Adversarial Security: Getting to the Root of the Problem

Raphael Phan, John N. Whitley, David J. Parish

Adversarial Security: Getting to the Root of the Problem[⋆]

Raphael C.-W. Phan, John N. Whitley, and David J. Parish

High Speed Networks (HSN) Research Group,
Electronic & Electrical Engineering,
Loughborough University,
LE11 3TU, UK
{R.Phan,J.N.Whitley,D.J.Parish}@lboro.ac.uk

Abstract. This paper revisits the conventional notion of security, and champions a paradigm shift in the way that security should be viewed: we argue that the fundamental notion of security should naturally be one that actively aims for the root of the security problem: the malicious (human-terminated) adversary. To that end, we propose the notion of adversarial security where non-malicious parties and the security mechanism are allowed more activeness; we discuss framework ideas based on factors affecting the (human) adversary, and motivate approaches to designing adversarial security systems. Indeed, while security research has in recent years begun to focus on human elements of the legitimate user as part of the security system's design e.g. the notion of ceremonies; our adversarial security notion approaches general security design by considering the human elements of the malicious adversary.

1 The General Security Problem

This paper sets out to revisit the conventional notion of security. In essence, conventional security represents the security advocate as a boxed-in non-initiator, in that (technical) security mechanisms therein aim to *protect* the good guy against or *cope* with, anticipated attacks. Quoting from [25], the general view is that "security is inherently about *avoiding* a negative". By design, the advocate is not equipped with the ability to initiate actions in the reverse direction towards the malicious adversary.

In that light, cryptographic techniques and network security techniques are traditionally *defensive* mechanisms in face of the malicious adversary. More precisely, confidentiality (encryption), integrity (message authentication codes, hash+signatures), authentication (signatures), non-repudiation (signatures) ensure that in the event of attacks, either data or identities are *protected from* unauthorized adversarial (read and/or write) access or at the very least that any attack is discovered by the victim parties; while intrusion detection or firewalls detect or block adversarial access.

[⋆] Part of this work adversarially motivated by coffee.

Intrusion tolerance, network resilience and proactive cryptography [8] (including forward security [9], key insulation [14], intrusion resilience [13], leakage resilience [26, 23]) techniques are of similar nature, emphasizing on being able to *cope* with or *survive* adversarial attacks.

While it must be said here that the network forensics approach does to some extent provide a channel to get back at the malicious adversary, this is via non-technical means, i.e. legal actions. Another emerging approach, non-technical as well, is that of security economics [7] that can also be seen as more proactive rather than simply defensive.

Taking a holistic view of the security problem, we would like to champion a paradigm shift in the way that security should be viewed, by arguing that the fundamental notion of security should naturally be one that *actively* aims to tackle the root of the security problem: the malicious adversary.

We also champion in this paper the fact that security should fully exploit the fact that the adversary is human-terminated; thus in terms of proactively addressing this root of the security problem, one should bear in mind that the human adversary lives in the real world and is thus influenced by real world factors aside from technical ones. Essentially, security pits human ingenuity (designer) against human ingenuity (adversary). While security research in recent years has begun to consider human factors within security designs in view that legitimate security users are often human (this makes a lot of sense since attackers have long been exploiting this weakness, e.g. social engineering), less research has concentrated on designing security by considering that adversaries are also human-initiated, although to some extent the research direction popularized by CAPTCHA [6] in considering how to identify if a human is present during web based authentication dates back to the work of Naor [22] in 1996.

2 Adversarial Security Design

We propose the notion of *adversarial security.* The adversarial angle of this notion is twofold.

First, it emphasizes on the ideal that security should be the resultant equilibrium established after fair play (to some extent) among all parties, whether honest or malicious. This is akin to the adversarial process e.g. in adversarial legal systems or adversarial politics which is game-like in nature. In contrast, the conventional notion of security does not really capture this since techniques therein are less symmetric in terms of the activeness, i.e. the malicious adversary is the active initiating party while the attacked party is the non-active defending or coping party. What is worse, the adversary bears no consequences from his/her actions nor from actions of the other non-malicious parties, while the non-adversarial parties bear the consequences of their own actions (e.g. lack of emphasis on security increases risk of being attacked) and even those of the adversary. Furthermore, although the provable security paradigm also adopts a game-like approach to defining security, it resembles less the fair play element between opposing sides that should be the nature of an adversarial process.

