

$$N_a = \alpha\Lambda_b[(n-1+\frac{r}{\beta})(\frac{1}{\tau}+\frac{1}{r})+\frac{1}{\tau}] \tag{9}$$

while the resulting signaling rate from the attack is:

$$\Lambda_a = \alpha\Lambda_b + \sum_{i=1}^{\infty}[\Lambda_{a_i}+\Lambda_{A_i}] = \alpha\Lambda_b[2n+1+\frac{2r}{\beta}] \tag{10}$$

A large value of n will improve the chances of correctly detecting a misbehaving mobile user, providing mitigation with full confidence to detach the misbehaving mobile from the network. If n is small we may have false positives, requiring analysis of the user's behaviour with other ongoing connections, or checking some data plane attributes such as destination IP addresses or port numbers that may be associated with malicious activities. Thus the higher the n, the faster the decision can be to disconnect the mobile, i.e. β increases with the threshold n, with a slope or derivative with respect to n expressed as $\beta^{'}$. W can show that the value n^* that minimises *both* N_a and Λ_a, satisfies:

$$\big(\beta(n^*)\big)^2 \approx r.\beta^{'}(n^*). \tag{11}$$

Figure 1 shows N_a and Λ_a versus n when $\beta(n) = 0.02n$ with $r = 0.5\ secs^{-1}$, and we see that $n^* = 5$ as predicted by (11).

A. Simulation

In this section, the joint detection and mitigation approach is evaluated using a mobile network simulator [31], representing both normal and attack events as in [32]. In Figure 2 the signaling misbehaviour increases from instants 2800 to 4000 secs after the simulation begins, and the mitigation scheme is activated at 7000 secs. Figure 2(a) shows the number of signaling messages sent and received by the RNC, while the response time for the application at a normal mobile is shown in Figure 2(b).

V. Conclusions

The recent growth in mobile data traffic is marked by an even greater surge in signaling loads due to interactions that include devices, apps, network configuration, cloud services and users. As mobile devices and apps increasingly access the Cloud in order to offload computationally intensive or

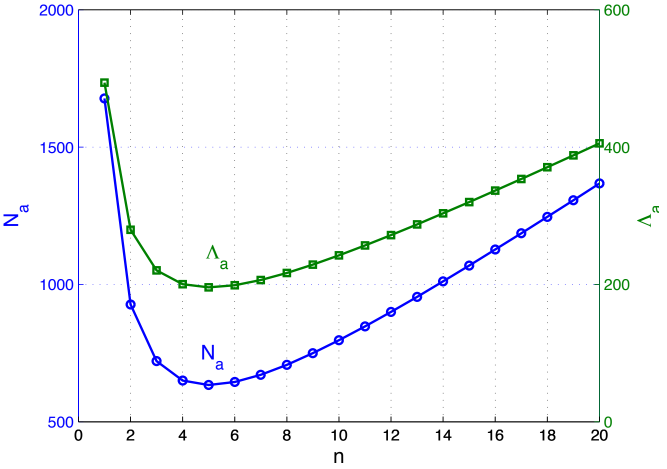

Fig. 1. Number of attacked mobiles (left) and resulting signaling overload (right) versus the number of false transitions that triggers the mitigation mechanism, when $\lambda = 10\ calls/s$, $\tau^{-1} = 5s$, $\alpha = 0.1$, $r^{-1} = 2s$, $\kappa^{-1} = 10s$, $\delta^{-1} = 5$mins, and $\mu^{-1} = 5s$. We observe a clear value of n which is optimum.

energy-costly activities, signaling storms can create heavy overloads that can significantly impair system performance and offer very poor quality of service to users. In addition, future machine to machine applications may be significantly impaired by such attacks, while Cloud services that are used by mobile devices can also become overloaded. We therefore propose a detection and mitigation technique for storms that uses a software counter for each mobile user, within mobile devices or in signaling system. When a maximum threshold of transitions to high bandwidth requests are detected which time-out because they do not make data or voice transfers, the mobile device is temporarily suspended, also limiting energy consumption in the network [30] and mobile battery depletion, and protecting against useless consumption of paid services. In future work, other techniques for dealing with detection and mitigation are worth investigation, such as collecting successive high bandwidth requests and serialising them [33] in the backbone network, and dropping them when they are so numerous that an indication of an attack can be suspected.

Acknowledgements

This research has benefitted from the EU FP7 NEMESYS Project "Enhanced Network Security for Seamless Service Provisioning in the Smart Mobile Ecosystem", grant agreement No. 317888 to Imperial College, $FP7-ICT-2011.1.4$ Trustworthy ICT.

