Layer 8 Tarpits: Overwhelming Malicious Actors With Distracting Information

Toni Virtanen and Petteri Simola
Finnish Defence Research Institute, Tuusula, Finland
toni.virtanen@mil.fi
petteri.simola@mil.fi

Abstract: This paper presents a concept for utilising falsified documents and disinformation as a security measure by diminishing the utility of the stolen information for the attacker. Classical definition of tarpitting honeypots is to create virtual servers attractive to worms and other malware that answer their connection attempts in such a way that the machine on the other end becomes stuck. A common extension to the OSI model is to refer the user as the layer 8 on top of the application layer. By generating attractive looking but falsified documents and datasets within our secured network along with the real information, we could be able to force the malicious user on the other end similarly to be 'stuck' as they need to dig through and verify all the information they have managed to steal. This in effect slows down the opponents' decision making speed, can make their activity in the network more visible and possibly even mislead them. The concept has similarities to the Canary trap or Barium Meal type of tests, and using Honey tokens to help identify who might be the leaker or from which database the data was stolen. However, the amount of falsified data or fake entries in databases in our concept is significantly larger and the main purpose is to diminish the utility of the stolen data or otherwise leaked information. The requirement to verify the information and scan through piles of documents trying to found the real information among them can give more time to the defender to react if the attack was noticed. It will also reduce the value of the information if it is just dumped in the open, as its contents and authenticity can be more easily questioned. AI powered methods such as the GPT-3 that can generate massive amounts very realistic looking text which is hard to differentiate from human generated texts could make this type of concept more feasible to the defender to utilise. The shortcoming of this concept is the risk that legitimate end-users could also confuse the real and falsified information together if that is not prevented somehow.

Keywords: tarpit, honeypot, information security, disinformation, decision making

1. Introduction

Securing critical information and assets is paramount for operational security in the military and other defence critical organizations. According to a recent Ponemon Institute study, 52% of all data breaches are caused by malicious or criminal attacks, 23% happen due to human error, and 25% are caused by both IT and business process failures and system glitches at an average total cost of $4.27 million (Ponemon Institute, 2020). Defenders have long been using monitoring and intrusion detection systems to alert against unauthorized access (Carol & Raouf, 2014). These can simply be sensors that are spread across the network to alert the Cyber Security Operations Centres (CSOC) if they detect traffic that are flagged as suspicious by the detection scripts monitoring the traffic. In addition to detecting anomalies within the network, various kinds of honeypots and honey tokens can be further used to bolster the detection capability of the defenders. These could be files, network locations and services, which look interesting to the hacker and their value lies in the fact that they are probed and attacked by the hackers. Honeypots can therefore be considered as a deceptive cyber security systems and are valuable in distracting attackers away from more valuable targets, provide early warnings on new exploitation methods and allow in-depth examination of adversaries tactics during and after an attack (Qassrawi & Hongli, 2010). Honeypots can be classified according to the level interaction they simulate for the attackers, from low-interaction honeypots that simulate a limited number of network services and vulnerabilities to medium- and high-interaction honeypots that provide increasing levels of emulation of a real operating system for the attacker to interact with. (Mokube & Adams, 2007)

The OSI model (Open Systems Interconnection model) is a seven-layer abstract model that describes an architecture of data communications for networked computers (ISO, 1994). It consists of physical layer, data link layer, network layer, transport layer, session layer, presentation layer and application layer. Although the seven-layer OSI model is not as common as the more widely used five-layer TCP/IP model (Braden, 1989), quite often the application layer is still referred as the seventh layer. The OSI model have also been extended by referring the layer 8 as the "user" layer. (Bauer & Patrick, 2004) This user layer represents the Human-Computer Interaction elements of the network communication architecture and some have even added political aspects to the layer 8. As the key factor for any honeypot is to deceive the attacker into believing the intrusion is real (Odemis, et al., 2018). Therefore, especially with the high interaction honeypots, it is crucial that in addition to looking technically realistic, they need to target the emotional and cognitive processes of the hackers through

deception. One such deceptive measure is to lure the attacker towards a honeypot by presenting that as a more easy or valuable target instead of the real system. This is often done by leaving easy, but not suspiciously easy, vulnerabilities in the honeypot. It is quite fitting therefore that the term honeypot comes from human intelligence (HUMINT) terminology, referring to a strategy where an attractive male or female agent is used to seduce individuals and exploit the created sexual relationship to coerce them to cooperate (Odemis et al 2018).

