2019 IEEE European Symposium on Security and Privacy (EuroS&P)

A Mechanised Cryptographic Proof of the WireGuard Virtual Private Network Protocol

Benjamin Lipp
INRIA Paris
benjamin.lipp@inria.fr

Bruno Blanchet
INRIA Paris
bruno.blanchet@inria.fr

Karthikeyan Bhargavan
INRIA Paris
karthikeyan.bhargavan@inria.fr

***Abstract*—WireGuard is a free and open source Virtual Private Network (VPN) that aims to replace IPsec and OpenVPN. It is based on a new cryptographic protocol derived from the Noise Protocol Framework. This paper presents the first mechanised cryptographic proof of the protocol underlying WireGuard, using the CryptoVerif proof assistant.**

We analyse the entire WireGuard protocol as it is, including transport data messages, in an ACCE-style model. We contribute proofs for correctness, message secrecy, forward secrecy, mutual authentication, session uniqueness, and resistance against key compromise impersonation, identity mis-binding, and replay attacks. We also discuss the strength of the identity hiding provided by WireGuard.

Our work also provides novel theoretical contributions that are reusable beyond WireGuard. First, we extend CryptoVerif to account for the absence of public key validation in popular Diffie-Hellman groups like Curve25519, which is used in many modern protocols including WireGuard. To our knowledge, this is the first mechanised cryptographic proof for any protocol employing such a precise model. Second, we prove several indifferentiability lemmas that are useful to simplify the proofs for sequences of key derivations.

***Index Terms*—security protocols, verification, computational model, VPN**

I. Introduction

The traditional distinction between a secure intranet and the untrusted Internet is becoming less relevant as more and more enterprises host internal services on cloud-based servers distributed across multiple data centres. Sensitive data that used to travel only between physically proximate machines within secure buildings is now sent across an unknown number of network links that may be controlled by malicious entities.

To maintain the security of such *distributed intranets*, the most powerful tools at the disposal of system administrators are Virtual Private Network (VPN) protocols that set up low-level secure channels between machines, and hence can be used to transparently protect all the data exchanged between them. Indeed, all leading cloud providers now offer VPN gateways, so that enterprises can treat cloud-based servers as if they were located within their intranet.[1]

Standards vs. Custom Protocols. Most popular VPN solutions are based on Internet standards like IPsec [1] and TLS [2], for several reasons. First, these protocols typically have multiple interoperable implementations that are available on all mainstream operating systems, so the VPN software can be easily built as a layer on top. Second, standards are designed to be future-proof by relying on versioning and *cryptographic agility*, so that a VPN protocol can easily move from one protocol version or cryptographic algorithm to another if (say) a weakness were found on some configuration. Third, published standards typically have been closely scrutinised by numerous interested parties, and hence are believed to be less likely to contain obvious security flaws.

Conversely, using a standard protocol also has its disadvantages. Standardisation takes time, and so a standard protocol may not use the most modern cryptographic algorithms. On the contrary, the need for interoperability and backwards compatibility often force implementations to continue support for obsolete cryptographic algorithms, leading to cryptanalytic attacks [3] and software flaws [4]. Over time, standards and their implementations can grow to an unmanageable size that can no longer be studied as a whole, allowing logical flaws to hide in unused corners of the protocol [5].

Consequently, many new secure channel protocols eschew standardisation in favour of a lean design that uses only modern cryptography and supports minimal cryptographic agility. The succinctness of the protocol description aids auditability, and the lack of optional features reduces complexity. Examples of this approach are the Signal protocol [6] used in many secure messaging systems and the Noise protocol framework [7].

WireGuard is a VPN protocol that adopts this design philosophy [8]. It implements and extends a secure channel protocol derived from the Noise framework, and it chooses a small set of modern cryptographic primitives. By making these choices, WireGuard is able to provide a high-quality VPN in a few thousand lines of code, and is currently being considered for adoption within the Linux kernel. The design of WireGuard is detailed and informally analysed in [8], but a protocol of such importance deserves a thorough security analysis.

A Need for Mechanised Proofs. Having a succinct, well-documented description is a good basis for understanding, auditing, and implementing a custom cryptographic protocol, but in itself is no guarantee that the protocol is secure. Symbolic analysis with tools like ProVerif [9] and Tamarin [10] can help find logical flaws, and WireGuard already has been analysed using Tamarin [11]. However, symbolic analyses do not constitute a full cryptographic proof. For example, they

[1] https://cloud.google.com/vpn/docs/concepts/overview, https://docs.aws.amazon.com/vpc/latest/userguide/vpn-connections.html, https://azure.microsoft.com/en-us/services/vpn-gateway/

DOI 10.1109/EuroSP.2019.00026

cannot demonstrate the absence of cryptanalytic attacks on secure channels and VPNs (e.g. [3].)

Cryptographic proofs provide the highest form of formal assurance, but writing proofs by hand requires significant expertise and effort, especially if the proof is to account for the precise low-level details of a real-world protocol. And as proofs get larger, the risk of introducing proof errors becomes non-negligible. All this effort is hard to justify for a custom protocol which may change as the software evolves. For example, a manual cryptographic proof for the WireGuard protocol appears in [12], but this proof would need to be carefully reviewed and adapted if the WireGuard protocol were to change in any way or if a variant of WireGuard were to be proposed.

We advocate the use of mechanised provers to build cryptographic proofs, so that they can be checked for errors, and can be easily modified to accommodate different variants of the protocol. In this paper, we rely on the CryptoVerif protocol verifier [13], [14] to build a proof of WireGuard. CryptoVerif relies on a computational model of cryptography, and generates machine-checkable proofs by sequences of games, like those manually written by cryptographers.

Uncovering Real-World Cryptographic Assumptions. A mechanised proof also allows the analyst to experiment with a variety of cryptographic assumptions and discover the precise set of assumptions that a protocol's security depends on.

In some cases, a protocol may require an unusual assumption about a hash function, or a stronger assumption about encryption than one may have expected, and these cases can provide a guide to implementers on what concrete cryptographic algorithms should or should not be used to instantiate the protocol. For example, in our analysis of WireGuard, we find that most of the standard properties require only standard assumptions about the underlying authenticated encryption scheme (AEAD) but identity hiding requires a stronger assumption, which is satisfied by the specific algorithms used by WireGuard, but may not be provided by other AEAD constructions.

In other cases, a protocol's use of a cryptographic primitive may motivate a new, more precise model of the primitive. Protocols like WireGuard seek to depend on a small set of primitives and reuse them in different ways. For example, WireGuard relies on the Curve25519 elliptic curve Diffie-Hellman operation for an ephemeral key exchange as well as for entity authentication. It uses Curve25519 public keys both as identities and as unique nonces to identify sessions. To verify that Curve25519 is appropriate for all these usages, and to prove the absence of attacks such as replays, identity mis-binding, and key compromise impersonation, we need to account for the details of the Curve25519 group, rather than rely on a generic Diffie-Hellman assumption. Hence, we propose a new model for Curve25519 in CryptoVerif and prove WireGuard secure against this model.

Contributions. We present the first mechanised proof for the cryptographic design of the WireGuard VPN, including the Noise IKpsk2 secure channel protocol it uses. Our analysis is done on WireGuard v1 as specified in [8]. In addition to classic key exchange security for IKpsk2, we examine the identity hiding and denial-of-service protections provided by WireGuard. We conclude with a discussion of the strengths and weaknesses of WireGuard, and propose improvements that would allow for stronger security theorems.

Our work also provides contributions reusable beyond the proof of WireGuard. To the best of our knowledge, this is the first mechanised proof for any cryptographic protocol that takes into account the precise structure of the Curve25519 group. We also prove a series of indifferentiability results that allow us to simplify sequences of random oracle calls, and we made several extensions to CryptoVerif that we mention in the rest of the paper when we use them. These extensions are included in CryptoVerif 2.01 available at https://cryptoverif.inria.fr/.

Our models of WireGuard are available at https://cryptoverif.inria.fr/WireGuard and more details are available in the long version of the paper [15].

