Revisiting the Limits of Steganography

Rainer Böhme

University of Innsbruck, Austria

Abstract. I select five key ideas from Ross Anderson's first paper on steganography and show how they have influenced the state of the art. The ideas are: content-adaptive steganography, the selection channel, the diminishing secure rate, generative steganography, and epistemic limits.

Keywords: Information Hiding · Steganography · Ross Anderson.

1 Hiding and Freedom

In 1996, Ross Anderson brought together researchers from five different subfields who had a common interest in hiding some information in other data (or noise) to achieve a security objective. This marked the beginning of a series of first bi-annual, then annual workshops on "Information Hiding" (IH), which continue to this day. The 2024 edition was held in Baiona, Spain and the 2025 edition will be held in San Jose, California. Incidentally, the forerunner of PETS[1] was also born out of this workshop series. Ross's own contribution to the first edition of IH was a short essay on "Stretching the Limits of Steganography" [2]. An extended journal version [3], co-authored with his student Fabien A. Petitcolas, is now Ross's third most cited paper.

Among other things, digital steganography was of interest in the first "crypto wars" [1] as an argument against restricting the use of cryptography. If a technology exists that allows undetectable secret communication, it becomes pointless to enforce laws that prohibit secret communication. Many of the cutting-edge technologies discussed at the workshop have societal implications. Sender anonymity supports freedom of expression, while watermarking for digital rights management enables applications that may deprive users of their autonomy.

For this piece, I re-read Ross's original work on steganography and point out striking technical insights that have led to the development of whole strands of literature, as well as some "rediscoveries," that are now linked to Ross's ideas. His 1996 paper appears surprisingly fresh after almost 30 years, and I would like to share some of my observations with the readers.

2 Groundbreaking Ideas for Digital Steganography

Recall that digital steganography aims to hide a secret message undetectably in inconspicuous cover objects. The sender and recipient share a key, and no

[1] Privacy Enhancing Technologies Symposium, https://petsymposium.org

one else should be able to tell whether an object contains any secret message or not [24]. The task is to find or modify an object so that it contains the message under the key and is indistinguishable from typical objects on the channel.

2.1 Content-Adaptive Steganography

Since finding a stego object by sampling quickly becomes inefficient for larger messages, the common approach for the sender is to sample a single cover object and modify it carefully. To reduce the risk of detection, local modifications should take into account the surrounding content. Ross described this for images in the spatial domain:[2]

> "Of course, not every pixel may be suitable for encoding ciphertext: changes to pixels in large fields of monochrome colour, or that lie on sharply defined boundaries, might be visible. So some systems have an algorithm that determines whether a candidate pixel can be used [...]"

All relevant embedding function today build on this observation and are content-adaptive. Nowadays suitability is no longer binary. Researchers are developing and comparing distortion functions that approximate the effect on detectability of changing an element in the cover [22]. What remains challenging is dealing with non-additive distortion, for example in cases where two or more pixels should better be changed together or not at all [23].

2.2 Selection Channel

Content-adaptive embedding makes the extraction more difficult. How does the recipient know where to look for the message? Simply applying the same suitability metric may not be successful, as the result may not be the same for the received object that has been modified during embedding. Ross's idea of a selection channel was to find an encoding that would not even require the recipient to know where the embedding was taking place:

> "We will use our keystream generator to select not one pixel but a set of them, and embed the ciphertext bit as their parity. This way, the information can be hidden by changing whichever of the pixels can be changed least obtrusively."

This idea became a game changer when it was generalized in *wet paper coding* [11], and later combined with syndrome coding [9]. In Ross's original scheme, it was difficult to find the right set size k. If it was too large, the capacity was reduced to $1/k$ of the available elements. If you make it too small, you increase the risk that at least one of the many sets doesn't have a good option for making an embedding change, and you get caught. Codes with low-density parity check matrices allow for larger overlapping sets, but require the sender to solve a system of equations. Doing this efficiently [10] while satisfying statistical properties of the change vector [17] is an open problem.

[2] For completeness, the idea can be found in an earlier (German) source [19] for audio, but the concept is less general there and the source is not not widely available.

2.3 Diminishing Secure Rate

Ross also reflected on the secure capacity of a channel. He had the right intuition that steganography should not be thought of as a one-shot game, because the adversary accumulates evidence.

> "Thanks to the Central Limit Theorem, the more covertext we give the warden, the better he will be able to estimate its statistics, and so the smaller the rate at which [the sender] will be able to tweak bits safely. The rate might even tend to zero."

The result that every (marginally) imperfect sender will be caught in the long run, and thus the secure rate is zero, was formalized in the *square root law* [14], first for the asymptotic case of $n \to \infty$ independent objects, and later empirically established for objects of varying sizes [16]. In the best case, a sender who can choose a cover object of size n must limit the number of embedding changes proportional to $\sqrt{n}$ in order to keep the risk of detection constant. The applicability of this law to content-adaptive embedding is an open problem [15].