Tarpit honeypots on the other hand are generally virtual servers attractive to worms and other malware that answer their connection attempts in such a way that the machine on the other end becomes stuck in a connection loop. Although especially effective against automated scripts that might be stuck for a very long time, tarpits can also delay human hackers significantly before they realize there is something wrong (Zobal, et al., 2019). A seminal example of a classical tarpit presented by Tom Liston was the LaBrea tarpit that tricked the attacker with a TCP three-way handshake by answering the connection establishing SYN-packets with a SYN/ACK packet while not actually opening any connection and ignoring the final ACK packet of the handshake. It was developed especially to delay and hinder worms and similar malware wasting the resources of those malicious actors using them. (Liston, 2002) For a more interactive tarpit honeypot, this type of delaying and hindering the attacker's reconnaissance of the system would be to use fake error messages and exaggerated delays (Rowe & Goh 2007).

As the title suggests, this paper discusses whether similar tactics used in honeypots and tarpits can be used against the actual malicious person or organization behind an attack, e.g. the 8-layer. By generating attractive looking but falsified documents and datasets within a secured network along with the real information, it could be possible to force the malicious user on the other end similarly to be 'stuck' as they need to dig through and verify all the information they might have managed to exfiltrate out of the secured network.

Although this concept is not technically a preventative measure, it could act as a deterrent by reducing the significance and utility of the exfiltrated data. It could also help to mitigate potential repercussions from a possible data leak, as it would be easier to question the authenticity of the leaked information. An example of this type of cyber-blurring has already been documented with the Macron's e-mail leak just 24 hours before the 2017 French presidential election (AFP, 2017). Macron's campaign team knew they could not protect themselves 100 % from hacking and sophisticated spear-phishing attacks, so they created false accounts, documents, and e-mails that would at least slow the attackers, as they would need to verify all of them. (Nossiter & Sanger, 2017) They were also proactive by often stating in the media that they have been under hacking attempts during the whole 2-year campaign period. Effectively making a narrative that some leaks are eventually going to happen, and when it did, they manage to take initiative by making statement just hours after the leak. Admitting that although there was a leak, it was mixed with falsified information "in order to sow doubt and disinformation" (Vilmer & others, 2019). With their fast reaction to the leak, they managed to change the discussion on the content of the leak to the trustworthiness of the leak. If the data were already mixed with falsified information, taking this type of initiative would be even easier to the defender.

2. Theory

2.1 Increasing opponent's workload

From a theoretical perspective, honeypots and tarpits can be consider as a form of deception. In military deception, it is all about gaining and maintaining initiative. Usually attackers have that initiative, and thus has an advantage over the defenders, who often are forced into a reactive position. The Macron´s campaign case presented earlier is a good example of such attempt to keep the initiative, even if a breach or leak does occur. There are several different ways in defining deception and military deception. One way is to categorise deception in two main categories: A-type and M-type deception. A-type, which is also called ambiguity increasing deception is designed to create general confusion and to distraction. M-type, which is also called misleading deception is designed to mislead an enemy into believing a specific deception plan (Daniel, et al., 1980). The Layer 8 tarpits can be considered to work mainly as A-type deception. Attackers are unable to distinguish what is real and relevant and what is fake or irrelevant and thus are forced to work through all documents, which is time and resource consuming. If the material is properly prepared, the selection of correct documents are always also prone to so-called human errors. In specific cases, the layer 8 tarpits can also be prepared to act as a type-M deception. However, such cases would require good threat intelligence on the malicious actors so that

misleading information could be created specifically to them. The type-M layer 8 tarpits would need more active curating and would be time limited, as they would need to be tightly linked to upcoming events.

The concept for layer 8 tarpits is on generating information overflow and on our humane inability to handle such overflow to the malicious actors. American psychological association defines information overflow as *"the state that occurs when the amount or intensity of information exceeds the individual's processing capacity, leading to anxiety, poor decision making, and other undesirable consequences"* (APA). When dealing with vast amount of information we are prone to use certain cognitive rules of thumb and shortcuts that ease our decision-making (Gigerenzer & Gaissmaier, 2011). However, these shortcuts also makes our decision making biased and vulnerable to errors. Knowing these biases can help defender create content that is most appealing to attackers.