Second, our notion is so-called adversarial in the sense it aims to emphasize on and get to the root of security problems, i.e. the malicious human-terminated adversary.

2.1 Framework

We start by thrashing out the framework that should influence the design of adversarial security mechanisms.

The root of the problem is the adversary, and the initiator at the adversarial side is a human (indeed, in our present times unlike science fiction where AI machines can be as malicious as humans, behind every adversary is a malicious human). Designing the human adversary in, rather than leaving this out as a non-technical social engineering issue, ties in well to the concept of ceremonial design [15] that is recently gaining popularity e.g. Bluetooth's pairing and schemes based on short authenticated strings [27], where humans are included in the security mechanism design.

Including non-technical issues into the design, including human factors, is also the approach taken by the discipline of systems engineering, which provides techniques to bring different areas of expertise, including both technical and non-technical, into a cross-disciplinary approach, or set of approaches, to solve problems. Using system engineering techniques to solve security problems is likely to provide more robust solutions. Typically the sectors included within a systems engineering brief are categorised into *lines-of-development*. The lines of development are classified by the acronym *TEPID OIL* - standing for: Training; Equipment; Personnel; Infrastructure; Doctrine and concepts; Organization; Information; and Logistics [3]. Taking a systems engineering approach to security means taking into account factors that are not just technical, so would include human factors like motivation and cost.

We can holistically model the adversarial side by classifying the factors that affect this human-involved adversary into differing layers of abstraction in top-down manner:

- Top layer: Psychological (human element)
  - motivation (to mount the attack)
    - benefit (derived from the attack)
    - cost (of attack actions)
    - risk (of being held responsible)
  - social implications
    - reputation (peer status)
- Middle layer: Real World
  - physical consequence to the human adversary
  - legal implications
- Bottom layer: Technical
  - hardness (of attack actions)
  - cost (of attack actions)
  - time (taken by attack actions)

Each of the layers need to be defined as time-varying. Indeed, it is easy to see that technology advances over time, and currently infeasible technical actions may not necessarily be so in the future. Real world factors may also vary with time, e.g. changes in cyberlaws. While motivational factors do tend to be time-invariant, yet for generality the top layer can also be defined as time-varying, to capture for instance the dwindling in value of the protected data e.g. stock market information is no longer useful when it becomes obsolete; thus there is no longer any motivation to mount attacks to acquire obsoleted market info.

Most of the conventional security techniques aim to tackle the security problem by targetting the bottom abstraction layer i.e. the technical layer. For instance, provably secure cryptographic constructions rely on the assumption that it is computationally infeasible (in terms of hardness, cost and/or time) for an adversary to mount technical attack actions.

In contrast, we can use the above listed layer abstractions to define a framework that is adversarial in the sense that the non-malicious parties in their interactions with the adversary are also provided the capability to mount active responses aimed at affecting the factors within those abstraction layers. This interaction can be modelled as an adversarial game. For instance, an adversarial security notion capturing the psychological factor is as follows. Let state_0^x denote the initial value assigned by the adversary to the motivation factor x, and state_1^x denote the state of this at the conclusion of the adversarial game. Then the adversarial advantage can be defined as the difference between state_0^x and state_1^x. The security mechanisms are deemed successful if $\mathsf{state}_1^x << \mathsf{state}_0^x$; i.e. the adversary's motivation is significantly reduced as a consequence of being involved in the attack.

Aside from influencing the design and redesign of security systems, this framework can also be used to create adversarial profiles, e.g. for network forensics and evidence construction.

2.2 Approaches

To understand this holistic, systems engineering-like approach to security systems, here we propose the following approaches be included in the overall view of solving security:

Approach 1: getting nearer to the adversarial source. We can do this in two ways.

– by abstraction: with reference to the abstraction layers, mechanisms can be designed to target factors nearest to the human adversary's mind, i.e. psychological factors. The lower down the abstraction layers, the more external and less attached it is to the adversary; e.g. if a security mechanism complicates the adversary's attack technically, then s/he can retry with another technical one. In contrast, if we design a mechanism that increases the adversary's risk of being caught for attacks s/he had mounted independent of the specific technicalities, this factor remains to internally affect the adversary's motivation. At the same time, different adversarial groups are presenting themselves, and as their motives are removed from previous adversaries, the technical implementation of their attacks are different too [4]. We suggest that involving social understanding and prevention of the motivation of an attack will lead to a stronger system.