II. WireGuard

WireGuard [8] establishes a VPN tunnel between two remote hosts in order to securely encapsulate all Internet Protocol (IP) traffic between them. The main design goals of WireGuard are to be simple, fast, modern, and secure. In order to establish a tunnel, a system administrator only needs to configure the IP address and long-term public key for the remote host. With this information, WireGuard can establish a secure channel, using a protocol derived from the Noise framework, instantiated with fast, modern cryptographic primitives like Curve25519 and BLAKE2. The full WireGuard VPN is implemented in a few thousand lines of code that can run on multiple platforms, but for performance, is usually run within the operating system kernel. In particular, WireGuard is in the process of being incorporated into the Linux kernel (most likely Linux 4.2/5.0), as an alternative to IPsec.

In this section, we focus on the cryptographic design of WireGuard. We begin by describing the secure channel component, then the extensions WireGuard makes for denial-of-service and stealthy operation. We end the section by detailing the concrete cryptographic algorithms used by WireGuard and the list of informal security goals it seeks to achieve.

A. Secure Channel Protocol: Noise IKpsk2

Noise [7] is a framework for building two-party cryptographic protocols that are secure by construction. Using the building blocks in this framework, a designer can create a new protocol that matches a desired subset of security guarantees: mutual or optional authentication, identity hiding, forward secrecy, etc. The Noise specification also includes a list of curated pre-defined protocols, with an informal analysis of their message-by-message security claims. WireGuard instantiates one of these protocols, which is called IKpsk2, and extends it to provide further guarantees needed by VPNs.

The secure channel protocol is depicted in Figure 1a, and the cryptographic computations are detailed in Figure 1b, using notations similar to [8]. Before the protocol begins, the initiator i and the responder r are assumed to have exchanged their long-term *static public keys* (S_i^{pub}, S_r^{pub}). Optionally, they may

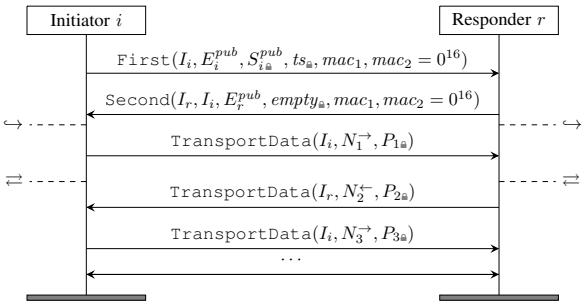
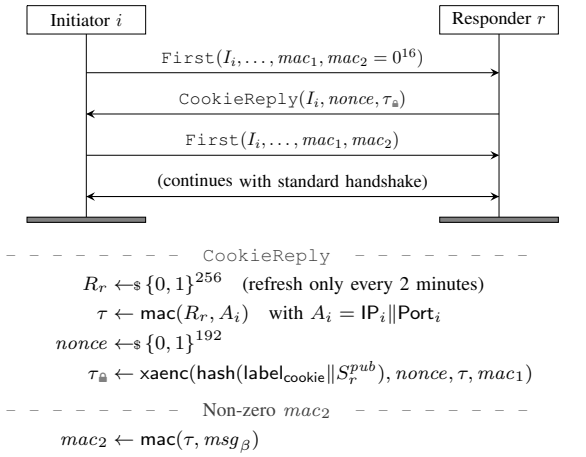


Figure 1a: WireGuard's protocol messages.



CookieReply

$R_r \leftarrow\!\$\ \{0,1\}^{256}$ (refresh only every 2 minutes)

$\tau \leftarrow \mathsf{mac}(R_r, A_i)$ with $A_i = \mathsf{IP}_i \| \mathsf{Port}_i$

$nonce \leftarrow\!\$\ \{0,1\}^{192}$

$\tau_{\text{🔒}} \leftarrow \mathsf{xaenc}(\mathsf{hash}(\mathsf{label}_{\mathsf{cookie}} \| S_r^{pub}), nonce, \tau, mac_1)$

Non-zero mac_2

$mac_2 \leftarrow \mathsf{mac}(\tau, msg_\beta)$

Figure 1c: Cookie mechanism under load.

First

E_i^{pub} : $(E_i^{priv}, E_i^{pub}) \leftarrow\!\$\ \mathsf{keygen}()$

$C_0 \leftarrow \mathsf{hash}(\mathsf{protocol\_name})$

$H_0 \leftarrow \mathsf{hash}(C_0 \| \mathsf{prologue})$

$H_1 \leftarrow \mathsf{hash}(H_0 \| S_r^{pub})$

$C_1 \leftarrow \mathsf{hkdf}_1(C_0, E_i^{pub})$

$H_2 \leftarrow \mathsf{hash}(H_1 \| E_i^{pub})$

$C_2 \| k_1 \leftarrow \mathsf{hkdf}_2(C_1, \mathsf{dh}(E_i^{priv}, S_r^{pub}))$

$S_{i\text{🔒}}^{pub}$: $S_{i\text{🔒}}^{pub} \leftarrow \mathsf{aenc}(k_1, 0, S_i^{pub}, H_2)$

$H_3 \leftarrow \mathsf{hash}(H_2 \| S_{i\text{🔒}}^{pub})$

$C_3 \| k_2 \leftarrow \mathsf{hkdf}_2(C_2, \mathsf{dh}(S_i^{priv}, S_r^{pub}))$

$ts_{\text{🔒}}$: $ts_{\text{🔒}} \leftarrow \mathsf{aenc}(k_2, 0, \mathsf{timestamp}(), H_3)$

$H_4 \leftarrow \mathsf{hash}(H_3 \| ts_{\text{🔒}})$

mac_1 : $mac_1 \leftarrow \mathsf{mac}(\mathsf{hash}(\mathsf{label}_{\mathsf{mac1}} \| S_r^{pub}), msg_\alpha)$

Second

E_r^{pub} : $(E_r^{priv}, E_r^{pub}) \leftarrow\!\$\ \mathsf{keygen}()$

$C_4 \leftarrow \mathsf{hkdf}_1(C_3, E_r^{pub})$

$H_5 \leftarrow \mathsf{hash}(H_4 \| E_r^{pub})$

$C_5 \leftarrow \mathsf{hkdf}_1(C_4, \mathsf{dh}(E_r^{priv}, E_i^{pub}))$

$C_6 \leftarrow \mathsf{hkdf}_1(C_5, \mathsf{dh}(E_r^{priv}, S_i^{pub}))$

$C_7 \| \pi \| k_3 \leftarrow \mathsf{hkdf}_3(C_6, psk)$

$H_6 \leftarrow \mathsf{hash}(H_5 \| \pi)$

$empty_{\text{🔒}}$: $empty_{\text{🔒}} \leftarrow \mathsf{aenc}(k_3, 0, empty, H_6)$

$H_7 \leftarrow \mathsf{hash}(H_6 \| empty_{\text{🔒}})$

mac_1 : $mac_1 \leftarrow \mathsf{mac}(\mathsf{hash}(\mathsf{label}_{\mathsf{mac1}} \| S_i^{pub}), msg_\alpha)$

Key Derivation

$T^{\rightarrow} \| T^{\leftarrow} \leftarrow \mathsf{hkdf}_2(C_7, empty)$

TransportData

$P_{1\text{🔒}} \leftarrow \mathsf{aenc}(T^{\rightarrow}, N_1^{\rightarrow} = 0, P_1, empty)$

$P_{2\text{🔒}} \leftarrow \mathsf{aenc}(T^{\leftarrow}, N_2^{\leftarrow} = 0, P_2, empty)$

$P_{3\text{🔒}} \leftarrow \mathsf{aenc}(T^{\rightarrow}, N_3^{\rightarrow} = 1, P_3, empty)$

Figure 1b: Cryptographic Computations for Protocol Messages.

Figure 1: (a) An overview of WireGuard's main protocol messages; (b) the cryptographic computations used to create these messages; they need to be adapted accordingly for the receiving side; and (c) the cookie mechanism used by WireGuard to protect hosts against Denial-of-Service attacks. We write $x_{\text{🔒}}$ for a variable containing an encryption of x; $x_{\text{🔒}}$ is just a variable identifier. msg_α refers to all the bytes of a message up to but not including mac_1, msg_β is the same but including mac_1. Session key derivation takes places after the second protocol message, symbolised by $\hookrightarrow$, at which point the initiator can send messages. The end of the handshake is symbolised by $\rightleftarrows$, after which transport data messages can be sent in both directions. The cookie mechanism is depicted in one direction, initiator to responder, but can actually be used by either initiator or responder, whichever is under load.

have also established a *pre-shared symmetric key* (psk); if this key is absent it is set to a key-sized bitstring of zeros.

