Cyber Security via Signaling Games: Toward a Science of Cyber Security

William Casey[1,2], Jose A. Morales[1,2], Thomson Nguyen[1,2], Jonathan Spring[1,2], Rhiannon Weaver[1,2], Evan Wright[1,2], Leigh Metcalf[3], and Bud Mishra[1,2]

[1] Courant Institute, NYU, New York
[2] Software Engineering Institute, CMU, Pittsburgh
[3] Software Engineering Institute, CMU, United States
{mishra,tvn210}@nyu.edu,
{wcasey,jamorales,jspring,rweaver,ewright}@cert.sei.edu,
lbmetcalf@cert.org
http://cs.nyu.edu/mishra/

Abstract. In March of 2013, what started as a minor dispute between Spamhaus and Cyberbunker quickly escalated to a distributed denial of service (DDoS) attack that was so massive, it was claimed to have slowed internet speeds around the globe. The attack clogged servers with dummy internet traffic at a rate of about 300 gigabits per second. By comparison, the largest observed DDoS attacks typically against banks had thus far registered only 50 gigabits per second. The record breaking Spamhaus/Cyberbunker conflict arose 13 years after the publication of best practices on preventing DDoS attacks, and it was not an isolated event.

Recently, NYU's Courant Institute and Carnegie Mellon Software Engineering Institute have collaboratively devised a game-theoretic approaches to address various cyber security problems involving exchange of information (asymmetrically). This research aims to discover and understand complex structures of malicious use cases within the context of secure systems with the goal of developing an incentives-based measurement system that ensures a high level of resilience to attack.

1 Introduction

In the 2010 JASON report [Mitre, 2010], the authors wrote "The need to secure computational infrastructure has become significant in all areas including those of relevance to the DoD and the intelligence community. Owing to the level of interconnection and interdependency of modern computing systems, the possibility exists that critical functions can be seriously degraded by exploiting security flaws." However, they also lamented, "While the level of effort expended in securing networks and computers is significant, current approaches in this area overly rely on empiricism and are viewed to have had only limited success." The following rationale was offered: "*The challenge in defining a science of cyber-security derives from the peculiar aspects of the field. The "universe" of cyber-security is an artificially constructed environment that is only weakly tied to the physical universe*."

R. Natarajan (Ed.): ICDCIT 2014, LNCS 8337, pp. 34–42, 2014.

Thus the difficulty in developing a *science of cyber security* (SCS) is thought to stem from its inherent Manicheanness [Mitre, 2010], where the adversary is strategic and utilitarian as opposed to being oblivious and stochastic (i.e. Augustine). However, it must also be noted that a significant fragment of a science of cyber security (SCS) has to be built upon a complex computational infrastructure that is amenable to reasoning and re-engineering based on logical models such as Kripke structures. Thus, it appears that a successful approach to the cyber security problem may come from an amalgamation of a dualistic approach, which are partly based on techniques from game theory (inspired and validated with the tools of systems biology, e.g. analysis of immune systems) and partly based on model building (e.g., machine learning and statistical inference) and model checking. In light of this discussion, it may be worth re-examining the strategic choices that entities such as SPAMHAUS and CYBERBUNKER made [Williams, 2013,Gallagher, 2013,Lee, 2013,Schwartz, 2013], despite the obvious fact that both parties must have been well-informed about the accepted norms and best practices that were incorporated in the hardware, software and protocol architectures; divorced from a model of the humans and the utilities they wished to derive from their strategic choices, the protocols, practices and norms [Saint-Andre, 2009] achieved precious little.

We propose a novel approach, in which we model cyber security in terms of classical Information-Asymmetry Games (also called Signaling Games) [Casey, 2013], where the players (i.e., agents) assume either a role of a sender (S) or that of a receiver (T). The sender has a certain type, t, for instance: *beneficent* (C for cooperator) or *malicious* (D for defector), which could be assumed to be given by nature. The sender observes his own type while the receiver does not know the type of the sender. Based on his knowledge of his own type, the sender chooses to send a message from a set of possible messages $M = \{m_1, m_2, m_3, \ldots, m_j\}$; these messages are allowed to be complex: for instance, an offer of a mobile app with certain advertised utility and a price. The receiver observes the message but not the type of the sender or the ability to fully verify the message. Then the receiver chooses an action from a set of feasible actions $A = \{a_1, a_2, a_3, \ldots, a_k\}$; the receiver may be *oblivious/trusting* (C for cooperator) or *vigilant/mistrustful* (D for defector) – for instance, the offer of a mobile app may be ignored, accepted, verified or rejected (with a possibility of a reputation-labeling of the app, the sender or the app-store, etc.). The two players receive payoffs dependent on the sender's type, the message chosen by the sender and the action chosen by the receiver. Examples of various modes of attacks and how they map to such abstract games will appear in the full paper. In this paper, we focus only on a simple model of transaction involving transfer of an app from a sender (an app store) to a receiver (an app user).

