2015

Steganography as a threat – fairytale or fact?

Tom Cleary

DOI: 10.4225/75/57b3f4c0fb888
13th Australian Digital Forensics Conference, held from the 30 November – 2 December, 2015 (pp. 49-53), Edith Cowan University Joondalup Campus, Perth, Western Australia.

STEGANOGRAPHY AS A THREAT - FAIRYTALE OR FACT?

Tom Cleary
Independent Researcher, Perth, Australia

Abstract

Almost since the birth of the Internet, there has been a fear that steganographically-encoded threats would be used to bring harm. Serious consideration has been given to the idea that merely downloading an image could introduce malware. Yet, for decades, evidence of this malware channel has been missing in action. There is still an unwritten assumption that images are harmless. Many vendors have implicitly avoided producing defences against steganographic threats. Is it truly impossible to make a widely harmful exploit this way or have malicious actors accepted general wisdom? Three recent papers suggest that there may be a new chapter ahead of us in this field. Practicable methods employing steganography for real world attacks are being uncovered. Early evidence supports the assertion we are seeing new developments in this field, thus far. So, are we seeing a new frontier in exploit development or is there another false dawn for steganographic threats?

Keywords
Steganography, Internet Threats, Exploit development

CONTEXT

Since very early in Internet development, scholars have considered the role that a harmful image could play in stealthy intrusion (Fabien, 1998). Early publications showed the results of surveys conducted to baseline the presence of steganography in collections of images (Provos & Honeyman, 2001). The results were unsurprising; very little evidence of this threat style was seen. Over the decades, other studies followed from these papers and confirmed that malicious actors do not appear to use steganography in the wild (Sujatha, Purushothaman, & Rajeswari, 2014). Yet, the idea that methods could be found to bypass all existing defences with a simple picture persist in the imagination of both defenders and attackers. Recently, there has been credible evidence published of developments being made using steganography to produce viable compromises in the real world.

As Internet defences evolved, the attacker community built improved malware in parallel with, and reflective of, defender activities. Where harm was caused, defences improved (Grimes, 2011). Since the malware community looks to gain maximal advantage and since most advantage was gained through viruses, trojans and remote access tools (RATs), these exploit methods were where most defensive effort was spent. Part of this evolution may be caused by Moore's Law preserving the balance between defnders and attackers, where costs remain equivalent. Part may be a reaction to organisational perceptions of benefit from investment - enough iron to defend against known threats is the risk tradeoff point. If defenders budget too much effort to repelling advanced but unlikely threats, then lack of reaction to ‘clear and present danger' from common threats will cause budgets to suffer. This tradeoff enforces preparation only for risks visible to fiscal managers.

Whilst justifiable and defensible, this is a position which clearly enables support from senior management and also fosters minimized spending on security defence. Management loves this idea. Unfortunately, the side effect is that nothing to counter novel threats will be supported.

In recent years, advanced persistent threats (APTs) have forced a change of approach onto businesses, intent on moving towards increased virtual presence and minimized costs. For security professionals the worrying aspect of this change is that business' reluctance to acknowledge novel threats prevents keeping pace with the reality of deployed malware. In fact, business is so resistant to confronting the changing landscape of cyber attacks, we are in an era where the only way to compel improvements in protection seems to be to remove chief executives from failing organizations (Basu, 2015). The experiences of Target, the U.S. Office of Personnel Management and Ashley-Madison show that this trend is alive, well and not limited to any sector of the market. To attackers, this probably means that improvements of organisational defences will be seen in the relatively near future.

At its simplest, steganography is a way of hiding one digital object inside another by taking advantage of the various slack spaces, redundancies and unused encoding spaces information standards provide. This allows a steganographer to hold malicious material in plain sight. It is increasingly popular in places where other, usually more direct, exploit methods would not succeed. The methods commonly used to encode steganographically concealed content into an image are many and varied, but began in earnest with "least significant bits" (LSBs). LSB encoding is available as a side effect of using colours in image files. Colours do not require a full byte to represent a pixel, but image rendering code takes sequences of bytes to represent those colours, for ease of