There are literally hundreds of cognitive biases and going through them all would require a whole book. Instead we will present three of the most well-known and studied biases that are relevant in the scope of this paper. One of the most studied bias is the so-called confirmation bias, which causes us to seek and accept information that we expect to find and that supports our previous assumptions (Nickerson, 1998; Tversky & Kahneman, 1973). Thus, if we know what attacker is hoping to find we can create fake data that fits in attackers assumptions. Second well-studied bias is the negativity bias. Negative information gains our attention more often than positive (Baumeister, et al., 2001; Rozin & Royzman, 2001). Therefore, the valence of information may be crucial when the malicious actors goes through the material and any negative information from their viewpoint gains more attention than other information. The third bias presented here is related to the social nature of our thinking. We prioritize social information over any other, so piece of information contains something social, for example information about relationships or scandals are more likely to gain attention thus to become selected as significant information (Raichle, et al., 2001; Schilbach, et al., 2008).

Our decision-making is not as rationale as we wish to think. In cognitive demanding cases such as information overflow, we are prone to use certain heuristics of which above-mentioned are just few examples. Knowing human decision-making can help us create tarpits that targets to the "layer 8", e.g. the malicious users and organizations behind the attacks. Obviously, one can argument that well educated attacker also knows these heuristics or biases. Fortunately, even though we are somewhat able to recognize erroneous thinking in others, we are often blind to our own vulnerabilities. (West, et al., 2012).

2.2 Adding more workload with encryption

Encrypting both the falsified and real documents in a similar manner would even further increase the time and resources needed to utilize the intercepted information by malicious actors. Even if they would be able to intercept and decrypt the information with massive computational resources, they still would need to use even more resources to try to verify which of the intercepted documents contain the genuine information and which are decoys. This type of encryption by obfuscation would add another layer of security especially for data transfers and could be used in a similar manner as with the canary trap and barium meal tests. Where sending varying versions to different receivers to reveal an insider leak. In those cases, the documents are purposefully varied by using synonyms and other small differences, without changing the actual meaning of the content. However, if information about the contents would leak out, one could see whose version matches the leaked document. One could also just send multiple distorted copies along with the real document and use shared symmetric-keys to identify which document is the genuine one and which are decoys, while also encrypting the whole message using asymmetric encryption methods. The drawback on this would be the increased requirements for bandwidth in data transfers and the need to have robust version control for the real documents and decoys for both parties.

2.3 Decreasing defenders workload

Machine Learning and Natural Language Processing systems could ease the workload of generating massive amounts very realistic looking text. Especially by utilizing Generative Pre-training, by first training the system with massive amounts of diverse corpus of unlabelled text, followed by a more context depended corpuses and fine tuning the model for a specific task. (Radford, et al., 2018) As an example, the current generation GPT-3 that has been claimed by its developers to generate so realistically looking text that it can be difficult to determine whether it was written by a human (Brown, et al., 2020). Using insights from deception and human decision making along with the typical document style and language within the organization, we could develop a process

that creates tempting and realistically looking false documents that tap on the cognitive biases and vulnerabilities of the potential malicious actors. To make the distinction even more difficult, it could be possible to use the real users' documents and file structures as a baseline and generate similar but falsified duplicates that mimic their style and behaviour. To further ensure that an opponent could not train their own algorithm to recognize the fake documents from real ones, it would be best to use more than one method of generate the fake documents.

3. Concept model for layer 8 tarpit

A simplified network topology layout for using deceptive tarpit networks is presented below. Real user data is fed to a pre-trained machine-learning algorithm as a context and situation specific training data (Distorted mirroring). The ML generates realistic but distorted images of the user data and virtual machine images, which are then mashed-up together with the original virtual workstations. Hardware or other means of authentication are then used to filter out the falsified entries from the authorized user's view. The layer 8 tarpit virtual desktops would also need to have some level of authentication, so that they would not be easy to distinct from the genuine virtual desktops. (Figure 1)