Because of the informational asymmetry, it is possible for a sender to be *deceptive*, as is often the case in the cyber context. Traditional techniques such as making the signaling somewhat "costly" for the sender can help, but must be engineered carefully, since otherwise the very information-sharing capabilities of the cyber system can be seriously compromised. There

have been proposals for new internet architecture, new internet protocols and "bandwidth-as-price" mechanisms [See [Walfish *et al.*, 2010], [Yau *et al.*, 2005], [Beitollahi and Deconinck, 2012], [Lee *et al.*, 2007], [Doron and Wool, 2011], [Fu *et al.*, 2011], [Kargl *et al.*, 2001], [Xie and Yu, 2009], [Bhatia *et al.*, 2012], and [Huang *et al.*, 2007]], but any such approach can burden the normal transactions with an unwelcome and unacceptably heavy overhead.

We, instead propose a system based on an explicit pricing, using *M-coins*[1]. The other key ingredient is based on mechanisms for credible deterrence. However, the focus of this paper will be on two topics: (1) a simplified model for a repeated game that results from our analysis and (2) the empirical results obtained from an agent based simulation.

2 The Game Theoretic Models

Below (in Table 1) we describe a parameterized payoff matrix associated with a single transaction, where a sender may act in the "*cooperate*" behavior mode by sending a useful app honestly or the "*defect*" behavior mode by sending a malicious app deceptively, and where a receiver may act in the "*cooperate*" behavior mode by accepting trusted or the "*defect*" behavior mode by responding with a challenge. The payoff-parameters in the table are as follows: $a =$ the *cost of app*, $b =$ the *value of app*, $c =$ the *cost of verification*, $d =$ the *benefit of hack*, $e =$ the *cost of getting caught*, $f =$ the *benefit of catching malicious user*, and $g =$ the *cost of challenging a sender*.

Table 1. Row player is the sender, column player is the receiver

Sender,Receiver	receive trusted	receive challenge
send clean	$(a, -a+b)$	$(a-c, -a-g)$
send malware	$(a+d, -a-d)$	$(a-c-e, -a+f-g)$

Table 2 simplifies the payoff matrix for the joint strategy considering both roles of sending and receiving per user in repetition of a single transaction:

3 The Results from Simulation

To examine the details of the potential dynamics of the resulting repeated game, we consider a reproducing population model where reproduction of a given strategy depends on its performance. Strategy mutation is possible in order to explore all possible finite strategies with mutation rates determined by a parameter μ. We include the population structure parameters δ and α, similar to how they

[1] M-coins have some resemblance to bit-coins and share many of the properties of bit-coins, but also differ significantly in the way they are acquired, in how the number in circulation is controlled and how they expire.

Table 2. Row player is the sender, column player is the receiver

receiver → CC sender ↓	CC	CD	DC	DD
CC	b b	$b-c$ $-g$	$-d$ $b+d$	$-c-d$ $d-g$
CD	$-g$ $b-c$	$-c-g$ $-c-g$	$f-g$ $b-c-e$	$-c+f-g$ $-c-e-g$
DC	$b+d$ $-d$	$b-c-e$ $f-g$	0 0	$-c-d-e$ $d+f-g$
DD	$d-g$ $-c-d$	$-c-e-g$ $-c+f-g$	$d+f-g$ $-c-d-e$	$-c-e+f-g$ $-c-e+f-g$

are used in [Traulsen and Nowak, 2007,van Veelen *et al.*, 2012] to explore reciprocity, and provide observations over a unit-square in $\delta \times \alpha$. Note that when $\delta = \alpha = 0$ the sender-receiver-pairs for each game are randomly chosen regardless of their types and change in every round; whereas when $\delta = \alpha = 1$ the sender-receiver-pairs remain constrained to similar types and unchanged from round to round. In general $(\delta, \alpha) \in [0,1]^2 \setminus \{(0,0),(1,1)\}$, the pairing is done with similar or dissimilar types for a round and remain fixed for a random number of rounds of the game.

