Concealing Content and Concealing Existence: Encryption, Steganography, and Circuit Models of Hidden Structure

Flyxion
Independent Researcher

October 2026

Abstract

Encryption conceals what a message says, while steganography conceals that a message exists. The two goals are routinely treated as points on a single scale of secrecy, with steganography as the weaker, more informal cousin. This essay argues that they are different security properties defined against different adversaries, that neither implies the other, and that the difference is most clearly seen when both are expressed as circuit families. An encryption circuit is judged against the uniform distribution, which every observer can construct. A steganographic embedding circuit is judged against a channel distribution that the embedder must be able to sample and that the adversary is assumed to model. The same distinction reappears in physical circuit design, where logic locking behaves like encryption of a function and where watermarks and hardware trojans behave like steganography in the space of functionally equivalent netlists. The essay states the formal definitions, proves two separation results and a composition result, and traces the consequences for hardware. It does not claim to resolve the open problems it identifies, and several of them are stated as such.

1 Two Kinds of Secrecy

A message can be protected in two ways that are easily confused. It can be transformed so that an observer who knows the message exists learns nothing about its content, or it can be embedded in innocuous traffic so that an observer does not learn that anything was sent. Shannon's treatment of secrecy systems addressed the first problem and left the second outside the model [10]. Simmons posed the second in the form of the prisoners' problem, in which two prisoners must coordinate an escape through a warden who reads every message and will block any communication that looks like coordination [11]. The warden in that story does not need to decrypt anything. Suspicion is sufficient, and a perfectly encrypted but visibly encrypted message fails.

The distinction would be of limited interest if one property could be obtained from the other by a simple construction. It cannot, and the reason is structural. Content secrecy is a statement about the relationship between the plaintext and the ciphertext, and it can be established with no reference to any environment in which ciphertexts appear. Existence secrecy is a statement about the relationship between the transmitted object and the population of objects that would have been transmitted anyway. It cannot be stated without naming that population.

Every formal treatment of steganography must therefore introduce a distribution that has no counterpart in the definition of encryption, and every difficulty particular to steganography descends from that fact.

Circuit models make this visible because a circuit is a concrete, costed object. An encryption scheme is a pair of circuit families whose size and depth are fixed by the key and message lengths. A steganographic embedder is a circuit that must also contain, or have oracle access to, a sampler for the channel, and its cost depends on the entropy and the sampling cost of that channel. Section 2 gives the definitions. Section 3 proves that encryption does not yield steganography and that the two compose in only one order. Section 4 examines rate and cost. Section 5 turns to physical circuit design, where the same distinction governs logic locking, watermarking, and trojan insertion. Section 6 collects the limits of the analysis.

2 Definitions

Security parameters are written n. All adversaries are probabilistic polynomial time unless stated otherwise, and $\mathsf{negl}(n)$ denotes a function smaller than every inverse polynomial.

Definition 1 (Encryption)**.** A symmetric encryption scheme is a triple $(\mathsf{Gen}, \mathsf{Enc}, \mathsf{Dec})$ of efficient algorithms with $\mathsf{Dec}_k(\mathsf{Enc}_k(m)) = m$. It has *pseudorandom ciphertexts under chosen plaintext attack* if no efficient adversary with oracle access to $\mathsf{Enc}_k(\cdot)$ distinguishes that oracle from an oracle returning, on each query m, a uniformly random string of length $|\mathsf{Enc}_k(m)|$, except with advantage $\mathsf{negl}(n)$.

This is a standard strengthening of semantic security [5]. It is used here because the comparison with steganography is cleanest when the ciphertext distribution is explicit and uniform.

Definition 2 (Channel)**.** A channel $\mathcal{C}$ is a family of distributions over documents, conditioned on a history of previously emitted documents. The embedder and the warden are both given sampling access to $\mathcal{C}$. The channel is *efficiently sampleable* if there is a polynomial-time algorithm that, given a history, outputs a document distributed as $\mathcal{C}$ conditioned on that history.

Definition 3 (Stegosystem and its security)**.** A stegosystem is a pair $(\mathsf{SE}, \mathsf{SD})$ of efficient algorithms with oracle access to $\mathcal{C}$ and a shared key k. SE_k takes a payload m and a history and returns a sequence of documents. SD_k recovers m from that sequence with overwhelming probability. The system is *secure against a warden* W if

$$\left|\Pr[W^{\mathcal{C}}(1^n) = 1] - \Pr[W^{\mathsf{SE}_k(\cdot)}(1^n) = 1]\right| \le \mathsf{negl}(n),$$

where in the second experiment the warden sees documents produced by embedding payloads of the warden's choosing, and in the first it sees documents drawn from the channel alone.