Message Exchange. The protocol begins when i sends the first handshake message to r, which includes the following components:

- I_i: a fresh session identifier, generated by i,
- E_i^{pub}: a fresh ephemeral public key, generated by i,
- $S_{i\text{🔒}}^{pub}$: i's static public key, encrypted for r,
- $ts_{\text{🔒}}$: a timestamp, encrypted with a key that can be computed only by i and r, and
- mac_1, mac_2: message authentication codes (see §II-B).

In response, r sends the second handshake message containing:

- I_i: i's session identifier,
- I_r: a fresh session identifier, generated by r,
- E_r^{pub}: a fresh ephemeral public key, generated by r,
- $empty_{\text{🔒}}$: an empty bytestring encrypted with a key that can be computed only by i and r, and
- mac_1, mac_2: message authentication codes (see §II-B).

The encrypted payloads in the two messages serve as authenticators: by computing the corresponding encryption key, each party proves that it knows the private key for its static public key. The encryption key for the second message

also requires knowledge of the optional psk providing an additional authentication guarantee. The two ephemeral keys add fresh session-specific key material that can be used to compute (forward) secret session keys known only to i and r.

At the end of these two messages, i and r derive authenticated encryption keys $(T^{\rightarrow}, T^{\leftarrow})$ that can be used to transport IP traffic in the two directions. Importantly, i sends the first transport message, hence confirming the successful completion of the handshake to r, before r sends it any encrypted traffic. Each of these transport messages includes:

- I_i or I_r: the recipient's session identifier,
- $N_j^{\leftarrow}$ or $N_j^{\rightarrow}$: the current message counter,
- P_j: an IP datagram, encrypted under the traffic key.

Cryptographic Computations. Figure 1b describes how each of these message components and traffic keys are computed. As the handshake proceeds, i and r compute a sequence of *transcript hashes* $(H_0, H_1, \ldots, H_7)$ that hashes in all the public data used in the two handshake messages, including:

- protocol_name, prologue: strings identifying the protocol,
- E_i^{pub}, E_r^{pub}: both ephemeral public keys,
- $S_r^{pub}, S_{i\,🔒}^{pub}$: both static public keys, but with the initiator's key in encrypted form,
- $ts_{🔒}, empty_{🔒}$: both encrypted handshake payloads, and
- π: an identifier derived from the pre-shared key.

These transcript hashes serve as unique identifiers for the current stage of the session. In particular, no two completed WireGuard sessions should have the same H_7.

Both parties also derive a sequence of *chaining keys* $(C_0, C_1, \ldots, C_7)$ by mixing in all the key material, including:

- protocol_name, E_i^{pub}, E_r^{pub},
- $\mathsf{dh}(E_i^{priv}, S_r^{pub}) = \mathsf{dh}(S_r^{priv}, E_i^{pub})$: the *ephemeral-static* Diffie-Hellman shared secret computed using the initiator's ephemeral key (named first in *ephemeral-static*) and the responder's static key (named second in *ephemeral-static*),
- $\mathsf{dh}(S_i^{priv}, S_r^{pub}) = \mathsf{dh}(S_r^{priv}, S_i^{pub})$: the static-static shared secret,
- $\mathsf{dh}(E_i^{priv}, E_r^{pub}) = \mathsf{dh}(E_r^{priv}, E_i^{pub})$: the ephemeral-ephemeral shared secret,
- $\mathsf{dh}(S_i^{priv}, E_r^{pub}) = \mathsf{dh}(E_r^{priv}, S_i^{pub})$: the static-ephemeral shared secret, and
- psk: the (optional) pre-shared key.

The function dh is the elliptic curve scalar multiplication, taking a private key and a public key as argument, permitting the computation of a *shared secret* [16]. In the preceding list, the initiator uses the first function call, and the responder the second one, respectively.

The protocol uses all four combinations of static and ephemeral Diffie-Hellman shared-secret computations to maximally protect against the compromise of some of these keys. The psk also serves as a defensive countermeasure against quantum adversaries who may be able to break the Diffie-Hellman construction, but not hkdf. Hence, by using a frequently updated psk, WireGuard users can protect current sessions against future quantum adversaries.

Each chaining key is mixed into the next chaining key via an hkdf key derivation that also outputs encryption keys as needed. This chain of key derivations outputs two encryption keys (k_1, k_2) for the first handshake message, an encryption key (k_3) and a PSK identifier (π) for the second message, and traffic keys $(T^{\leftarrow}, T^{\rightarrow})$ for all subsequent transport messages.

To encrypt each message, WireGuard uses an authenticated encryption scheme with associated data (AEAD) that takes a key, a counter, a plaintext (padded up to the nearest blocksize) and an optional hash value as associated data. The encryptions in the handshake messages use the current transcript hash (H_2, H_3, H_6) as associated data, which guarantees that the two participants have a consistent session transcript. Transport messages use an empty string as associated data. The message counter is initially set to 0 for each AEAD key and incremented by 1 every time the key is reused.

Relationship with IKpsk2. The secure channel protocol described above is a direct instantiation of Noise IKpsk2, with five notable differences. First, WireGuard adds local session identifiers (I_i, I_r) for the initiator and responder. Second, WireGuard fixes the payload of the first message to a timestamp, and the one of the second message to the empty string. Third, WireGuard stipulates that the first traffic message is sent from the initiator to the responder. Fourth, WireGuard excludes zero Diffie-Hellman shared secrets to avoid points of small order, while Noise recommends not to perform this check. Fifth, WireGuard adds two message authentication codes to the handshake messages, to provide stealth and to protect against DoS, as described in the next section. We also observe that although this protocol is superficially similar to other popular Noise protocols like IK (which is used in WhatsApp), there are important differences between these variants and a proof for one does not translate to the other.

B. Extensions for Stealth and Denial-of-Service

A VPN protocol operates at a low-level in the networking stack and hence needs to not only protect against cryptographic attacks, but also real-world network-level attacks such as *denial of service* (DoS). Indeed, a cryptographic protocol like IKpsk2 that needs to perform two expensive Diffie-Hellman operations before it can authenticate a handshake message is even more vulnerable to DoS: an adversary can send bogus messages that tie up computing resources on the recipient. A further security goal for WireGuard is that its VPN endpoints should be *stealthy*, in the sense that it should not be possible for a network adversary to blindly scan for WireGuard services.

To support stealthy operation, WireGuard endpoints do not respond to any handshake message unless the sender can prove that it knows the static public key of the recipient. This proof is incorporated in the mac_1 field included in each handshake message, which contains a message authentication code (MAC) computed over the prefix of the current handshake message up to but not including mac_1, using a MAC key derived from the recipient's static public key. The recipient verifies this MAC before processing the message, and stays silent if the MAC fails. Hence, a network adversary who does not know the

public key cannot detect whether WireGuard is running on a machine, and at the same time cannot force the recipient to perform two finally useless Diffie-Hellman operations.

To protect more actively against DoS, WireGuard incorporates a cookie-based protocol (depicted in Figure 1c) that a host can use when it is under load. For example, if the responder suspects it is under a DoS attack, it can refuse to process the first handshake message and instead send back an initiator-specific fresh *cookie* (τ) that is computed from a frequently rotated secret key (R_r) (known only to the responder) and the initiator's IP address (IP_i) and source port (Port_i). The responder encrypts this cookie for the initiator, using a key derived from the initiator's static public key, a fresh nonce, and the mac_1 field of the first message as associated data.

The initiator decrypts τ and then retries the handshake by sending the first message again, but this time with a second field mac_2 that contains a MAC over the full message up to and including mac_1, using τ as the MAC key. After verifying this MAC, the responder continues with the standard handshake.

However, to obtain τ, an adversary must be able to read messages on the network path between the initiator and responder and must also know the initiator's static key (which is never sent in the clear by the protocol). And even if the adversary has both these capabilities, it is required to perform session specific cryptographic computations for every handshake message it sends to the responder, significantly limiting its ability to mount a DoS attack. Hence, this cookie protocol protects the recipient from brute-force network attacks.

Note that the mac_2 field is included in both handshake messages, and hence can be used in both directions, to protect both the initiator and responder from DoS attacks.