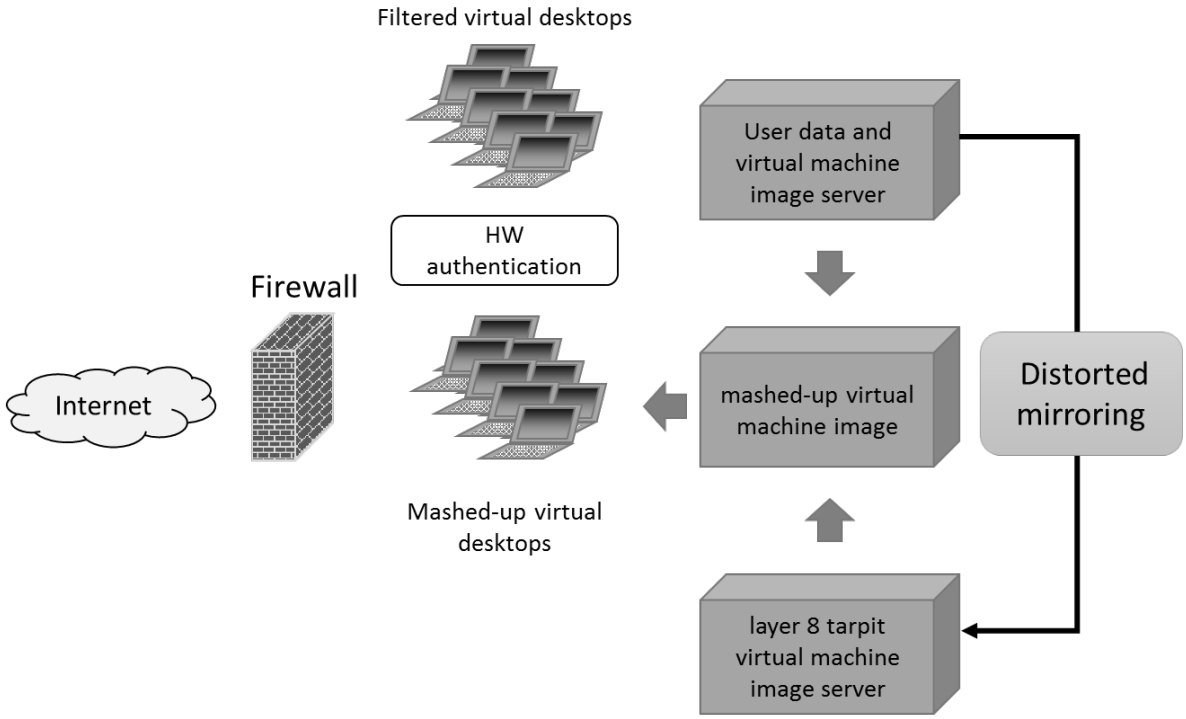

Figure 1: Simplified virtualized network topology model for layer 8 tarpits. User data is mirrored through a ML system that distorts most of the information to a realistically looking layer 8 tarpit counterpart. The real virtual machine image is then mashed-up with the layer 8 tarpit virtual machine. Hardware authentication is then used to give filter out the falsified entries from authorized users view.

4. Discussion

This paper presented a concept for using deception and falsified information as a way to delay and hinder the malicious actors behind and attack. In other words, targeting the human from the extended OSI network communications layer model e.g. the 8-layer. It discussed on using theories and understanding from psychology to enhance the falsified information so that it would create an information overflow to the opponent depleting their resources and reducing the utility of the information. AI powered methods such as the GPT-3 that can generate massive amounts very realistic looking text which is hard to differentiate from human generated texts could make this type of concept more feasible to the defender to utilise in the future. The presented concept for a network topology model illustrates a hypothetical application for using disinformation and obfuscation as a security measure. Overwhelming the attacker with similar looking targets and introducing ambiguity for not knowing whether they have managed to exfiltrate genuine information or just information from a decoy tarpit.

The obvious shortcoming of this type of defence by obfuscation is the added costs of space and bandwidth that the deceptive information would take within the defender's own network. Our estimate is that this type of system would require as much resources as running a state-of-the-art honeynet system. This type of defensive measure should also never be used as the only line of defence, but as an additional layer security that mainly would mitigate the damage after a successful attack. Another problem comes with the risk of authorized users confusing the real and falsified information if there is not good processes to keep them separated or identify which is real and which is not. The final issue with using deception as a defensive measure comes from potential legal ramifications that this type of approach might have on corporate, national or international level. As an example, there might be national legislations on misconduct where a public servant should not mislead citizens or give a false statement when in office.