The simulation model is as follows:

Initialization: Create a random population of N users who choose a repeated-game strategy randomly over a set of seed-strategies. This set of agents provides the population at time $k = 0$.

The simulation model is constructed with the following update-cycle:

Pairing: Using the population at time $(k-1)$ we create $N/2$ random pairings. *Population Structure parameter*: For each pair with probability α one strategy is selected with the other removed and replaced with a copy of the selected strategy. Therefore for a given strategy s within the population its probability of playing itself is $\alpha + (1-\alpha)p_s$ where p_s is the frequency of strategy s's occurrences in the population at time $(k-1)$. Parameter α allows for an investigation into a spectrum of possible population structures from $\alpha = 0$ (random pairing), to $\alpha = 1$ (stronger and general forms of kinship and spatial/network-connectivity-based closeness for $\alpha > 0$).

Strategize: Each selected pair will play a repeated game with a number of plays dependent on a geometric distribution with continuation parameter δ. The expected number of plays per game is $1/(1-\delta)$, for example $\delta = 0$ reduces to single shot games.

Determine Payoff: Strategy payoff is determined using automata and payoff matrix; a multiplicative discount factor for payoff may be introduced (omitted here).

Next Round: A population of size N is re-created by sampling the strategies at time $(k-1)$ using a distribution whose density is computed as proportional to population normalized performances. This set of agents constitutes the population at time k.

Mutate: Each user-agent is subject to the possibility of mutation with mutation rate μ; a mutation creates a strategy one-mutation step from its previously selected strategy determined in the preceding step. Mutation steps may add or delete a state, re-label a state or re-assign an edge destination. Mutation rates are performed in-situ on the population and update the population at time k.

3.1 Behavior Modes (dependent on parameters d, e, f, g)

We summarize the results from our simulation as shown below:

3.2 Strategies

See Figure 1, for a list of strategies whose fitness is studied during the simulation.

We list in Figures 1(a), 1(b), 1(c) and 1(d) strategy-profiles with single state. In the rows below these figures, we list in Figures 1(e), 1(f), 1(g), 1(h), 1(i), 1(j), 1(k) and 1(l) several more strategy-profiles with two states.

3.3 Equilibrium Strategies at a Glance

Figure 2 shows the asymptotic structures of the strategic behavior of the population.

3.4 Limiting Measures of Send Cooperate and Receive Cooperatively

Figure 3 examines the nature of cooperative behavior[2] as a function of the parameters δ and α that jointly determine "correlation of encounters."

[2] Note that, when the cost of checking g is sufficiently large (in particular compared to the other which provide benefit or shifting burden to the attacker) the population will simply penalize any strategy that does so to such an extent that survival of a checking strategy among competing non-checking strategies is extremely rare. The data includes a few thousand runs for which challenging strategies are eliminated from the population (because of the high cost, without a commensurate benefits for doing so): see Fig3(b) [1118], where since the values are constant and zero they are all mapped to the mean of the jet color map (green). Note further that the act of challenging must have a price that coincides with a benefit for doing so (for example when $g = f$) or a means of shifting the cost burden to the attacker (for example when $g = e$).