The two MACs are WireGuard-specific mechanisms which are not present in IKpsk2. Since they do not use any of the session keys (or hashes or chaining keys) that are used in IKpsk2, adding these mechanisms should, in principle, not affect the security of the secure channel protocol. However, since the static public keys of the two hosts are used in the two MACs, we need to carefully study their impact on the identity-hiding guarantees of IKpsk2.

C. Instantiating the Cryptographic Algorithms

WireGuard uses a small set of cryptographic constructions and instantiates them with modern algorithms, carefully chosen to provide strong security as well as high performance:

- dh: all Diffie-Hellman operations use the Curve25519 elliptic curve [16];
- hash: the BLAKE2s hash function [17];
- aenc: authenticated encryption for handshake and traffic message uses the AEAD scheme ChaCha20Poly1305 [18], using the message counter as nonce;
- xaenc: cookie encryption uses an *extended* AEAD construction using XChaCha20Poly1305, which incorporates a 192-bit random nonce [19] into the standard ChaCha20Poly1305 construction;
- mac: all MAC operations use the keyed MAC variant of the BLAKE2s hash function;
- hkdf_n: all key derivations use the HKDF construction [20], using BLAKE2s as the underlying hash function.

The values $\mathsf{label}_{\mathsf{mac1}}$ and $\mathsf{label}_{\mathsf{cookie}}$ are distinct constants.

D. Security Goals, Informally

Using the mechanisms described in this section, WireGuard seeks to provide the following set of strong security guarantees, inheriting the security claims of Noise IKpsk2 [7] and extending them with the additional DoS and stealth goals of WireGuard [8]. In the following, we use *honest* to refer to a party that follows the protocol specification, and *dishonest* to a party that doesn't, i.e. that is controlled by the adversary. Most properties are defined to hold within a *clean* session; we define this notion formally in Section §V-A.

- **Correctness:** If an honest initiator and an honest responder complete a WireGuard handshake and the messages are not altered by an adversary, then the transport data keys ($T^{\rightarrow}, T^{\leftarrow}$) and the transcript hash H_7 are the same on both hosts.
- **Secrecy:** If a transport data message P is sent over a tunnel between two honest hosts, then this message is kept confidential from the adversary. Furthermore, the traffic keys for this tunnel are also confidential.
- **Forward Secrecy:** Secrecy for a session holds even if both the static private keys (S_i^{priv}, S_r^{priv}) and the pre-shared key (psk) become known to the adversary, but only after the session has been completed and all its traffic keys and chaining keys are deleted by both parties.
  Secrecy also holds even if the static and ephemeral keys are compromised (e.g. by a quantum adversary), as long as the pre-shared key is not compromised.
- **Mutual Authentication:** If an honest initiator (resp. responder) completed a handshake (ostensibly) with an honest peer, then that peer must have participated in this handshake. Moreover, if a host A receives a plaintext message over a WireGuard tunnel that claims to be from host B, then B must have (intentionally) sent this message to A.
- **Resistance against Key Compromise Impersonation (KCI):** The recipient of a message can authenticate the message's sender even if the recipient's static key is compromised.
- **Resistance against Identity Mis-Binding:** If two honest parties derive the same traffic keys in some WireGuard session, then they agree on each other's identities, even if one or both of them have been interacting with a dishonest party or an honest party with compromised keys. This property is also called resistance against unknown key-share attacks.
- **Resistance against Replay:** Any protocol message sent may be accepted at most once by the recipient.
- **Session Uniqueness:** There is at most one honest initiator session and at most one honest responder session for a given traffic key. Similarly, there is at most one honest initiator session and at most one honest responder session for given handshake messages.

- **Channel Binding:** Two sessions that have the same final session transcript hash H_7 share the same view and the same session keys.
- **Identity Hiding:** Just by looking at the messages transmitted over the network, a passive adversary cannot infer the static keys involved in a session. (However, these identities are not forward secret: If the responder's static key gets compromised, the adversary can later decrypt the initiator's static public key that was transmitted in the first message.)
- **DoS Resistance:** The adversary cannot have a message accepted by a recipient under load without having first made a round trip with that recipient. In practice, this means that the adversary has to be at the claimed address. Because we assume that the adversary controls the network, we cannot prove more than enforcing a round trip.

The security goals above are stated in terms of completed WireGuard sessions, with most security guarantees only applying after the third message, when both initiator and responder start freely sending and receiving data. In particular, the first transport data message (i.e. the third message) serves as key confirmation to the responder, and is needed to prove that the initiator has control over its ephemeral key. This is why, in WireGuard, the responder does not send any data until it sees this third message. In the rest of this paper, we investigate whether WireGuard achieves the goals set out above.

III. Cryptographic Assumptions

This section presents the assumptions that we make on the cryptographic primitives used by WireGuard. For most primitives, the desired assumption is already present in the library of primitives of CryptoVerif, so we just need to call a macro to use that assumption. Still, we had to design a new model for Curve25519, detailed below.

A. Random Oracle Model

We assume that BLAKE2s is a random oracle [21]. This assumption is justified in [22] using a weak ideal block cipher.

B. IND-CPA and INT-CTXT for AEAD

We assume that the ChaCha20Poly1305 AEAD scheme [18] is IND-CPA (indistinguishable under chosen plaintext attacks) and INT-CTXT (ciphertext integrity) [23], provided the same nonce is never used twice with the same key. IND-CPA means that the adversary has a negligible probability of distinguishing encryptions of two distinct messages of the same length that it has chosen. INT-CTXT means that an adversary with access to encryption and decryption oracles has a negligible probability of forging a ciphertext that decrypts successfully and has not been returned by the encryption oracle. These properties are justified in [24], assuming ChaCha20 is a PRF (pseudo-random function) and Poly1305 is an ϵ-almost-Δ-universal hash function. The latter property is shown to hold in [25].

C. Curve25519 and Gap Diffie-Hellman

WireGuard uses the elliptic curve Curve25519 [16] for Diffie-Hellman key exchanges. This curve is a group G of order kq where $k = 8$ (cofactor) and q is a large prime. The base point g has prime order q; we denote by G_{sub} the prime order subgroup generated by g. In WireGuard and typical implementations of Curve25519 as specified by RFC 7748 [16], the incoming public keys are not verified, so they may be any element of G and may not belong to G_{sub}, and all exponents are non-zero multiples of k modulo kq. For each public key X in G, there are k public keys Y in G such that $X^k = Y^k$ and only one of these public keys is in G_{sub}. (We write point multiplication exponentially.) We say that public keys X and Y such that $X^k = Y^k$ are *equivalent*, because they yield the same Diffie-Hellman shared secrets: for any exponent $z = kz'$, $X^z = X^{kz'} = Y^{kz'} = Y^z$. Moreover, the public keys may be 0, the neutral element of G and G_{sub}, and $0^x = 0$ for all x.

While most proofs of Diffie-Hellman key agreements assume a prime order group, that assumption is not correct for most implementations of Curve25519. For instance, the identity misbinding issue that we discuss in Section VI would not appear in a prime order group. Therefore, we need to provide a new model that takes into account the properties mentioned above.

The main idea of our model is to rely on a Diffie-Hellman assumption in the prime order subgroup G_{sub}, and so to work as much as possible with elements in G_{sub}. We rewrite the computations in G into computations in G_{sub} by first raising the public keys to the power k, and we rely on standard properties of prime order groups for G_{sub}.

In CryptoVerif, we first define the following types:

type G [bounded, large].
type G_{sub} [bounded, large].
type Z [bounded, large, nonuniform].

The type G represents the group G; it is bounded because it is represented by bitstrings of bounded length, and large because collisions between randomly chosen elements in G have a negligible probability. Similarly, the type G_{sub} represents the group G_{sub}, and the type Z corresponds to non-zero integers multiple of k modulo kq. When honest participants choose exponents, they are chosen uniformly in a *subset* of Z: they are of the form $2^{254} + 8n$ for $n \in \{0, \ldots, 2^{251} - 1\}$ and $kq > 2^{255}$. Therefore, the distribution for choosing random exponents inside the whole Z is non-uniform, which is indicated by the annotation nonuniform.

We define functions:

fun $exp(G, Z) : G$.
fun $mult(Z, Z) : Z$.
equation builtin commut($mult$).