This follows the computational formulation of Hopper, Langford, and von Ahn [6], which refines Cachin's information-theoretic definition, in which perfect security means that the relative entropy between the channel distribution and the stego distribution is zero [2]. In both, the benchmark for indistinguishability is $\mathcal{C}$ rather than the uniform distribution. That single substitution is the formal content of the claim that existence secrecy is a different property from content secrecy.

3 Separation and Composition

Proposition 1 (Encryption does not give steganography)**.** *Let $\mathcal{C}$ emit n-bit documents, and suppose there is an efficiently decidable set $S \subseteq \{0,1\}^n$ with $|S| \leq 2^{n/2}$ and $\Pr_{\mathcal{C}}[d \in S] \geq 1/2$. Then for any encryption scheme with pseudorandom ciphertexts of length n, a warden that outputs 1 exactly when the observed document lies in S distinguishes ciphertexts from the channel with advantage at least $1/2 - 2^{-n/2} - \mathsf{negl}(n)$.*

Proof. Under the channel, the warden outputs 1 with probability at least $1/2$. A uniform string lies in S with probability at most $|S|/2^n \leq 2^{-n/2}$. By pseudorandomness of ciphertexts, a ciphertext lies in S with probability at most $2^{-n/2} + \mathsf{negl}(n)$, since membership in S is efficiently decidable and therefore itself a distinguisher. The advantage is the difference of the two probabilities. □

The set S stands for any structure that real documents have and random strings lack: a valid file header, grammatical text, a plausible instruction sequence, a netlist that parses. The proposition is elementary, but its consequence is not. Strengthening the encryption scheme does not help, because the failure lies in the distribution against which ciphertexts are compared. A scheme can be perfectly secret in Shannon's sense and perfectly conspicuous.

Proposition 2 (Steganography does not give confidentiality)**.** *Suppose SD is keyless, so that anyone who holds the documents can run it. Then the system may be secure in the sense of Definition 3 and yet provide no confidentiality against a party who merely suspects that a payload is present.*

Proof. Security in Definition 3 constrains only a warden who must distinguish stego documents from channel documents. A party who obtains the documents by other means, or who runs SD on every document as a matter of routine, recovers m at no cost. Nothing in Definition 3 forbids this, because the definition says nothing about what a party learns once existence is granted. □

The two propositions together show that neither property subsumes the other. The following result shows how they combine.

Theorem 1 (Composition)**.** *Let $(\mathsf{Gen}, \mathsf{Enc}, \mathsf{Dec})$ have pseudorandom ciphertexts and let $(\mathsf{SE}, \mathsf{SD})$ be secure for payloads drawn uniformly at random from $\{0,1\}^\ell$, where ℓ is the ciphertext length. With independent keys k_1, k_2, the system that transmits $\mathsf{SE}_{k_2}(\mathsf{Enc}_{k_1}(m))$ is a secure stegosystem. If in addition a party learns k_2 or learns that SD is keyless, the payload remains protected by the chosen-plaintext security of Enc_{k_1}.*

Proof. Two hybrids suffice. In the first, $\mathsf{Enc}_{k_1}(m)$ is replaced by a uniform string of the same length. The change is undetectable by pseudorandomness of ciphertexts. In the second, the stegosystem is applied to this uniform string, and its output is indistinguishable from channel output by the assumed security of $(\mathsf{SE}, \mathsf{SD})$ on uniform payloads. The confidentiality claim follows because a party who can run SD obtains only $\mathsf{Enc}_{k_1}(m)$. □

Remark 1. The order matters. Embedding first and encrypting afterwards produces a pseudorandom string, and by Proposition 1 such a string is conspicuous on any channel with structure. Encryption belongs inside the embedding because the embedding step requires a

uniform payload as its input distribution, which is exactly what pseudorandom ciphertext provides. This is the sense in which encryption serves steganography as a preprocessing step, and it explains why practical hiding systems run encryption first even though encryption alone is never sufficient.

4 Cost, Rate, and the Sampler

The cleanest known stegosystem is the rejection-sampling construction of Hopper, Langford, and von Ahn [6]. To embed one bit b in one document, the embedder draws up to λ documents d from the channel until a keyed pseudorandom function satisfies $F_k(d) = b$, and emits the last draw. If F_k is balanced on the support of the channel, each draw succeeds with probability one half, the expected cost is two samples per bit, and the output distribution differs from the channel by a term that falls exponentially in λ plus the bias of F_k on the channel. In circuit terms the embedder is a depth-λ loop whose body contains one copy of the sampler and one copy of F_k. Its size is therefore $\lambda(|\mathcal{S}| + |F|)$, where $|\mathcal{S}|$ is the size of the sampler, and the sampler term typically dominates.