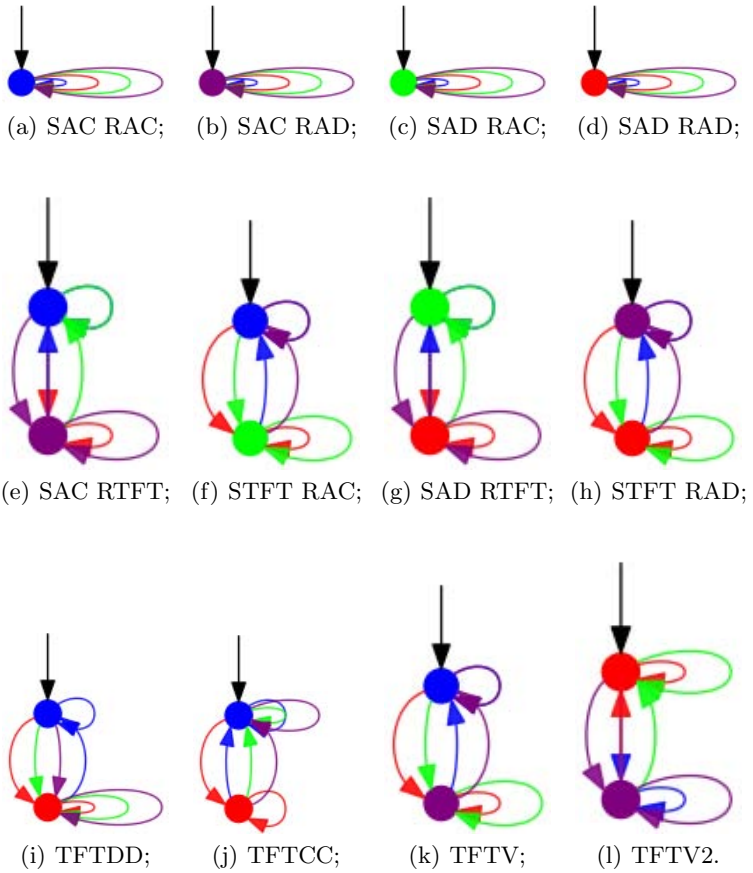
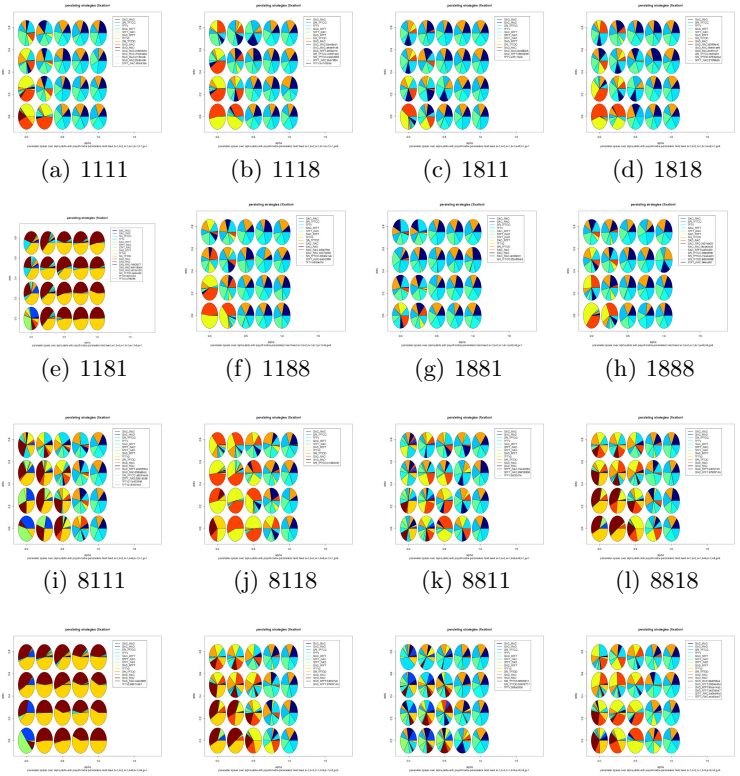


Fig. 1. Repeated game strategy encoded as *finite state automata*. Black arrows indicate initial state. Blue indicates a play of sending cooperatively and receiving trusted. Purple indicates a play of sending cooperatively and receiving untrusted (defect action may challenge reputation of sender). Green indicates a play of sending defect (attacks) and receiving trusted. Finally red indicates a play of sending defect and receiving untrusted. Arrows indicate the transition taken depending on an opponent's previous play. A repeated game may occur for any pairs of agents; the number of plays determined by a geometric distribution continuation parameter δ. Above: twelve seed strategies for population dynamics with evolution pressures for strategy fitness.