We have $exp(X, y) = X^y$, and $mult$ is the product modulo kq, in Z. Since its two arguments are non-zero multiples of k, so is its result, and it is in Z. The last line states that the function $mult$ is commutative. (We could add associativity and other properties, like existence of inverses, but commutativity is typically sufficient to prove security of basic Diffie-Hellman

key exchanges. More algebraic properties may be needed to prove group Diffie-Hellman protocols, for instance. Note that not modelling these does not restrict the adversary in the computational model.)

fun $pow\_k(G) : G_{sub}$.
fun $exp\_div\_k(G_{sub}, Z) : G_{sub}$.
fun $G_{sub}2G(G_{sub}) : G$ [data].
equation forall $x : G_{sub}, x' : G_{sub}$;
$(pow\_k(G_{sub}2G(x)) = pow\_k(G_{sub}2G(x'))) = (x = x')$.

We have $pow\_k(X) = X^k$, and it is in G_{sub} for all X in G. We have $exp\_div\_k(X, y) = X^{y/k}$. This function operates on G_{sub} and is convenient since the exponents in Z are always multiples of k. The function $G_{sub}2G$ is the identity from G_{sub} to G; it is necessary to convert elements of type G_{sub} to type G. The annotation data tells CryptoVerif that it is injective. The last equation says that pow_k is injective when restricted to the subgroup G_{sub}, of order q. Indeed, k is prime to q, so it can be inverted modulo q.

We also define constants:

const $zero : G$. **const** $zero_{sub} : G_{sub}$.
equation $zero = G_{sub}2G(zero_{sub})$.
const $g : G$. **const** $g\_k : G_{sub}$.
equation $pow\_k(g) = g\_k$. **equation** $g\_k \neq zero_{sub}$.

The neutral element is $zero$ as an element of G and $zero_{sub}$ as an element of G_{sub}. The base point is g, and $g\_k = g^k$.

We also state equations that hold on these functions:

equation forall $X : G, y : Z$;
$exp(X, y) = G_{sub}2G(exp\_div\_k(pow\_k(X), y))$. (1)

equation forall $X : G_{sub}, y : Z, z : Z$;
$exp\_div\_k(pow\_k(G_{sub}2G(exp\_div\_k(X, y))), z) = exp\_div\_k(X, mult(y, z))$. (2)

Equation (1) says that $X^y = (X^k)^{y/k}$ and Equation (2) that $((X^{y/k})^k)^{z/k} = X^{y.z/k}$. Equation (2) applies in particular to simplify $exp(exp(X, y), z)$ after applying (1): $exp(exp(X, y), z) = G_{sub}2G(exp\_div\_k(pow\_k(G_{sub}2G(exp\_div\_k(pow\_k(X), y))), z)) = G_{sub}2G(exp\_div\_k(pow\_k(X), mult(y, z)))$. These equations are used by CryptoVerif as rewrite rules, to rewrite the left-hand side into the right-hand side. They allow rewriting computations in the group G into computations that happen in the subgroup G_{sub}, after raising the public key to the power k. In particular, $exp(g, y) = G_{sub}2G(exp\_div\_k(g\_k, y))$ and $exp(exp(g, y), z) = G_{sub}2G(exp\_div\_k(g\_k, mult(y, z)))$.

The next equation allows CryptoVerif to simplify equality tests with the neutral element, which are used by some protocols, including WireGuard, to exclude that element from the allowed public keys.

equation forall $X : G_{sub}, y : Z$;
$(exp\_div\_k(X, y) = zero_{sub}) = (X = zero_{sub})$.

When $y \in Z$, $y = ky'$ for some y' not multiple of q, so y' is invertible modulo q. Therefore, $X^{y/k} = 0$ if and only if $X^{y'} = 0$ if and only if $(X^{y'})^{1/y'} = 0^{1/y'}$, that is, $X = 0$.

Other properties serve to simplify equalities between Diffie-Hellman values in G_{sub}, with the goal of showing that these equalities are false. When the Diffie-Hellman shared secrets are passed to a random oracle, these equality tests appear after using the random oracle assumption: we compare the arguments of each call to the random oracle with arguments of previous calls, to know whether the random oracle should return the result of a previous call.

equation forall $X : G_{sub}, X' : G_{sub}, y : Z$;
$(exp\_div\_k(X, y) = exp\_div\_k(X', y)) = (X = X')$. (3)

equation forall $X : G_{sub}, y : Z, z : Z$;
$(exp\_div\_k(X, y) = exp\_div\_k(X, z)) = ((y = z) \vee (X = zero_{sub}))$. (4)

collision $x \xleftarrow{R} Z$; **forall** $X : G_{sub}, Y : G_{sub}$;
return$(exp\_div\_k(X, x) = Y) \approx_{\mathrm{Pcoll1rand}(Z)}$ **return**$((X = zero_{sub}) \wedge (Y = zero_{sub}))$
if X independent-of $x \wedge Y$ independent-of x. (5)

Equation (3) holds because $y = ky'$ for some y' invertible modulo q as shown above. In particular, using (1), injectivity of $G_{sub}2G$, and (3), $exp(X, y) = exp(X', y)$ simplifies into $pow\_k(X) = pow\_k(X')$. In contrast, in a prime order group, $exp(X, y) = exp(X', y)$ implies $X = X'$. This is the reason why, in the identity mis-binding issue of Section VI, we fail to prove equality of the public keys $X = X'$ and can only prove $pow\_k(X) = pow\_k(X')$.

Equation (4) holds because, when $X \in G_{sub}$ is different from 0, X is a generator of G_{sub}, so all elements $X^{y'}$ for $y' \in [1, q-1]$ are distinct, hence all elements $X^{y/k}$ as well.

In the collision statement (5), $\mathrm{Pcoll1rand}(Z)$ is the probability that a randomly chosen element x in Z is equal to an element of Z independent of x. For Curve25519, since random exponents are chosen uniformly among a set of 2^{251} elements, $\mathrm{Pcoll1rand}(Z) = 2^{-251}$. Statement (5) means that the probability of distinguishing $exp\_div\_k(X, x) = Y$ from $(X = zero_{sub}) \wedge (Y = zero_{sub})$ is at most $\mathrm{Pcoll1rand}(Z)$ assuming X is chosen randomly in Z ($x \xleftarrow{R} Z$) and X and Y are independent of x. Indeed, suppose that $exp\_div\_k(X, x) = Y$ differs from $(X = zero_{sub}) \wedge (Y = zero_{sub})$. If $X = 0$, then $X^{x/k} = 0$, so both expressions reduce to $Y = 0$, so they cannot differ. Therefore, $X \neq 0$. The second expression is then false. Moreover, X is a generator of G_{sub}, so $Y = X^y$ for some y independent of x. The equality $X^{x/k} = Y = X^y$ holds if and only if $x/k = y \mod q$ so $x = ky \mod kq$ with ky independent of x, so this happens with probability $\mathrm{Pcoll1rand}(Z)$. So the first expression is true with probability $\mathrm{Pcoll1rand}(Z)$, and the two expressions differ with that probability. The support for side-conditions in collision statements is an extension of CryptoVerif that we implemented.

Our model includes a few other properties, detailed and justified in the long version of the paper. In particular, it includes properties for simplifying equalities between products in Z. Such equalities appear for instance after simplification of equalities $exp\_div\_k(g\_k, mult(x, y)) = exp\_div\_k(g\_k, mult(x', y'))$. For instance, we model that $mult(x, y) =$

$mult(x, y')$ if and only if $y = y'$ and that, when x is chosen randomly in Z and y and z are independent of x, we have $mult(x, y) = z$ with probability at most $\mathrm{Pcoll1rand}(Z)$.

This model is included as a macro in CryptoVerif's library of cryptographic primitives, so that it can easily be reused. It also applies to other curves that have a similar structure, for instance Curve448, which is also used by the Noise framework, and by other protocols like TLS 1.3.

We assume that the prime order subgroup G_{sub} satisfies the gap Diffie-Hellman (GDH) assumption [26]. This assumption means that given a generator g, g^a, and g^b for random a, b, the adversary has a negligible probability to compute g^{ab}, even when the adversary has access to a decisional Diffie-Hellman oracle, which tells him given G, X, Y, Z whether there exist x, y such that $X = G^x$, $Y = G^y$, and $Z = G^{xy}$. It was already modelled in CryptoVerif.