alignment. Many tools can be used to stuff additional content into an image through LSBs, undetectable to humans. Pattern detection by code is effective, to a degree, because whatever convention the attacker uses to represent his own information, the original image and the altered image are unlikely to contain similar variations in content, in the same places. So, adding differing LSBs to an area where colours are the same, for example, should be detectable to a program. Whichever method is used to conceal content, differences in entropy should be detectable, especially if one can compare to the original image. Sadly, those opportunities rarely present themselves real-time, on the Internet. Methods that stand alone are less effective in finding steganography in use. The basics of practical steganography include elements of robustness, where common image transformations should not eliminate hidden content. However, using cryptography in addition to steganographic methods tends to reduce robustness whilst assisting in hiding content. Making hidden content appear random through cryptographic means, increasing the difficulty in detecting steganography has the disadvantage that anything which alters encrypted content may make it irretrievable (Al-Ani, Zaidan, Zaidan, & Alanazi, 2010) For attackers intent on compromising Cloud hosted systems, when defenders are conditioned to trust the efficacy of preventive controls and where volatile populations of images accompany end user behaviours, faith in harmless images may be misplaced. Although known instances of detected steganographic methods in the wild are rare, more are being uncovered. As organisations lose faith that they are safe because evidence of compromise is not visible to management, they may need to look more deeply at their stored content. In many cases, evidence of compromise may not be visible to anyone, especially if no-one is actively seeking it out. Prevailing attitudes to security practice may need to adapt to increasing use of hidden attacks, when the penalty is attrition at the top levels.

So, from the perspective of an attacker, the time has come to start a new development push and implement tools to exploit different channels. This might include channels that have a long history, but have been ignored by the industry - steganographic channels.

CURRENT SITUATION

It is widely acknowledged that the major players in the malware industry are either backed by organized crime or Nation States ("Cyber and Technology Enabled Crime," 2015). The ‘hobbyist' culture that prevailed in preceding decades is largely extinct. It has been replaced by a world where the individual hacker is paid to develop exploits to the specification of an employer. Whether driven by bug bounties, crimeware scams or military/defence organizations, few truly independent, significant developers of malware remain. The nature of the generated malware is also different. The reason that noisy attacks like ‘Code Red' and ‘Slammer' have vanished isn't likely to be because Law Enforcement has succeeded in removing the originators from the scene (despite the hype that Kevin Mitnick's incarceration once generated..) It is more likely because attackers have sought to monetise what they find during their activities. What would be the point in owning someone else's machine and not extracting as much value as you can? Attackers do not want the ‘shoutz and greetz' any more. They would much prefer to get paid. Mainstream businesses are finally waking up to the changed opposition they face. With the move to ‘the Cloud' ahead of us, we are about to experience a sea change in approaches to defence against malware. If the thesis presented here is correct, we should also see evidence that a change of approach is being developed amongst attackers. The principle is that we need three recent, concrete examples to suggest an idea is not entirely attributable to chance. I will present three recent instances where the idea of using steganographic means to compromise organizational defences have proven to be tangible, instead of imaginary.

‘STRAWS IN THE WIND'

Case One

Saumil Shah, an independent security researcher, has published a paper that discusses using malicious images as an attack vector for a complete compromise in an image (Shah, 2015). He develops a thesis that states using ‘slack space' in design formats of a variety of popular image standards permits pictures to be used as malware delivery containers. To demonstrate the proof of concept he has published detailed methods for using PNG and JPEG images that can be rendered as visible images and as Javascript, simultaneously. The idea is that a web page provides instructions to visiting browsers, which automatically run the exploit. The exploit is hidden in a real image, which can pass any of the tests currently in use for edge protection - it is a real image, for most purposes. The trick in this method is to use some area of the file containing the image data to expose attributes recognizable as a browser script language, for example Javascript, close to the start of file. Unlike previous methods, this one doesn't depend on the mismatch between auto-compressed attributes at the end of a file and most image formats which conform to the ‘Tag, Length, Value' sequential evaluation method from the start of

file. This means that precautions taken for prior generations of steganographic attack detect these files poorly. That said, this is not the ‘silver bullet' of steganographic exploitation. One of the main benefits for attackers would be if they could use the almost universal JPEG format alone for delivering attacks. Indeed, this proof of concept does permit JPEG to contain the attack. However, the published method depends on having the JPEG ‘size' field contain characters interpretable as Javascript comments. This means that any file which has the field set to a value other than hex "00 10" would be suspect. Being able to easily spot likely attack images would negate most of the advantage of the method. Examination of all the JPEG files cached on an active Proxy server showed no files with a header value other than that above. This suggests that a viable exploit may not come through this path unless a valid and widely adopted reason to change header fields emerges. At present, this seems unlikely.