4 Discussion

In the JASON report, the committee addressed the following question ($Q2$ on page 4): *Are there "laws of nature" in cyber space that can form the basis of scientific inquiry in the field of cyber security? Are there mathematical abstractions or theoretical constructs that should be considered?* The answer they provided



(a) 1111 (b) 1118 (c) 1811 (d) 1818

(e) 1181 (f) 1188 (g) 1881 (h) 1888

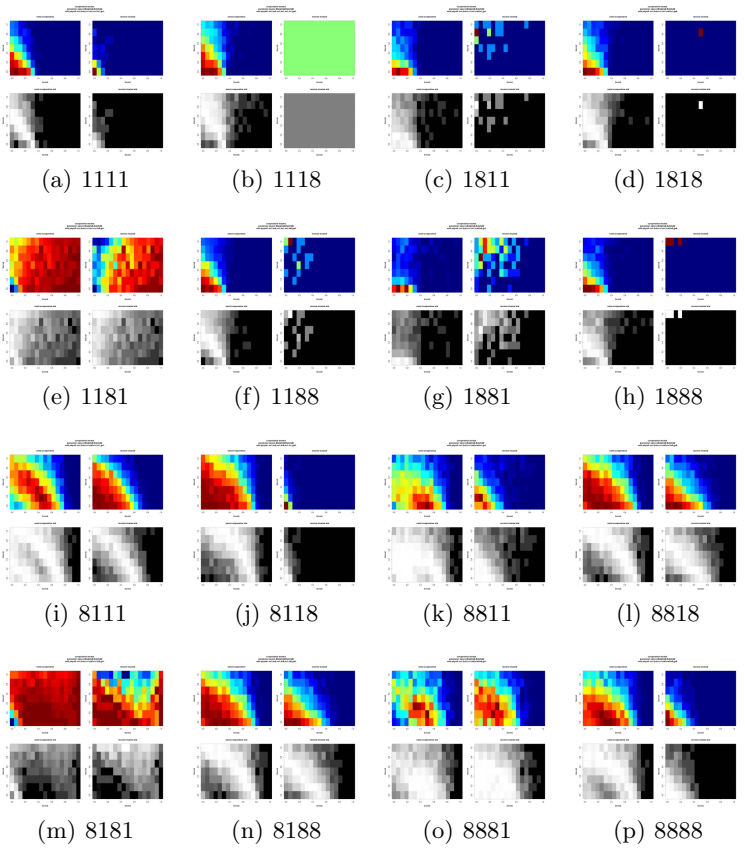
(i) 8111 (j) 8118 (k) 8811 (l) 8818

(m) 8181 (n) 8188 (o) 8881 (p) 8888

Fig. 2. Infrequent mutation rates applied to populations of twelve seed strategies provide a notion of what strategies have advantages and are culled for various environments or settings of payoff matrix values. Exploration of d, e, f, g are shown above. Each chart provides a view of which strategy fixate in the population at various values of d, e, f, g, pie charts are organized over the unit square of α, δ.

is rather pessimistic: "There are no intrinsic "laws of nature" for cyber-security as there are, for example, in physics, chemistry or biology. Cyber-security is essentially an applied science that is informed by the mathematical constructs of computer science such as theory of automata, complexity, and mathematical logic." In contrast, we show that by suitably modeling the agents of a system and the utilities they wish to achieve in cyber space, and under the standard assumptions of "common knowledge of rationality," a suitable law can be imposed on the system, which can evolve to a desirable equilibrium.

We believe that, although our work is preliminary and require further research, it is promising and could prove to be immensely useful, especially to policy makers in the security community.



(a) 1111 (b) 1118 (c) 1811 (d) 1818

(e) 1181 (f) 1188 (g) 1881 (h) 1888

(i) 8111 (j) 8118 (k) 8811 (l) 8818

(m) 8181 (n) 8188 (o) 8881 (p) 8888

Fig. 3. Charts of aggregate population behavior at various values of d, e, f, g showing overall percentage of time a population sends cooperatively and receives trusted. Each chart has four sub charts with average percentage send cooperatively plays shown in the upper left, average percentage of receive cooperatively plays shown in the upper right and standard deviation for each percentage shown below. Each quadrant provides a view for simulations over the α, δ parameter unit square.

Acknowledgements. We would like to thank members of the Software Engineering Institute, and in particular two colleagues: Bill Scherlis and Dean Sutherland, for creating the opportunities for this collaboration. The research reported here was supported by a joint CMU-SEI-NYU grant.