In contrast, in their cryptographic proof of WireGuard, Dowling and Paterson [12] use the PRF-ODH assumption. We use the GDH and random oracle assumptions instead because CryptoVerif cannot currently use the PRF-ODH assumption in scenarios with key compromise. While in principle the PRF-ODH assumption is weaker, Brendel et al. [27] show that it is implausible to instantiate the PRF-ODH assumption without a random oracle, so our assumptions and the one of [12] are in fact fairly similar.

IV. Indifferentiability of Hash Chains

Before modelling WireGuard, we first present a different, equally precise, formulation of hash chains that is more amenable to a mechanised proof in CryptoVerif. Indeed, WireGuard makes many hash oracle calls to BLAKE2s, and at each call to a random oracle, CryptoVerif tests whether the arguments are the same as in any other previous random oracle call (to return the previous result of the random oracle). Therefore, using directly BLAKE2s as a random oracle would introduce a very large number of cases and yield exaggeratedly large cryptographic games. In order to avoid that, we simplify the random oracle calls using indifferentiability lemmas. These lemmas are not specific to WireGuard and can be used to simplify sequences of random oracle calls in other protocols, including other Noise protocols and Signal [6]. In the future, these lemmas may serve as a basis for an indifferentiability prover inside CryptoVerif, which would simplify random oracle calls before proving the protocol.

Specifically, WireGuard uses HKDF in a chain of calls to derive symmetric keys at different stages of the protocol:

$$
\begin{array}{llll}
C_0 & \leftarrow \mathsf{const} & C_5 & \leftarrow \mathsf{hkdf}_1(C_4, v_4) \\
C_1 & \leftarrow \mathsf{hkdf}_1(C_0, v_0) & C_6 & \leftarrow \mathsf{hkdf}_1(C_5, v_5) \\
C_2 \| k_1 & \leftarrow \mathsf{hkdf}_2(C_1, v_1) & C_7 \| \pi \| k_3 & \leftarrow \mathsf{hkdf}_3(C_6, v_6) \\
C_3 \| k_2 & \leftarrow \mathsf{hkdf}_2(C_2, v_2) & T^{\rightarrow} \| T^{\leftarrow} & \leftarrow \mathsf{hkdf}_2(C_7, v_7) \\
C_4 & \leftarrow \mathsf{hkdf}_1(C_3, v_3) & &
\end{array}
$$

We show, using the indifferentiability lemmas of this section, that hkdf_n is indifferentiable from a random oracle, and that the chain above is indifferentiable from:

$$
\begin{array}{ll}
k_1 & \leftarrow \mathsf{chain}'_1(v_0, v_1) \\
k_2 & \leftarrow \mathsf{chain}'_2(v_0, v_1, v_2) \\
\pi \| k_3 \| T^{\rightarrow} \| T^{\leftarrow} & \leftarrow \mathsf{chain}'_6(v_0, v_1, v_2, v_3, v_4, v_5, v_6)
\end{array} \quad (6)
$$

Thus, we obtain a much simpler computation, which we use in our CryptoVerif model of WireGuard. Previous analyses of WireGuard did not use such a result because they do not rely on the random oracle model: [12] relies on the PRF-ODH assumption, [11] uses the symbolic model.

A. Definition of Indifferentiability

Indifferentiability can be defined as follows. This definition is an extension of [28] to several independent oracles. We give an asymptotic definition here. In the long version, we give explicit probabilities and proofs for all results.

Definition 1 (Indifferentiability)**.** *Functions $(F_i)_{1 \le i \le n}$ with oracle access to independent random oracles $(H_j)_{1 \le j \le m}$ are indifferentiable from independent random oracles $(H'_i)_{1 \le i \le n}$ if for each value of the security parameter η, there exists a simulator S that runs in time $P_1(\eta)$ such that for any distinguisher D that runs in time $P_2(\eta)$,*

$$|\Pr[D^{(F_i)_{1 \le i \le n}, (H_j)_{1 \le j \le m}} = 1] - \Pr[D^{(H'_i)_{1 \le i \le n}, S} = 1]| \le f(\eta)$$

where P_1 and P_2 are polynomials and f is a negligible function. The simulator S has oracle access to $(H'_i)_{1 \le i \le n}$.

In the game $G_0 = D^{(F_i)_{1 \le i \le n}, (H_j)_{1 \le j \le m}}$, the distinguisher interacts with the real functions F_i and the random oracles H_j from which the functions F_i are defined. In the game $G_1 = D^{(H'_i)_{1 \le i \le n}, S}$, the distinguisher interacts with independent random oracles H'_i instead of F_i, and with a simulator S, which simulates the behaviour of the random oracles H_j using calls to H'_i. Indifferentiability means that these two games are indistinguishable. We assume that the output length of each random oracle depends only on the security parameter.

B. Basic Lemmas

In this section, we show several basic indifferentiability lemmas, which are not specific to WireGuard.

Lemma 1 ([29, Lemma 2])**.** *If H is a random oracle, then the functions $H_1, \ldots, H_n$ defined as H on disjoint subsets $D_1, \ldots, D_n$ of the domain D of H are indifferentiable from independent random oracles, assuming one can determine in polynomial time to which subset D_i an element belongs.*

Lemma 2. *The concatenation of two independent random oracles with the same domain is indifferentiable from a random oracle.*

Lemma 3 ([29, Lemma 3])**.** *The truncation of a random oracle is indifferentiable from a random oracle.*

Lemmas 4 and 5 deal with the composition of two random oracle calls in sequence; they have been proved using

CryptoVerif. We extended CryptoVerif to be able to prove indistinguishability between two games given by the user. With this extension, CryptoVerif shows the indistinguishability result between the games G_0 and G_1 described in Section IV-A, which implies the indifferentiability result.

Lemma 4. *If $H_1 : S_1 \to S_1'$ and $H_2 : S_1' \times S_2 \to S_2'$ are independent random oracles, then H_3 defined by $H_3(x, y) = H_2(H_1(x), y)$ is indifferentiable from a random oracle.*

Lemma 5. *If $H_1 : S_1 \to S_1'$ and $H_2 : S_1' \times S_1 \to S_2'$ are independent random oracles, then $H_1' = H_1$ and H_2' defined by $H_2'(x) = H_2(H_1(x), x)$ are indifferentiable from independent random oracles.*

C. Indifferentiability of HKDF

The hkdf key derivation function is defined as follows [20]:

$$\begin{aligned}&\mathsf{hkdf}_n(salt, key, info) = k_1 \| \ldots \| k_n \text{ where}\\ &\quad prk = \mathsf{hmac}(salt, key)\\ &\quad k_1 = \mathsf{hmac}(prk, info \| i_0)\\ &\quad k_{i+1} = \mathsf{hmac}(prk, k_i \| info \| i + i_0) \text{ for } 1 \le i < n\end{aligned}$$

where $n \le 255$, and $i_0 = 0x01$ and i are of size 1 byte. In WireGuard, *info* is always empty, so we omit it in Section II.

We suppose that hmac is a random oracle, and we show that hkdf_n is indifferentiable from a random oracle, with the additional assumption that the calls to hmac use disjoint domains. (We show that this assumption is necessary and give a full proof of the result in the long version of the paper [15].) Let $\mathcal{S}$, $\mathcal{K}$, and $\mathcal{I}$ be the sets of possible values of *salt*, *key*, and *info* respectively, and $\mathcal{M}$ the output of hmac.

Lemma 6. *If* hmac *is a random oracle and $\mathcal{K} \cap (\mathcal{I} \| i_0 \cup \bigcup_{i=1}^{n-1} \mathcal{M} \| \mathcal{I} \| i + i_0) = \emptyset$ then hkdf_n with domain $\mathcal{S} \times \mathcal{K} \times \mathcal{I}$ is indifferentiable from a random oracle.*

This result extends the proof given for hkdf_2 in [29, Lemma 1]. Moreover, our proof is modular and partly made using CryptoVerif, thanks to the basic lemmas of Section IV-B.