This is also an appropriate approach when attempting to prosecute suspects of network attacks: the gathered evidences are considered not only with respect to the technical layer, e.g. packet headers, but evidences are also considered at the psychological layer e.g. the suspect's intent.

- spatially: alternatively, approach 1 could mean getting nearer to the adversary's location, e.g. rather than having firewalls essentially boxing in the attacked machine, to instead have the adversary's network provider or nearby intermediaries boxing in the adversary at his machine. This is beneficial for the network security setting. More precisely, less inconvenience is caused to the attacked machine because it is not penalized for being attacked; instead, the adversary feels the consequences of his/her attack.

  There is one caveat of this approach if only a technical solution is employed. Trying to contain the adversary at the source would work for a conventional Denial of Service (DoS) attack, but is much harder to see effect if it is a Distributed Denial of Service (DDoS) attack where the adversary uses multiple distributed locations to launch the attack. In contrast, adopting approach 1 in the sense of abstraction would mean targeting a higher abstraction layer, i.e. the motivation of the adversary, and this will be equally effective to both DoS and DDoS attacks.

Approach 2: legitimate fightbacks. The idea of fighting back or striking back has been proposed in network security literature [16, 30, 20], albeit sparingly. Quoting [12]: "returning fire has a benefit in suppressing attack activity...a strong offense is a good defense". For instance, [16] reports on how Fortune 500 companies have admitted that they have the capability (via installed software) of counter-attcking hackers. Such measures include flooding the hacker's system. The study also reported that many companies would rather trust their own strikeback capabilities than summon the enforcement authorities. The caveat is that there may be legal issues [21], and furthermore one cannot always ascertain who the adversary is.

[28, 29] suggest to fight fire with fire, i.e. combat DoS attacks by having the legitimate users also launch a kind of DoS on the network resources. This is legitimate since the attack is not directly targeted at the adversary per se, although it causes the same kind of inconvenience to legitimate users as would a DoS attack since a lot of retransmissions are involved and thus wastage of network resources. There are network protocols to help the victim require the adversary to desist, for example IPCAF [5], although an obvious problem with any such protocols is general take-up: it would require all Internet Service Providers (ISPs) to implement additional protocols within their networks.

Rather than actively 'fighting back' towards the adversary using the same weapon as the adversary, which may lead to indefinite vengeful cyber-warfare where the non-malicious parties will be no better than the adversary, we advocate instead to consider legitimate ways and fighting fire with non-fire. Essentially, this means that we use ways that are different from the ones used by the adversary; e.g. we do not launch DoS style responses to DoS attacks, or upon viral infection we do not attempt to propagate the virus back to the source. This sort of legitimate fightback is an interesting open research direction.

Some ways of legitimate fightback can be designed, with reference to the framework of abstraction layers:

- de-motivation: we can design a system such that the benefit derived from an attack is much less than the attack cost. For instance, the business model used by Apple where third party developers build applications that sell for a song; malicious users are then no longer motivated to copy nor pirate these applications since the cost of doing so is not significantly less than actually buying the applications legitimately.
- social pressure: humans tend to bow down to peer pressure. As an analogy: rather than directly punish a misbehaving pupil, a class teacher could subtly get the class to frown on (e.g. laugh at) the pupil by making a cynical remark. In similar vein, the social networking site Facebook.com retracted its new policy to retain user data online after public protest [17].
- removal of cheap resource: distributed attacks require a number of slave hosts in the network, commonly the owners of these machines would not intend their machines to be part of a distributed attack. The implementation of ingress firewalls to prevent attacks on a system is well understood and well used in many Internet sectors [1]. The implementation of ingress firewalls as part of a domestic and commercial Internet connection is now thought of as normal and expected. This has prevented many simple virus attacks on otherwise vulnerable machines. We suggest that egress firewall filtering, if implemented on the scale that ingress firewall filtering has been, will have a similar tide-changing effect in reducing adversarial network traffic.

Approach 3: human involvement. Keeping in mind our approach to design the human element of the adversary right into a security system, the human involvement approach advocates the following design strategies:

- human-tractable but machine-intractable tasks: the strategy here is to have security-critical actions require human mediations; this means the human adversary cannot fully automate his/her attack steps since the tasks cannot be easily performed by machines, and therefore affects the cost and time subfactors as well as increases attack consequences on the adversary. Examples of security systems that use this approach are systems based on short authenticated strings, for example Bluetooth, which requires human-communication channels; and CAPTCHA to combat machine-automated spamming or DoS attacks. While the approach is nice, particular techniques to implement such an approach should be designed with care. For instance, it was recently shown that typical CAPTCHA techniques are known to be weak [2].