References

Beitollahi and Deconinck, 2012. Beitollahi, H., Deconinck, G.: Review: Analyzing Well-known Countermeasures Against Distributed Denial of Service Attacks. Comput. Commun. 35(11), 1312–1332 (2012)

Bhatia *et al.*, 2012. Bhatia, S., Schmidt, D., Mohay, G.: Ensemble-based DDoS Detection and Mitigation Model. In: Proceedings of the Fifth International Conference on Security of Information and Networks, SIN 2012, pp. 79–86. ACM, New York (2012)

Casey, 2013. Casey, W.: Deterrence for Malware: Towards a Deception-Free Internet (2013), http://blog.sei.cmu.edu/archives.cfm/author/will-casey+

Doron and Wool, 2011. Doron, E., Wool, A.: WDA: A Web Farm Distributed Denial of Service Attack Attenuator. Comput. Netw. 55(5), 1037–1051 (2011)

Fu *et al.*, 2011. Fu, Z., Papatriantafilou, M., Tsigas, P.: CluB: A Cluster Based Framework for Mitigating Distributed Denial of Service Attacks. In: Proceedings of the ACM Symposium on Applied Computing, SAC, pp. 520–527. ACM, New York (2011)

Gallagher, 2013. Gallagher, S.: How Spamhaus' Attackers Turned DNS into a Weapon of Mass Destruction. arstechnica.com (2013), http://arstechnica.com/information-technology/2013/03/how-spamhaus-attackers-turned-dns-into-a-weapon-of-mass-destruction/

Huang *et al.*, 2007. Huang, Y., Geng, X., Whinston, A.B.: Defeating DDoS Attacks by Fixing the Incentive Chain. ACM Trans. Internet Technol. 7(1) (February 2007)

Kargl *et al.*, 2001. Kargl, F., Maier, J., Weber, M.: Protecting Web Servers from Distributed Denial of Service Attacks. In: Proceedings of the 10th International Conference on World Wide Web, WWW 2001, pp. 514–524. ACM, New York (2001)

Lee *et al.*, 2007. Lee, K.-W., Chari, S., Shaikh, A., Sahu, S., Cheng, P.-C.: Improving the Resilience of Content Distribution Networks to Large Scale Distributed Denial of Service Attacks. Comput. Netw. 51(10), 2753–2770 (2007)

Lee, 2013. Lee, D.: Global Internet Slows after Biggest Attack in History. BBC news (2013), http://www.bbc.co.uk/news/technology-21954636

Mitre, 2010. Mitre. Science of Cyber-security. JASON, MITRE Corporation (2010), https://www.fas.org/irp/agency/dod/jason/cyber.pdf

Saint-Andre, 2009. Saint-Andre, P.: Best Practices to Discourage Denial of Service Attacks. XSF XEP (2009), http://xmpp.org/extensions/xep-0205.html

Schwartz, 2013. Schwartz, M.J.: DDoS Spam Feud Backfires: Bulletproof Cyberbunker Busted. Informationweek.com (2013), https://www.informationweek.com+/security/attacks/ddos-spam-+feud-backfires-+bulletproof-cyb/240151895

Traulsen and Nowak, 2007. Traulsen, A., Nowak, M.A.: Chromodynamics of Cooperation in Finite Populations. PLoS One 2(3), e270 (2007)

van Veelen *et al.*, 2012. van Veelen, M., García, J., Rand, D.G., Nowak, M.A.: Direct Reciprocity in Structured Populations. Proceedings of the National Academy of Sciences 109(25), 9929–9934 (2012)

Walfish *et al.*, 2010. Walfish, M., Vutukuru, M., Balakrishnan, H., Karger, D., Shenker, S.: DDoS Defense by Offense. ACM Trans. Comput. Syst. 28(1), 3:1–3:54 (2010)

Williams, 2013. Williams, R.: DDoS Attack Against spamhaus Exposes Huge Security Threat on DNS Servers. hothardware.com (2013), http://hothardware.com/News/DDoS-Attack-Against-Spamhaus-Exposes-Huge-Security-Threat-On-DNS-Servers/

Xie and Yu, 2009. Xie, Y., Yu, S.-Z.: Monitoring the Application-layer DDoS Attacks for Popular Websites. IEEE/ACM Trans. Netw. 17(1), 15–25 (2009)

Yau *et al.*, 2005. Yau, D.K.Y., Lui, J.C.S., Liang, F., Yam, Y.: Defending Against Distributed Denial-of-Service Attacks with Max-Min Fair Server-Centric Router Throttles. IEEE/ACM Trans. Netw. 13(1), 29–42 (2005)