This paper did not focus on how deception could be used as a pre-emptive defensive measure, by introducing false error messages, demands and stalling in high-interaction honeypot systems. This type of active honeypots could be effective in giving more time to the defenders to react once an attack has been detected. (Rowe & Goh, 2007). However, although our concept of targeting the human factor of the malicious actors and using deception to overwhelm the opponent with falsified information is mostly reactive. It can level the cyber battlefield by hindering the opponents' decision-making speed, alter their situational awareness and give an opportunity for the defender to claim the initiative to mitigate the damages from a leak in the information space.

References

AFP, 2017. Macron says hacked documents have been mixed with false ones to 'sow doubt and disinformation'. TheJournal.ie, 6 5.

APA, n.d. APA Dictionary of Psychology. [Online] Available at: https://dictionary.apa.org/

Bauer, B. & Patrick, A. S., 2004. A human factors extension to the seven-layer OSI reference model. Retrieved January, Volume 6.

Baumeister, R. F., Bratslavsky, E., Finkenauer, C. & Vohs, K. D., 2001. Bad is stronger than good. review of general psychology.

Braden, R., ed., 1989. RFC1122: Requirements for Internet hosts-communication layers. s.l.:IETF.

Brown, T. B. et al., 2020. Language models are few-shot learners. arXiv preprint arXiv:2005.14165.

Carol, F. & Raouf, B., 2014. Intrusion Detection Networks : A Key to Collaborative Security.. s.l.:Auerbach Publications.

Daniel, D. C. et al., 1980. Multidisciplinary perspectives on military deception, s.l.: Calhoun.

Gigerenzer, G. & Gaissmaier, W., 2011. Heuristic decision making. Annual review of psychology, Volume 62, p. 451–482.

ISO, 1994. ISO/IEC 7498-1: 1994. Information technology-Open systems interconnection-Basic reference model: The basic model.

Liston, T., 2002. Tom Liston talks about LaBrea. [Online] Available at: https://labrea.sourceforge.io/Intro-History.html

Mokube, I. & Adams, M., 2007. Honeypots: concepts, approaches, and challenges. s.l., s.n., p. 321–326.

Nickerson, R. S., 1998. Confirmation bias: A ubiquitous phenomenon in many guises. Review of general psychology, Volume 2, p. 175–220.

Nossiter, A. & Sanger, D. E. &. P. N., 2017. Hackers Came, but the French were prepared. The New York Times, 9 5.

Odemis, M., Yucel, C., Koltuksuz, A. & Ozbilgin, G., 2018. Suggesting a Honeypot Design to Capture Hacker Psychology, Personality and Sophistication. s.l., s.n., p. 432–438.

Ponemon Institute, 2020. 2020 cost of data breach study: Global analysis. [Online] Available at: www.ponemon.org

Qassrawi, M. T. & Hongli, Z., 2010. Deception Methodology in Virtual Honeypots. s.l., s.n., pp. 462-467.

Radford, A., Narasimhan, K., Salimans, T. & Sutskever, I., 2018. Improving language understanding by generative pre-training.

Raichle, M. E. et al., 2001. A default mode of brain function. Proceedings of the National Academy of Sciences, Volume 98, p. 676–682.

Rowe, N. C. & Goh, H. C., 2007. Thwarting cyber-attack reconnaissance with inconsistency and deception. s.l., s.n., p. 151–158.

Rozin, P. & Royzman, E. B., 2001. Negativity bias, negativity dominance, and contagion. Personality and social psychology review, Volume 5, p. 296–320.

Schilbach, L. et al., 2008. Minds at rest? Social cognition as the default mode of cognizing and its putative relationship to the "default system" of the brain. Consciousness and cognition, Volume 17, p. 457–467.

Tversky, A. & Kahneman, D., 1973. Availability: A heuristic for judging frequency and probability. Cognitive psychology, Volume 5, p. 207–232.

West, R. F., Meserve, R. J. & Stanovich, K. E., 2012. Cognitive sophistication does not attenuate the bias blind spot.. Journal of personality and social psychology, Volume 103, p. 506.

Vilmer, J.-B. J. & others, 2019. "Macron Leaks" Operation: A Post-Mortem. Atlantic Council.

Zobal, L., Kolář, D. & Fujdiak, R., 2019. Current State of Honeypots and Deception Strategies in Cybersecurity. s.l., s.n., pp. 1-9.