– human-involved ceremonies: the strategy of designing systems to include the human terminals, so-called ceremonies [15, 18, 19] allows to capture not just the technical issues but the human elements of users of such systems as well, and therefore relevant issues such as social engineering no longer need to be regarded as out of scope during the design stage.

Approach 4: every action has a reaction. This strategy is to design a system where every action by a party (irrespective of whether it is an honest party or a malicious one) causes a reaction i.e. leads to an effect on the party itself, e.g. each action costing monetarily or resource-wise. For instance, in some security systems for wireless networks [31], incentive based schemes are designed where credit is globally distributed to all parties during setup or on enrollment into the system, and where credits are awarded to or deducted from a party based on its actions.

In this setting, malicious parties will be penalized directly from their attack actions, even if an attack attempt is not successful. This is also particularly relevant in the network security scene, e.g. DoS attacks, spams or spits, where each generated traffic towards the target victim machine incurs a reaction back to the adversary. What is more, non-attacking parties who wrongly accuse others will want to think carefully because their accusing actions will also cause a reaction e.g. cost.

Approach 5: being stateful and bearing grudges. Related to approach 4, in the sense that each attack attempt should update the system state so that there is a reaction back to the adversary that affects the adversary's subsequent actions, the approach here tackles attacks of the exhaustive type, e.g. brute force password (or secrets of low entropy) guessing attacks. The gist is that the security system should remember each attack attempt (even if the attempt does not lead to a successful attack), and be stateful such that the subsequent attack attempt would require significantly more adversarial effort to mount. This idea is used by Bluetooth [10] to discourage bruteforce guessing, by having each repeated attempt lead to a waiting time that is exponentially increasing. On a related note, the idea of bearing grudges has also been applied to discourage misbehaviour by selfish parties rather than discourage attacks, e.g. [11].

3 Concluding Remarks

We have advocated a paradigm shift in the way we address the security problem, i.e. taking a holistics systems engineering-like approach and in doing so including an additional focus on the adversarial angle. Adversarial angle in the sense of fair play between the adversarial and attacked sides, and in the sense of getting right to the source of the problem i.e. the human-terminated adversary. We proposed to treat the factors affecting an adversary as time-varying layers of abstraction; and discussed five approaches with this in mind.

Security should not remain as a purely defensive strategy, quoting [30]: "sitting back and waiting for attackers is a strategy doomed to failure. . . defensive wars are not winnable".

Acknowledgement

We thank the iNetSec 2010 anonymous reviewers and non-anonymous attendees for comments and interest in this research direction.

References

1. D.B. Chapman, E.D. Zwicky, and D. Russell, Building Internet Firewalls. O'Reilly & Associates, Inc. Sebastopol, CA, USA, 1995.
2. C.J. Hernandez-Castro and A. Ribagorda, "Remotely Telling Humans and Computers Apart: An Unsolved Problem," Proc. iNetSec '10, IFIP Advances in Information and Communication Technology, Volume 309, Springer-Verlag 2009, p. 9-26.
3. C. Kerr, R. Phaal and D. Probert, "A Framework for Strategic Military Capabilities in Defense Transformation," International Command and Control Research and Technology Symposium, 2006.
4. BBC News. Political Hacktivists Turn To Web Attacks, 2010. http://news.bbc.co.uk/1/hi/technology/8506698.stm. This is an electronic document. Date of publication: February 10, 2010. Date retrieved: February 10, 2010. Date last modified: February 10, 2010.
5. C.-H. Wu, C.-C.A. Huang and J.D. Irwin, "Using Identity-Based Privacy-Protected Access Control Filter (IPACF) to Against Denial Of Service Attacks and Protect User Privacy," Proc. SpringSim '07, San Diego, CA, USA, 2007, pp. 362–369.
6. L. von Ahn, M. Blum, N.J. Hopper and J. Langford, "CAPTCHA: Using Hard AI Problems for Security," *Advances in Cryptology - Eurocrypt '03*, LNCS 2656, 2003, pp. 294-311.
7. R. Anderson and T. Moore, "The Economics of Information Security," Science, vol. 314, no. 5799, 2006, pp. 610-613.
8. B. Barak, A. Herzberg, D. Naor and E. Shai, "The Proactive Security Toolkit and Applications," Proc. ACM CCS '99, 1999, pp. 18-27.
9. M. Bellare and S. Miner, "A Forward-Secure Digital Signature Scheme," Advances in Cryptology - Crypto '99, LNCS 1666, 1999, pp. 431-448.
10. Bluetooth SIG, "Bluetooth Core Specifications v4.0," 17 December 2009.
11. S. Buchegger and J.Y. Le Boudec, "Nodes Bearing Grudges: Towards Routing Security, Fairness, and Robustness in Mobile Ad Hoc Networks," Proc. PDP '02, 2002, pp. 403-410.
12. F. Cohen, "Managing Network Security: Returning Fire," Network Security, vol. 1999, no. 2, 1999, pp. 11-15.
13. Y. Dodis, M.K. Franklin, J. Katz and M. Yung, "Intrusion-Resilient Public-Key Encryption," Topics in Cryptology - CT-RSA '03, LNCS 2612, 2003, pp. 19-32.
14. Y. Dodis, J. Katz, S. Xu and M. Yung, "Key-Insulated Public-Key Cryptosystems," Advances in Cryptology - Eurocrypt '02, LNCS 2332, 2002, pp. 65-82.
15. C. Ellison, "UPnP Security Ceremonies: Design Document", October 2003. Available online at http://www.upnp.org/download/standardizeddcps/UPnPSecurityCeremonies_1_0secure.pdf