References

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

[2] P. Traynor, M. Lin, M. Ongtang, V. Rao, T. Jaeger, P. McDaniel, and T. La Porta, "On cellular botnets: measuring the impact of malicious devices on a cellular network core," in *Proceedings of the 16th ACM conference on Computer and communications security (CCS '09)*. Chicago, Illinois, USA: ACM, 2009, pp. 223–234.

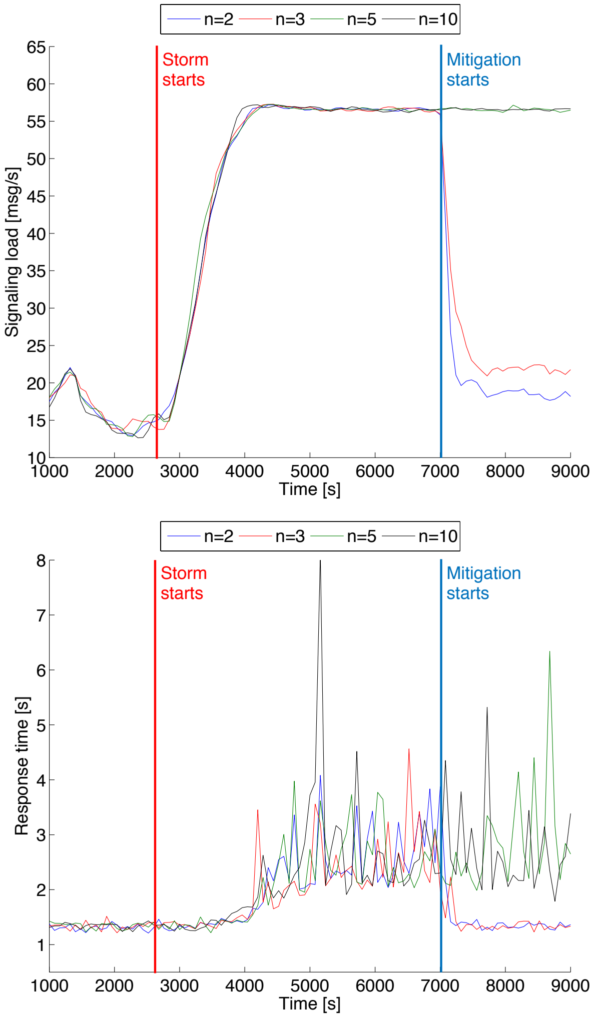

Fig. 2. Simulation results: time variation of (a) the total signaling load at the RNC, and (b) the observed response time for applications. In order to show how the storm builds up, the storm is launched (as shown) but the mitigation mechanism is initially disabled. At 7000 seconds, the mitigation mechanism is enabled, and the mitigation counter (the number of transitions that trigger the mitigation) is set. Rapidly after the mitigation mechanism is enabled, *if the mitigation counter is set to the optimum values of* 2 *or* 3, the signaling load and the application response times drop to their values before the attack began. However if the mitigation counter is set to higher values, its usefulness is not apparent.

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

[4] O. H. Abdelrahman and E. Gelenbe, "Signalling storms in 3G mobile networks," in *Proc. IEEE International Conference on Communications (ICC'14)*. Sydney, Australia: IEEE, June 2014, pp. 1023–1028.

[5] S. Corner, "Angry birds + android + ads = network overload," Jun 2011. [Online]. Available: http://goo.gl/2dSf9F

[6] M. Donegan, "Operators urge action against chatty apps," Light Reading Report, Jun 2011. [Online]. Available: http://goo.gl/vjLf1T

[7] Rethink Wireless, "DoCoMo demands Google's help with signalling storm," Jan 2012. [Online]. Available: http://goo.gl/pQjsAm

[8] Arbor Networks, "Worldwide infrastructure security report," 2012. [Online]. Available: http://goo.gl/2GZpP

[9] G. Reddig, "OTT service blackouts trigger signaling overload in mobile networks," Sep 2013. [Online]. Available: http://goo.gl/tJDx9p

[10] D. Maslennikov, "Mobile malware evolution: Part 6," Kaspersky Lab, Tech. Rep., Feb 2013. [Online]. Available: http://goo.gl/rXJ8J

[11] E. Gelenbe and Y. Cao, "Autonomous search for mines," *European Journal of Operational Research*, vol. 108, no. 2, pp. 319–333, 1998.