Case Two

Operation ‘Tropic Trooper' - this campaign was brought to light by Trend Micro earlier this year. It has all the hallmarks of a ‘tried and true' phishing exploit, with the interesting exception that it uses steganographic methods to evade the edge protections of some sensitive organizations in Asia, notably Taiwan. The initial response to the phishing campaign is an image file that is decoded with a string known to the attacking script. Once inside an infested client, the typical sequence of securing a backdoor, hiding itself and pivoting to entrench itself in the organizational fleet takes place alongside ‘calling home' to its' C&C herder. The paper published by Trend suggests that this APT has been in place since 2012. (Alintanahin, 2015) The use of a typical botnet style mechanism, linked to the use of steganography in the course of an APT infestation would seem to be an increase in sophistication over the traditional view of steganography. Typically, it has been used merely to hide passive information in transit. The paper also hints that the use of steganographic techniques is gaining popularity with the more esoteric protagonists on the Net. The worry is that where they lead, others follow.

Case Three

One of the recent Mandiant papers concerns a tool they have called ‘Hammertoss' ("HAMMERTOSS: Stealthy Tactics Define a Russian Cyber Threat Group," 2015) Mandiant attribute ‘Hammertoss' to ‘APT29', a group believed to be a Russian Government team. The malware is described as being a binary attack. That means the attack is not complete until both initialization code and the actual attack are present inside the victim organisation's defences. They are inserted independently of one another. Interestingly, the latter element is delivered as a steganographically-encoded malware image. The malware is decoded by a Javascript fragment cached previously, which is activated by a string from an external source matching a value computable by the script. It collects a decoding value, extracts the malware from the image and triggers the attack. The actions performed by the script appear to be a simple end user lookup to a popular social site like github or twitter.

This is a significant development in the history of steganographic malware, because it has been found in the wild by a significant player in the security field and uses a mechanism that would defeat the vast majority of currently deployed technologies. It extends the repertoire of those responsible for APTs and may be the vanguard of a new class of attacks.

Mandiant is a unit of Fireye, a "rising star" of the security industry. The publicity they gained during the Target compromise was widespread and favourable. Indeed, the fact that a Fireye product raised alerts to Target support personnel about the ongoing intrusion the moment it occurred, was one of the most telling points in the subsequent post compromise review (Riley, Elgin, Lawrence, & Matlack, 2014). The fact that no-one reacted to those alerts may be a significant factor in the downfall of the then CEO. Mandiant have continued to publish research papers relating to previously unknown attacks and attackers, raising awareness of the torrent of novel attacks pouring onto the Net. In addition to expanding awareness of the new methods in use, they highlight the insidious nature of these stealthy intrusions. This reinforces the view that current preventive measures are worthless against a skilled and persistent attack. The current crops of defensive tools no longer suffice. They must be replaced by more effective controls.

As has happened with ‘malvertising' in recent years, improvements in detection by defenders are blunted by changes of tactics by attackers. Attackers are traditionally seen as competing for ‘contested ground' (Tzu, ?), avoiding open assault, instead choosing methods that play to their own strengths. This reveals the nature of Internet defences in clear detail. In 1984, the IRA said, "Today we were unlucky, but remember we only have to be lucky once. You will have to be lucky always" (Taylor, 2002) with reference to the ‘Brighton bombing' which nearly destroyed the U.K. Government. It epitomizes the asymmetric situation that attackers seek. The anonymity provided by the Internet linked to the visibility and stability prized by those promoting an organization means that true parity for defenders is unlikely to exist. On the Internet, you can't dodge the bullets

or you lose your followers. To ensure longevity for an Internet based organization, the Target example suggests investment in modern detection methods will be a source of value. This will be true so long as incident response is properly resourced. It seems likely that we are about to find plenty of obscurely hidden attacks leaking past our traditional tools. ‘Hammertoss' may be the first public indication that attackers are seeking more productive methods.