Conversely, a strictly positive secure rate can only be attained if the steganography is perfect, i. e., the distributions of the cover and stego objects are identical.

2.4 Generative Steganography

There are channels where this is possible in principle because the distribution of objects is defined, e. g., by a generative language model or a generative adversarial network [12]. If you assume such a channel, it becomes possible to make stego objects indistinguishable from covers and thus attain a positive secure rate [13]. But why would such channels exist in the first place? Wouldn't it be easier to exchange the latent space of the model and 'decompress' it at the recipient's end? Ross clearly saw the link between perfect compression and undetectability:

> "Information theorists assume that any signal can in theory be completely compressed. But if this could ever be done in practice, then the steganography problem would become trivial: [The sender] can just 'uncompress' her ciphertext getting a comprehensible message, and [the adversary] would have to pass the result."

With recent developments in learned large language models and neural image compression [4], it may be within reach to iterate over the values of the latent space in order to generate objects close enough to the channel distribution to be indistinguishable from real objects. Similarly, setting the latent space to the (encrypted) message should result in a secure stego object. What remains a challenge is to exactly retrieve the latent space from the generated object, as the generation involves floating-point operations and rounding losses. Until this problem is solved, coding is required to make the message extractable [20].

2.5 Epistemic Limits

The assumption of a channel with a defined distribution, even if it is encoded in an incomprehensible way in billions of trained parameters, is arguably an escape from solving a steganographic problem.[3] One could also imagine a channel where mathematicians exchange random numbers, so any encryption scheme would provide secure steganography in this channel.

Ross acknowledged that the channel distribution is not under our control and is generally not fully understood. What is known about it needs to be captured in models:

> "Performance of [the adversary's] job depends on his having a model of the source, and the danger to Alice and Bob is that his model might be better than theirs."

Today's models are inferred from data. Detecting stego objects with learned classifiers was proposed in 2003 [18] and is now the standard. Machine learning is also increasingly used on the sender's side [6]. The race for the better (trained) model turns into a race for access to more and better training data, and for efficient ways to closely approximate the underlying distributions. This is a problem common to many fields, most notably machine learning.

One difference is that approximation errors are not just "challenging cases" to be buried in supplementary material, but security vulnerabilities. When discovered, the system is "broken." Good designs provide evidence of their absence.

To me, this shows how fundamental steganography research is. It is about the ability to decide on hypotheses, to learn about reality from incomplete observations, including understanding what information is lost during processing, and to do all this efficiently and, if possible, with guarantees. Because every gap gives an advantage to the adversary. Moreover, the scope is not limited to message exchange. Making something artificial indistinguishable from something real appears in many corners of security [21]. Steganography remains fascinating.

3 Concluding Remarks

In this area, as in many others, Ross did what he liked best, doing groundbreaking research "with shovels." This paved the way for others to fill in the details "with pincers," including myself with a dissertation on the epistemic limits of steganography [7, 8]. Ross also brought people together and created a community. I recommend that readers attend a future edition of the (now) *ACM Workshop on Information Hiding and Multimedia Security*.[4]

Working on this contribution to the Festschrift has reminded me of the value of reading original work. I encourage all researchers to trace ideas back to their source by following the citation trail (and finding ways to fill in the gaps).

[3] The GPT-2 channel assumed in [13] was distinguishable from real text at the time the paper was written: https://huggingface.co/openai-detector/ (accessed: July 2021). Neural compression can be distinguished from conventional compression [5].