Proof sketch. Since the domains are disjoint, by Lemma 1, the $(n + 1)$ calls to hmac are indifferentiable from independent random oracles $H_0, \ldots, H_n$. The constant $i + i_0$ can be removed from the arguments of H_{i+1} since it is fixed for a given H_{i+1}. By Lemma 5, the computation of $k_2 = H_2(H_1(prk, info), prk, info)$ is indifferentiable from a random oracle $k_2 = H_2'(prk, info)$. Applying this reasoning n times, the computation of k_i for $1 \le i \le n$ is indifferentiable from independent random oracles $k_i = H_i'(prk, info)$. By Lemma 2, concatenation of H_i' for $1 \le i \le n$ is indifferentiable from a random oracle H, so $\mathsf{hkdf}_n(salt, key, info) = k_1 \| \ldots \| k_n = H(prk, info)$, where $prk = H_0(salt, key)$. By Lemma 4, we conclude that hkdf_n is indifferentiable from a random oracle. □

D. Indifferentiability of a Chain of Random Oracle Calls

In this section, we prove the indifferentiability of a chain of random oracle calls defined as follows.

Definition 2 (Chain)**.** *Let $m \ge 1$ be a fixed integer, let C and C_j with $0 \le j \le m + 1$ be bitstrings of length l', let v_j with $0 \le j \le m$ be bitstrings of arbitrary length, let l be the length of the output of $\mathsf{H}(C_j, v_j)$, and let r_j with $0 \le j \le m$ be bitstrings of length $(l - l')$. (l and l' are functions of the security parameter.) We define the functions $\mathsf{chain}_n, 0 \le n < m$ and the function chain_m in the following way:*

$$\begin{aligned}&\mathsf{chain}_n(v_0, \ldots, v_n) =\\ &\quad C_0 = \mathsf{const}\\ &\quad \textit{for } j = 0 \textit{ to } n \textit{ do } C_{j+1} \| r_j = \mathsf{H}(C_j, v_j)\\ &\quad \textit{return } r_n\end{aligned} \tag{7}$$

$$\begin{aligned}&\mathsf{chain}_m(v_0, \ldots, v_m) =\\ &\quad C_0 = \mathsf{const}\\ &\quad \textit{for } j = 0 \textit{ to } m \textit{ do } C_{j+1} \| r_j = \mathsf{H}(C_j, v_j)\\ &\quad \textit{return } C_{m+1} \| r_m\end{aligned} \tag{8}$$

The functions $\mathsf{chain}_n, n < m$, have an output of length $(l - l')$, and the output length of chain_m is l.

Lemma 7. *If H is a random oracle, then chain_n, for $n \le m$, are indifferentiable from independent random oracles.*

We could probably prove this lemma for small values of m using CryptoVerif, but the generic result requires a manual proof because CryptoVerif does not support loops.

E. Application to WireGuard

WireGuard employs BLAKE2s [30] both directly as the function hash and indirectly as hash function in hmac and thus also in hkdf. The domains of these two uses are disjoint in WireGuard, as shown by an easy inspection of the length of the arguments. Then by Lemma 1, we can consider two independent random oracles, hash for the direct uses and hash' for the uses via hkdf. Since hash is a random oracle, it is a fortiori collision-resistant. We use that assumption for hash in our CryptoVerif proof.

Since hash' is a random oracle, hmac-hash' is indifferentiable from a random oracle by [31, Theorem 3]. Using Lemma 6, hkdf_n is indifferentiable from a random oracle. Since Lemma 7 assumes a chain of calls to the same hkdf_n function, we rewrite the chain of hkdf calls in WireGuard to use only calls to hkdf_3, as 3 is the maximum number of outputs needed, and discard the unused suffix: by definition of hkdf_n, this yields the same result. By Lemma 7, this computation can be replaced with:

$$\begin{aligned}k_1 \|\_ &\leftarrow \mathsf{chain}_1(v_0, v_1)\\ k_2 \|\_ &\leftarrow \mathsf{chain}_2(v_0, v_1, v_2)\\ \pi \| k_3 &\leftarrow \mathsf{chain}_6(v_0, v_1, v_2, v_3, v_4, v_5, v_6)\\ T^{\rightarrow} \| T^{\leftarrow} \|\_ &\leftarrow \mathsf{chain}_7(v_0, v_1, v_2, v_3, v_4, v_5, v_6, v_7)\end{aligned}$$

where $\|$ concatenates blocks of length the output length of hmac, _ is an unnamed block, and chain_i for $i \le 7$ are independent random oracles. (The result of chain_i for $i = 0, 3, 4, 5$ is not used.) The output of the random oracles can be truncated by Lemma 3 to avoid having to discard parts of the output. Moreover, in WireGuard, $v_7 = empty$, so $T^{\rightarrow}$

and $T^{\leftarrow}$ only depend on $v_0, \ldots, v_6$, as do π and k_3 in the previous line. By Lemma 2, we concatenate $\mathsf{chain}_6(\ldots)$ and $\mathsf{chain}_7(\ldots, empty)$, and thus obtain (6). The long version of the paper details this proof.

V. Modelling WireGuard

This section presents our model of the WireGuard protocol in CryptoVerif. We prove security properties for that model in Section VI.

A. Execution Environment

In our model, we consider two honest entities A and B. In the initial setup, we generate the static key pairs for these two entities and publish their public keys, so that the adversary can use them. After this setup, we run parallel processes that represent a number of executions of A and B polynomial in the security parameter.

The entities A and B can play both the initiator and responder role. These two entities can run WireGuard between each other, but also with any number of dishonest entities included in the adversary: for each session, the adversary sends to the initiator its *partner public key*, that is, the public key of the entity with which it should start a session; the adversary sends to the responder the set of partner public keys that it accepts messages from.

This setting allows us to prove security for any sessions between two honest entities, in a system that may contain any number of (honest or dishonest) other entities. We prove security for sessions in which A is the initiator and B is the responder. We do not explicitly prove security for sessions in which B is the initiator and A is the responder, but the same security properties hold by symmetry.

The processes for the entities A and B model the entire protocol, including the first two protocol messages, the key confirmation message from the initiator, and then a number of transport data messages polynomial in the security parameter, in both directions between initiator and responder. The model also includes random oracles, and we allow the adversary to call any of the random oracles that we use.

We consider 3 variants of this model:

Variant 1. This variant does not rely at all on the pre-shared key for proving security, so A and B receive a pre-shared key chosen by the adversary at the beginning of each execution. That allows the adversary to model both the absence of a pre-shared key (by choosing the value 0) or a compromised pre-shared key of its choice.

We model the dynamic compromise of the private static key of A (resp. B) by a process that the adversary can call at any time and that returns the private key of A (resp. B) and records the compromise by defining a particular variable, so that it can be tested in the security properties that we consider.

In WireGuard, four Diffie-Hellman operations and the pre-shared key contribute to the session keys. If the pre-shared key is not used or compromised, security is based on the four Diffie-Hellman operations. If one of them cannot be computed by the adversary, then the session keys are secret. Therefore, we consider all combinations of compromises but those where both keys on one side are compromised, that is:

1) A and B's private static keys may be dynamically compromised;
2) A's private static key may be dynamically compromised and B's private ephemeral key is compromised (by sending it to the adversary as soon as it is chosen);
3) B's private static key may be dynamically compromised and A's private ephemeral key is compromised;
4) A and B's private ephemeral keys are compromised.

We prove most security properties for *clean* sessions, that is, intuitively, sessions between honest entities; cleanliness is the minimal assumption needed to hope for security. A session of A is clean when either B's private static key is not compromised yet and A's partner public key is equivalent to B's static public key, or B's private static key is compromised and the public ephemeral key received by A is equivalent to a non-compromised ephemeral generated by B. B's session cleanliness is defined symmetrically. Intuitively, when B's private static key is not compromised, A can rely on that key to authenticate B, so A thinks she talks to B when she runs a session with B's public key. We consider a public key equivalent to B's public key rather than equal to B's public key to strengthen the properties: the authentication property shown in Section VI then implies that when A successfully runs a session with a partner public key equivalent to B's public key, then these two keys are in fact equal. (We find an interesting scenario concerning equivalent public keys and identity mis-binding with variant 3 of our model, we discuss it in §VI.) When B's private static key is compromised, A cannot authenticate B, but we can still prove security when the ephemeral key received by A has been generated by B. Like for static keys, when A successfully runs a session with a received ephemeral equivalent to an ephemeral generated by B, then these two ephemerals are in fact equal. (Instead of considering compromised ephemeral keys, we could also have modelled dishonestly generated ephemeral keys. We expect that some properties shown in §VI, such as session uniqueness, would not hold in this case.)