Two quantities govern what can be achieved. The first is the min-entropy of the channel per document, $H_\infty(\mathcal{C})$. A channel with no entropy admits no embedding, because the only document the embedder can emit is the one that was going to be emitted anyway. The second is the cost of sampling. In the black-box setting, in which the embedder can only draw from the channel, the number of samples required grows exponentially in the number of bits embedded per document [3]. Higher rates require access to the channel's structure, such as the ability to sample conditioned on a prefix, and this is where the channel stops being a mere oracle and becomes a component of the construction.

Encryption has no corresponding term. An encryption circuit has size polynomial in message length, rate arbitrarily close to one up to a nonce, and no dependence on any external distribution. The comparison is summarized in Table 1.

	Encryption	Steganography	Physical analogue
Conceals	Content	Existence	Function (locking); presence (watermark, trojan)
Benchmark distribution	Uniform	Channel $\mathcal{C}$	Tool output or test distribution
Rate limit	Near 1	$H_\infty(\mathcal{C})$ per document	Entropy of equivalent implementations
Extra cost	None	Sampler for $\mathcal{C}$	Synthesis or fabrication flow
Failure mode	Key compromise	Channel mismatch	Tool fingerprint, inspection

Table 1: Encryption and steganography compared, with the corresponding objects in circuit design.

The last row of the table names the principal practical hazard. Because the benchmark is the channel, the warden's model of the channel determines what counts as a deviation. An embedder that samples correctly with respect to one model of the channel can be conspicuous

with respect to a better one. Kerckhoffs's principle applies to the key and the algorithm [8], but it cannot be extended to the channel, which is a fact about the world rather than a design choice, and which a capable warden will usually know better than the embedder does.

5 Physical Circuit Design

Physical design supplies the setting in which these abstractions acquire concrete counterparts. Three practices are considered: logic locking, watermarking, and trojan insertion.

5.1 Logic locking as encryption of a function

A locked circuit has a function $L(x, k)$ such that $L(x, k^*) = f(x)$ for a correct key k^* and $L(x, k) \neq f(x)$ for many incorrect keys [9]. The designer's aim is that a holder of the netlist, who does not hold k^*, cannot recover f. This is encryption of a function. The locking is overt, since the key inputs are ports on the netlist, and no attempt is made to hide that a key exists. The relevant adversary model is correspondingly the one of Definition 1, with the added complication that the adversary may hold a working unlocked chip as an oracle. Under that stronger model, satisfiability-based attacks recover keys for many locking schemes in a small number of oracle queries [12], which mirrors the familiar lesson that encryption security depends on the access model and not only on the construction.

Obfuscation theory sharpens the point. Virtual black-box obfuscation, which would make a circuit reveal nothing beyond its input-output behavior, is impossible for general program families [1]. Indistinguishability obfuscation offers a weaker guarantee, that two functionally identical circuits of the same size have indistinguishable obfuscations [4]. It follows that obfuscation can conceal *which implementation* of a function was supplied, but it provides no protection when the two circuits differ even on a single input. This limit matters in the discussion of trojans below.

5.2 Watermarking as steganography over equivalent netlists

Constraint-based watermarking embeds an ownership signature in the free choices that a synthesis or placement procedure leaves open [7]. In the vocabulary of Section 2, the cover is the set $N(f)$ of netlists realizing a function f within given area and delay constraints, the channel is the distribution with which the design flow selects from $N(f)$, and the payload is a signature recovered by a keyed rule applied to the chosen element.

Seen this way, the capacity and security of a watermark are controlled by the same quantities as in Section 4. The raw capacity is bounded by $\log |N(f)|$, but the secure capacity is bounded by the min-entropy of the flow's output distribution. A deterministic synthesis tool has a point-mass output distribution on a given source description. Against a warden who holds the source and the tool, the channel has no entropy, and any watermark is detectable by rerunning the tool and comparing. Against a warden who holds only the finished netlist, the channel is the distribution over source descriptions the warden considers plausible, and its entropy is large. The warden's knowledge therefore determines the secure capacity of an identical embedding, which is a concrete instance of the channel-dependence emphasized above and one that has no analogue in encryption.