[4] https://www.ihmmsec.org/

References

1. Abelson, H., Anderson, R., Bellovin, S.M., Benalob, J., Blaze, M., Diffie, W., Gilmore, J., Neumann, P.G., Rivest, R.L., Schiller, J.I., Schneier, B.: The risks of key recovery, key escrow, and trusted third-party encryption. World Wide Web Journal **2**(3), 241–257 (1997)
2. Anderson, R.J.: Stretching the limits of steganography. In: Anderson, R.J. (ed.) Information Hiding (1st International Workshop). Lecture Notes on Computer Science, vol. 1174, pp. 39–48. Springer (1996)
3. Anderson, R.J., Petitcolas, F.A.P.: On the limits of steganography. IEEE Journal on Selected Areas in Communications **16**, 474–481 (1998)
4. Ballé, J., Minnen, D., Singh, S., Hwang, S.J., Johnston, N.: Variational image compression with a scale hyperprior. In: International Conference on Learning Representations (ICLR). OpenReview.net (2018), https://openreview.net/forum?id=rkcQFMZRb, (accessed: December 2024)
5. Bergmann, S., Moussa, D., Brand, F., Kaup, A., Riess, C.: Forensic analysis of AI-compression traces in spatial and frequency domain. Pattern Recognition Letters **180**, 41–47 (2024)
6. Bernard, S., Bas, P., Klein, J., Pevný, T.: Backpack: A backpropagable adversarial embedding scheme. IEEE Transactions on Information Forensics and Security **17**(9), 3539–3554 (2022)
7. Böhme, R.: Improved statistical steganalysis using models of heterogeneous cover signals. Ph.D. thesis, Technische Universität Dresden, Department of Computer Science, Dresden, Germany (2008)
8. Böhme, R.: An epistemological approach to steganography. In: Katzenbeisser, S., Sadeghi, A.R. (eds.) Information Hiding (IH). Lecture Notes in Computer Science, vol. 5806, pp. 15–30. Springer (2009)
9. Crandall, R.: Some notes on steganography. Mimeo posted to a mailing list (1998), online available at http://dde.binghamton.edu/download/Crandall_matrix.pdf (accessed: November 2024)
10. Filler, T., Judas, J., Fridrich, J.: Minimizing additive distortion in steganography using syndrome–trellis codes. IEEE Transactions on Information Forensics and Security **6**(3-2), 920–935 (2011)
11. Fridrich, J., Goljan, M., Soukal, D.: Perturbed quantization steganography with wet paper codes. In: Dittmann, J., Fridrich, J.J. (eds.) ACM Multimedia and Security Workshop (MM&Sec). pp. 4–15. ACM (2004)
12. Hayes, J., Danezis, G.: Generating steganographic images via adversarial training. In: Advances in Neural Information Processing Systems. pp. 1954–1963 (2017)
13. Kaptchuk, G., Jois, T.M., Green, M., Rubin, A.D.: Meteor: Cryptographically secure steganography for realistic distributions. In: Kim, Y., Kim, J., Vigna, G., Shi, E. (eds.) ACM Conference on Computer and Communications Security (CCS). pp. 1529–1548. ACM (2021)
14. Ker, A.: Batch steganography and pooled steganalysis. In: Camenisch, J., Collberg, C., Johnson, N., Sallee, P. (eds.) Information Hiding (IH). Lecture Notes in Computer Science, vol. 4437, pp. 265–281. Springer (2007)
15. Ker, A.: The square root law of steganography: Bringing theory closer to practice. In: Stamm, M.C., Kirchner, M., Voloshynovskiy, S. (eds.) ACM Workshop on Information Hiding and Security Workshop (IH&MMSec). pp. 33–44. ACM (2017)
16. Ker, A., Pevný, T., Kodovský, J., Fridrich, J.: The square root law of steganographic capacity. In: Ker, A., Dittmann, J., Fridrich, J. (eds.) ACM Multimedia and Security Workshop (MM&Sec). pp. 107–116. ACM (2008)

17. Köhler, O.M., Pasquini, C., Böhme, R.: On the statistical properties of syndrome trellis coding. In: Krätzer, C., Shi, Y.Q., Dittmann, J., Kim, H.J. (eds.) Digital Forensics and Watermarking (IWDW). Lecture Notes in Computer Science, vol. 10431, pp. 331–346. Springer (2017)
18. Lyu, S., Farid, H.: Detecting hidden messages using higher-order statistics and support vector machines. In: Petitcolas, F.A.P. (ed.) Information Hiding (IH). Lecture Notes in Computer Science, vol. 2578, pp. 340–354. Springer (2003)
19. Möller, S., Pfitzmann, A., Stierand, I.: Rechnergestützte Steganographie: Wie sie funktioniert und warum folglich jede Reglementierung von Verschlüsselung unsinnig ist *[Computer-based steganography: how it works and why any regulation of encryption is therefore nonsensical]*. Datenschutz und Datensicherung **18**(6), 318–326 (1994)
20. Nakajima, T., Ker, A.D.: The syndrome-trellis sampler for generative steganography. In: IEEE Workshop on Information Forensics and Security (WIFS). IEEE (2020)
21. Pasquini, C., Schöttle, P., Böhme, R.: Decoy password vaults: At least as hard as steganography? In: di Vimercati, S.D.C., Martinelli, F. (eds.) ICT Systems Security and Privacy Protection (IFIP SEC). IFIP Advances in Information and Communication Technology, vol. 502, pp. 327–340. Springer (2017)
22. Pevný, T., Filler, T., Bas, P.: Using high-dimensional image models to perform highly undetectable steganography. In: Böhme, R., Fong, P., Safavi-Naini, R. (eds.) Information Hiding (IH). Lecture Notes in Computer Science, vol. 6387, pp. 161–177. Springer (2010)
23. Pevný, T., Ker, A.D.: Exploring non-additive distortion in steganography. In: Böhme, R., Pasquini, C., Boato, G., Schöttle, P. (eds.) ACM Workshop on Information Hiding and Multimedia Security (IH&MMSec). pp. 109–114. ACM (2018)
24. Simmons, G.J.: The prisoners' problem and the subliminal channel. In: Chaum, D. (ed.) Advances in Cryptology, Proceedings of CRYPTO '83. pp. 51–67. Plenum Press (1984)

Acknowledgements

I would like to thank Benedikt Lorch for useful comments on a draft of this work, and the organizing team, especially Frank Stajano, for their efforts in putting together the Rossfest Symposium.