EXTRAPOLATION

Steganography is considered too expensive to defend against, because reliable methods for detecting it do not exist. Absence of demonstrable assurance leads many organisations to conclude that defending against steganography is a waste of resources. This was recently confirmed in an Intel Malware Security group newsletter (Intel, 2015). A statement in the newsletter relating to use of images that contain an extra steganographic payload which "doesn't affect the content itself, it is hard to distinguish and difficult to detect" simply states the implied threat that malware authors use of steganography may present. It's hard to defend against and no fully assured detection exists. The examples above would indicate that there are perceptible moves by the attacker community to remain "ahead of the game" when the industry is moving to Cloud based services. This evidence implies that the attacker community pays close attention to reactions by defenders and looks forward to likely changes in behaviour of their true targets - naïve end users. Examples like the compromised businesses discussed above make clear that absence of risk appetite in commercial organisations presents a great opportunity for attackers prepared to invest in novel methods. End users who have been conditioned to use images in high volume uses (e.g. Facebook, Snapchat etc.) add to business incentive for "giving a bye" to "passive content". The sheer volume of images handled by most conventional organisations in using Web based services may make any realistic defences economically infeasible. Until such time as a practical method to reliably detect hidden content is available advantage lie with attackers. For many businesses, the window of opportunity presented to hostile players may not only lose them board members, it may include the need to pay enough compensation to harmed customers that the affected companies could fold.

Whether the content is hidden in LSBs, Frequency Domains or any other organised encoding artefact, keeping long lasting integrity of malware would be a real concern for attackers. Instead of actively spreading through a population like a worm or virus, passive spread through user behaviour would enhance attacks. Attackers would keep pressure on maintaining positive detection rather than attracting attention to their tools. This implies that stealthy content must survive common image operations performed in Web services (such as format conversion, re-sizing, cropping etc.) One of the obvious needs for an attacker travelling this route would be something that can restore the integrity of their product in the event that an impaired payload is placed on an important target. This may be the crucial distinction between previous attempts to provide an "all-in-one" payload and the steganographic successes of the future. A binary attack might bring an "on-demand" botnet within the grasp of attacker communities. A major problem with many attacks is that once they are in place, they are detectable per se. Standard defensive mechanisms can then be prepared to detect and eradicate them. Even with the recent proliferation of polymorphic and overwhelmingly individualised payloads, traditional Vendors have not been displaced from their market. Once a situation where attacks can be hidden "in plain sight" and target systems can be brought within grasp of an anonymised controller who can control a botnet on no notice exists, then any "clean bill of health" for a system can be true only for a very short period of time. With the ability to produce a payload inside a trust boundary with only "normal" Web traffic necessary to activate it by methods that must be discovered in advance of the threat being deployed, the reaction time of defenders is minimised and the effort expended to uncover threats is raised. This is not a recipe for minimised cost. If organisations are not prepared to build effective defences for novel threats, they will need to budget for increased levels of incident response or, perhaps, die. A side effect of this scenario may be that the life span of a "zero-day" exploit is extended, because once a successful compromise is in place, only those who have prepared to deal with novel threats may be able to uncover them. The investment made by attackers in the tools of their trade may be more easily defensible, under those circumstances. Perhaps the idea is to avoid having to buy so much expensive attack code, whilst remaining effective in compromising target organisations. Maybe it's purely to commodify intrusion to the point that any unscrupulous individual can run a profitable ransomware campaign. Whatever the motive, it seems that a level playing field for defenders remains a distant dream.