16. B. Gengler, “Strikeback,” Computer Fraud & Security, vol. 1999, no. 1, 1999, pp. 8-9.
17. B. Johnson and A. Hirsch, “Facebook Backtracks after Online Privacy Protest,” Guardian.co.uk, 19 February 2009.
18. C. Karlof, J.D. Tygar and D. Wagner, “Conditioned-safe Ceremonies and a User Study of an Application to Web Authentication,” Proc. NDSS ’09, 2009.
19. C. Karlof, J.D. Tygar and D. Wagner, “Conditioned-safe Ceremonies and a User Study of an Application to Web Authentication,” Proc. SOUPS ’09, 2009.
20. V. Jayawal, W. Yurcik and D. Doss, “Internet Hack Back: Counter Attacks as Self-Defense or Vigilantism?,” Proc. ISTAS ’02, 2002.
21. J.H. Matsuura, “Digital Victim or ”Vigilante”: Legal and Ethical Limits to Online Self-Defense,” Proc. Ethicomp ’04, 2004, pp. 629-634.
22. M. Naor, “Verification of a Human in the Loop, or Identification via the Turing Test”, September 1996. Available online at http://www.wisdom.weizmann.ac.il/∼naor/PAPERS/human_abs.html
23. R.C.-W. Phan, K.-K.R. Choo and S.-H. Heng, “Security of a Leakage-Resilient Protocol for Key Establishment and Mutual Authentication,” Proc. ProvSec ’07, LNCS 4784, 2007, pp. 169-177.
24. B. Schneier, “The Psychology of Security,” *Communications of the ACM*, vol. 50, no. 5, 2007, pp. 128.
25. B. Schneier, “How the Human Brain Buys Security,” IEEE Security & Privacy, vol. 6, no. 4, 2008, pp. 80.
26. S. Shin, K. Kobara and H. Imai, “Leakage-Resilient Authenticated Key Establishment Protocols,” *Advances in Cryptology - Asiacrypt ’03*, LNCS 2894, 2003, pp. 155-172.
27. S. Vaudenay, “Secure Communications over Insecure Channels based on Short Authenticated Strings,” Advances in Cryptology - Crypto ’05, LNCS 3621, 2005, pp. 309-326.
28. M. Walfish, H. Balakrishnan, D. Karger and S. Shenker, “DoS: Fighting Fire with Fire,” *Proc. HotNets ’05*, 2005.
29. M. Walfish, M. Vutukuru, H. Balakrishnan, D. Karger and S. Shenker, “DDoS Defense by Offense,” ACM SIGCOMM Computer Communication Review, vol. 36, no. 4, 2006, pp. 303-314.
30. D.J. Welch, N. Buchheit and A. Ruocco, “Strike Back: Offensive Actions in Information Warfare,” Proc. NSPW ’99, 1999, pp. 47-52.
31. Y. Zhang, W. Lou and Y. Fang, “SIP: a Secure Incentive Protocol against Selfishness in Mobile Ad Hoc Networks,” Proc. IEEE WCNC ’04, 2004, pp. 1679-1684.