[12] E. Gelenbe, K. Hussain, and V. Kaptan, "Simulating autonomous agents in augmented reality," *Journal of Systems and Software*, vol. 74, no. 3, pp. 255–268, 2005.

[13] E. Gelenbe, "Search in unknown random environments," *Physical Review E*, vol. 82, no. 6, 2010.

[14] NSN Smart Labs, "Understanding smartphone behavior in the network," White paper, Jan 2011. [Online]. Available: http://goo.gl/jMtXu

[15] C. Schwartz, T. Hoßfeld, F. Lehrieder, and P. Tran-Gia, "Angry apps: The impact of network timer selection on power consumption, signalling load, and web QoE," *Journal of Computer Networks and Communications*, vol. 2013, no. 176217, 2013.

[16] Y. Choi, C. hyun Yoon, Y. sik Kim, S. W. Heo, and J. Silvester, "The impact of application signaling traffic on public land mobile networks," *IEEE Commun. Mag.*, vol. 52, no. 1, pp. 166–172, Jan 2014.

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

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

[19] E. Gelenbe, "Dealing with software viruses: a biological paradigm," *Information Security Technical Report*, vol. 12, no. 4, pp. 242–250, 2007.

[20] Z. Wang, Z. Qian, Q. Xu, Z. Mao, and M. Zhang, "An untold story of middleboxes in cellular networks," in *Proc. ACM SIGCOMM*, Toronto, Canada, Aug 2011, pp. 374–385.

[21] Z. Qian, Z. Wang, Q. Xu, Z. M. Mao, M. Zhang, and Y.-M. Wang, "You can run, but you can't hide: Exposing network location for targeted DoS attacks in cellular networks," in *Proc. Network and Distributed System Security Symp. (NDSS'12)*, San Diego, CA, Feb 2012, pp. 1–16.

[22] J. Li, W. Pei, and Z. Cao, "Characterizing high-frequency subscriber sessions in cellular data networks," in *Proc. IFIP Networking Conf.*, Brooklyn, NY, May 2013, pp. 1–9.

[23] Cisco, "Cisco visual networking index: Forecast and methodology, 2013-2018," White Paper, Jun 2014. [Online]. Available: http://goo.gl/xoBrTA

[24] M. Z. Shafiq, L. Ji, A. X. Liu, J. Pang, and J. Wang, "A first look at cellular machine-to-machine traffic: Large scale measurement and characterization," *SIGMETRICS Perform. Eval. Rev.*, vol. 40, no. 1, pp. 65–76, Jun 2012.

[25] 3GPP, "Machine-type and other mobile data applications communications enhancements (release 12)," 3GPP Mobile Competence Centre c/o ETSI, 650 route des Lucioles 06921 Sophia-Antipolis, FRANCE, TR 23.887, Dec 2013.

[26] E. Gelenbe, "Probabilistic models of computer systems ii," *Acta Informatica*, vol. 12, no. 4, pp. 285–303, 1979.

[27] E. Gelenbe and J.-M. Fourneau, "Random neural networks with multiple classes of signals," *Neural Computtaion*, vol. 11, no. 4, pp. 953–963, May 1999.

[28] E. Gelenbe, "The first decade of g-networks," *European Journal of Operational Research*, vol. 126, no. 2, pp. 231–232, October 2000.

[29] E. Gelenbe and S. Timotheou, "Random neural networks with synchronized interactions," *Neural Computation*, vol. 20, no. 9, pp. 2308–2324, 2008. [Online]. Available: http://dx.doi.org/10.1162/neco.2008.04-07-509

[30] E. Gelenbe and C. Morfopoulou, "A framework for energy-aware routing in packet networks," *Comput. J.*, vol. 54, no. 6, pp. 850–859, 2011. [Online]. Available: http://dx.doi.org/10.1093/comjnl/bxq092

[31] G. Gorbil, O. H. Abdelrahman, and E. Gelenbe, "Storms in mobile networks," in *Proc. 9th ACM Symposium on QoS and Security for Wireless and Mobile Networks (Q2SWinet'14)*, Montreal, Canada, Sep. 2014, pp. 119–126. [Online]. Available: http://dx.doi.org/10.1145/2642687.2642688

[32] G. Gorbil, O. H. Abdelrahman, M. Pavloski, and E. Gelenbe, "Modeling and analysis of rrc-based signalling storms in 3g networks," *IEEE Transactions on Emerging Topics in Computation, accepted*, 2014.

[33] F. Baccelli, E. Gelenbe, and B. Plateau, "An end to end approach to the resequencing problem," *Journal of the ACM*, vol. 31, no. 3, pp. 474–485, 1984.