CONCLUSION

The major conclusion we can draw from these examples is that steganography has not lost its persistent attraction. Its appeal to cyber attackers has not grown less over time. Indeed, the fascination is still strong for steganographic methods, particularly amongst those engaged in finding ways to avoid or misdirect defender reactions. Part of this fascination is probably innate. The simple, childlike joy found in magicians or ‘trompe

l'oeuil' artefacts is reflected in the attention paid to a simple mathematical hallucination. Steganography is easily accessible to ordinary people. Whether for preserving confidentiality in the exchange of messages or in smuggling malware past vigilant defences, steganography presents a tangible challenge to defenders. This makes it worthy of serious consideration in planning. Particularly when the blind spot for "passive content" threatens to provide an entirely new field of attacks against systems already being moved to more available platforms, ignoring novel threats in plans may be fatal. We would be wise to consider developing effective ways to detect malware that arrives as an image. A likely consequence will be a need to scale up investment in hardware to reduce latency when crossing boundaries. The average acceptable wait for content is still four seconds. (Strange how that doesn't increase?) So, if we are to continue defending against old threats, as well as commissioning capability to deal with novel threats, then more machinery, resources and time will have to be allocated. Most of it will be devoted to improving reactions against previously unknown threats. The question becomes how we do that?

REFERENCES

Al-Ani, Z. K., Zaidan, A. A., Zaidan, B. B., & Alanazi, H. O. (2010). Overview: Main Fundamentals for Steganography. *CoRR, abs/1003.4086*. Retrieved from http://arxiv.org/abs/1003.4086

Alintanahin, K. (2015, May 14). Operation Tropic Trooper. Retrieved from http://www.trendmicro.com/cloud-content/us/pdfs/security-intelligence/white-papers/wp-operation-tropic-trooper.pdf

Basu, E. (2015, June 15). Target CEO Fired - Can You Be Fired If Your Company Is Hacked? *Forbes*. Retrieved from http://www.forbes.com/sites/ericbasu/2014/06/15/target-ceo-fired-can-you-be-fired-if-your-company-is-hacked/

Cyber and Technology Enabled Crime. (2015, October 18). [Government Agency]. Retrieved from https://www.crimecommission.gov.au/publications/intelligence-products/crime-profile-fact-sheets/cyber-and-technology-enabled-crime

Fabien, R. A. (1998). *On The Limits of Steganography*.

Grimes, R. A. (2011). Fixing the Internet: A Security Solution.

HAMMERTOSS: Stealthy Tactics Define a Russian Cyber Threat Group. (2015, July 29). [Vendor]. Retrieved September 28, 2015, from https://www2.fireeye.com/rs/848-DID-242/images/rpt-apt29-hammertoss.pdf

Intel, S. G. (2015, October 29). MSO Newsletter [Vendor Blog]. Retrieved from https://community.mcafee.com/groups/global-malware-support-operations/blog/2015/10/29/malware-support-operations-newsletter-q3-2015

Provos, N., & Honeyman, P. (2001). *Detecting Steganographic Content on the Internet*. In ISOC NDSS'02.

Riley, M., Elgin, B., Lawrence, D., & Matlack, C. (2014, March 13). Missed Alarms and 40 Million Stolen Credit Card Numbers: How Target Blew It. *Bloomberg Features*. Retrieved from http://www.bloomberg.com/bw/articles/2014-03-13/target-missed-alarms-in-epic-hack-of-credit-card-data

Shah, S. (2015). Stegosploit: Hacking with Pictures. In *6th Annual HITB Security Conference in The Netherlands*. De Beurs van Berlage, Netherlands. Retrieved from https://conference.hitb.org/hitbsecconf2015ams/wp-content/uploads/2015/02/D1T1-Saumil-Shah-Stegosploit-Hacking-with-Pictures.pdf

Sujatha, P., Purushothaman, S., & Rajeswari, R. (2014). Performance Study of Combined Artificial Neural Network Algorithms for Image Steganalysis. In S. Sathiakumar, L. K. Awasthi, M. R. Masillamani, & S. S. Sridhar (Eds.), *Proceedings of International Conference on Internet Computing and Information Communications* (Vol. 216, pp. 441-451). Springer India. Retrieved from http://dx.doi.org/10.1007/978-81-322-1299-7_41

Taylor, P. (2002). *The war against the IRA*. London: Bloomsbury Publishing PLC : [distributor] Macmillan Distribution (MDL).

Tzu, S. (2012). *The art of war*. e-artnow.