5.3 Trojans as keyed deviations from a test distribution

A hardware trojan is a circuit C' that agrees with a specified circuit C except on a trigger set K [13]. Under black-box functional testing with inputs drawn from a distribution D, the probability that a test hits the trigger set is $\Pr_D[x \in K]$. A trigger implemented as a k-bit comparator on otherwise uniform inputs fires with probability 2^{-k}, so T random test vectors detect it with probability at most $T2^{-k}$. Choosing k somewhat larger than $\log T$ renders the trojan invisible to the test regime while leaving it fully controllable by the party who knows the trigger.

The structure is the keyed-embedding structure of Definition 3, with the test distribution D playing the role of the channel and the trigger playing the role of the key. The analogy also marks the point at which the black-box abstraction fails. A warden restricted to sampling the channel cannot see the structure of the sampler, but a physical inspector can examine layout, timing, power draw, and structural features that lie outside any input-output distribution. Obfuscation cannot close this gap, because as noted above indistinguishability obfuscation does not apply when the circuits differ functionally. The correct statement is that trojan security in the black-box model is a matter of rare events, while trojan security against white-box inspection is a different and unresolved problem.

6 Limits and Open Problems

Several limits qualify the analysis. The composition theorem assumes independent keys and a payload that is uniform at the point of embedding, and it says nothing about robustness against an active warden who modifies the transmitted documents. Robust steganography is a harder problem with a smaller set of positive results, and it is not addressed here. The treatment of channels as sampleable distributions idealizes natural data, for which the true distribution is unknown and which no embedder samples exactly. The assumption of efficient sampleability is therefore an assumption about the relationship between the embedder's model and the warden's model, and the proofs above give no guidance on how closely the two must agree.

The physical analogies are structural rather than quantitative. The equivalence between netlist watermarking and steganography over a channel of equivalent implementations is an equivalence of form, and measuring the entropy of a real synthesis flow is an empirical task that depends on the tool and the design. The rare-event model of trojans similarly treats tests as independent samples from a fixed distribution, which is a simplification of directed test generation. Whether a theory of white-box concealment can be built that reduces to the present black-box results in the appropriate limit is an open question. The essay states it as one and does not supply an answer.

What the analysis does establish is narrower and, it is hoped, firm. Hiding content and hiding existence are not the same security property. They differ in the distribution against which indistinguishability is measured, in the resources their circuits consume, and in the way they fail. Encryption fails when a key is lost. Concealment fails when the adversary's model of the channel is better than the embedder's. Treating the two as stages of one scale of secrecy obscures that difference, whereas treating them as separate properties that compose in a specified order makes both the strengths and the costs of each explicit.

References

[1] B. Barak, O. Goldreich, R. Impagliazzo, S. Rudich, A. Sahai, S. Vadhan, K. Yang. On the (im)possibility of obfuscating programs. *CRYPTO*, 2001.

[2] C. Cachin. An information-theoretic model for steganography. *Information Hiding*, 1998.

[3] N. Dedić, G. Itkis, L. Reyzin, S. Russell. Upper and lower bounds on black-box steganography. *Journal of Cryptology*, 2009.

[4] S. Garg, C. Gentry, S. Halevi, M. Raykova, A. Sahai, B. Waters. Candidate indistinguishability obfuscation and functional encryption for all circuits. *FOCS*, 2013.

[5] S. Goldwasser, S. Micali. Probabilistic encryption. *Journal of Computer and System Sciences*, 1984.

[6] N. Hopper, J. Langford, L. von Ahn. Provably secure steganography. *CRYPTO*, 2002.

[7] A. Kahng, J. Lach, W. Mangione-Smith, S. Mantik, I. Markov, M. Potkonjak, P. Tucker, H. Wang, G. Wolfe. Watermarking techniques for intellectual property protection. *DAC*, 1998.

[8] A. Kerckhoffs. La cryptographie militaire. *Journal des sciences militaires*, 1883.

[9] J. Roy, F. Koushanfar, I. Markov. EPIC: Ending piracy of integrated circuits. *DATE*, 2008.

[10] C. Shannon. Communication theory of secrecy systems. *Bell System Technical Journal*, 1949.

[11] G. Simmons. The prisoners' problem and the subliminal channel. *CRYPTO*, 1983 (published 1984).

[12] P. Subramanyan, S. Ray, S. Malik. Evaluating the security of logic encryption algorithms. *HOST*, 2015.

[13] M. Tehranipoor, F. Koushanfar. A survey of hardware trojan taxonomy and detection. *IEEE Design & Test of Computers*, 2010.