Variant 2. This variant relies exclusively on the pre-shared key for security. In that variant, we consider all private static and ephemeral keys as always compromised. We choose a pre-shared key randomly in the initial setup, and run sessions between A and B with that pre-shared key. In this model, A's partner public key is always B's public key and symmetrically, and these sessions between A and B are always considered clean. The adversary can run A's and B's sessions with other entities since A and B's private static keys are compromised and these sessions use a different pre-shared key.

Variant 3. In this variant, all keys are compromised: all private static and ephemeral keys are always compromised and the pre-shared key is chosen by the adversary for each session. This model is useful for proving properties that do not rely on session cleanliness, that is, properties that hold even for sessions involving dishonest participants.

With this model, we analyse the whole WireGuard protocol as it is, tying together the authenticated key exchange and the transport data phase. A similar approach was chosen by the creators of the Authenticated and Confidential Channel Establishment (ACCE) [32] model to analyse TLS. Instead of reasoning about key indistinguishability, ACCE looks at the security of the messages exchanged encrypted using the key. We do the same, for the key confirmation and all subsequent transport data messages.

In ACCE, the adversary has to choose one clean test session in which it tries to break security by determining the secret bit. In all other sessions, it is allowed to reveal the session keys. In our model, all clean sessions are test sessions, and we explicitly reveal the session keys in sessions that are not clean.

B. Modelling Tricks

Apart from the HKDF chains where we prove that the way we model them is indifferentiable from the real protocol in Section IV, we use the following modelling tricks:

Timestamps. CryptoVerif has no support for time, so instead of generating the timestamp, we input it from the adversary. In other words, we delegate the task of timestamp generation to the adversary. In order to model replay protection for the first message, the responder stores a global table (that is, a list) of triples containing the received timestamp, the partner public key for that session, as well as its own public key. (This is equivalent to having a distinct table of timestamps and partner public keys for each responder, represented by its public key.) The responder rejects the first message when the triple (received timestamp, partner public key, and responder public key) is already in the table.

Nonces for the AEAD scheme. The nonces in WireGuard are computed by incrementing a counter. CryptoVerif has no support for that, so we receive the desired value of the counter from the adversary. We guarantee that the same counter is never used twice in the same session for sending messages by storing all counters used for sending messages in a table of pairs (session index, counter), where the session index identifies the session uniquely: it indicates whether A or B is running, as initiator or as responder, and contains a unique integer index for the execution of that entity in that role. This is equivalent to having a distinct table of counters for each session. The message is not sent when the adversary provides a counter that is already in the table. We guarantee that the same counter is never used twice for receiving messages in the same way, using a separate table.

MACs. We omit the MACs mac_1 and mac_2 in our model. This simplifies the proof but preserves its soundness, since they can be computed and verified by the adversary: we deliver the messages without MACs to the adversary, and the adversary can add the MACs; conversely, the adversary can remove the MACs before delivering messages to the protocol model. We let the adversary choose the key R_r that the responder uses for computing cookies. All other elements needed to compute the MACs are public: constants and static public keys. We reintroduce the MACs in a separate model that we use for proving resistance against DoS.

Importantly, these modelling tricks increase the power of the adversary: the implementation done in WireGuard is a particular case of what the adversary can do in our model, in which the adversary chooses the current time as timestamp, increases the counter for sending messages at each emission, accepts incoming counters in a sliding window, and computes and verifies mac_1 and mac_2 by itself. As a result, a security proof in our model remains valid in WireGuard.

VI. Verification Results

In order to prove authentication properties, we insert events in our model, to indicate when each message is sent or received by the protocol. Specifically, we insert events sent1, sent2, sent_msg_initiator, and sent_msg_responder just before sending message 1, message 2, and transport messages on the initiator and responder sides respectively, and corresponding events rcvd1, rcvd2, rcvd_msg_responder, and rcvd_msg_initiator when these messages have been received and successfully decrypted. The event rcvd2 and the events for transport messages are executed only in clean sessions.

Mutual key and message authentication, resistance against KCI, resistance against replay from message 2. We show authentication for all messages starting from the second protocol message, by proving the following correspondence properties between events, in the first two variants of our CryptoVerif model of Section V-A:

$$\begin{aligned}
&\mathsf{inj\text{-}event}(\mathsf{rcvd2}(S_r^{pub}, E_i^{pub}, S_{i\,\text{🔒}}^{pub}, S_i^{pub}, ts_{\text{🔒}}, ts,\\
&\qquad\qquad E_r^{pub}, empty_{\text{🔒}}, T^{\rightarrow}, T^{\leftarrow}))\\
&\Rightarrow \mathsf{inj\text{-}event}(\mathsf{sent2}(S_r^{pub}, E_i^{pub}, S_{i\,\text{🔒}}^{pub}, S_i^{pub}, ts_{\text{🔒}}, ts,\\
&\qquad\qquad E_r^{pub}, empty_{\text{🔒}}, T^{\rightarrow}, T^{\leftarrow}))\,,
\end{aligned}$$

$$\begin{aligned}
&\mathsf{inj\text{-}event}(\mathsf{rcvd_msg_responder}(\\
&\qquad S_r^{pub}, E_i^{pub}, S_{i\,\text{🔒}}^{pub}, S_i^{pub}, ts_{\text{🔒}}, ts,\\
&\qquad E_r^{pub}, empty_{\text{🔒}}, T^{\rightarrow}, T^{\leftarrow}, N^{\rightarrow}, P_{\text{🔒}}, P))\\
&\Rightarrow \mathsf{inj\text{-}event}(\mathsf{sent_msg_initiator}(\\
&\qquad S_r^{pub}, E_i^{pub}, S_{i\,\text{🔒}}^{pub}, S_i^{pub}, ts_{\text{🔒}}, ts,\\
&\qquad E_r^{pub}, empty_{\text{🔒}}, T^{\rightarrow}, T^{\leftarrow}, N^{\rightarrow}, P_{\text{🔒}}, P))\,,
\end{aligned}$$

We also prove a third query (similar to the second one above) for transport data messages in the other direction, with events rcvd_msg_initiator and sent_msg_responder. A proven correspondence between two injective events (inj-event) means that each execution of the left-hand event corresponds to a distinct execution of the right-hand event.

The first query means that, if the initiator session is clean and the initiator has received the second message, then the responder sent it, and initiator and responder agree on their static and ephemeral public keys, session keys, timestamp, and communicated ciphertexts. This authenticates the responder to the initiator.

The second and third queries mean that, if the receiver session is clean and the receiver received a transport packet,

then a sender sent that transport packet, and the receiver and the sender agree on their static and ephemeral public keys, session keys, timestamp, sent plaintext, message counter, and communicated ciphertexts. In particular, for the key confirmation message, this authenticates the initiator to the responder. These queries also provide message authentication for the transport data messages.

All these properties hold when the pre-shared key is not compromised (variant 2 of Section V-A). They also hold when neither both S_i^{priv} and E_i^{priv} nor both S_r^{priv} and E_r^{priv} are compromised and the receiver session is clean; this is true, in particular, when the sender's static private key is not compromised yet (variant 1 of Section V-A).

The above queries include resistance against replays because the correspondences are injective: each reception corresponds to a *distinct* emission. They also include resistance against KCI attacks because the rcvd$*$ events are issued even if the receiver's static key has already been compromised: the receiver session is still clean in this case. Note that, for the responder, resistance against KCI attacks only starts after it receives the first data transport message. Indeed, the first protocol message is subject to a KCI attack: if the private static key of the responder (S_r^{priv}) is compromised, then the adversary can forge the first message and impersonate the initiator to the responder.

Secrecy and forward secrecy. We show secrecy of transport data messages in clean sessions by a left-or-right message indistinguishability game. In the initial setup, we randomly choose a secret bit. For each transport data message in a clean session, the adversary provides two padded plaintexts of the same length, and we encrypt one of them depending on the value of that bit. CryptoVerif proves the secrecy of that bit, in variants 1 and 2 of Section V-A, showing that the adversary cannot determine which of the two plaintexts was encrypted.

The secrecy query includes forward secrecy, because we allow dynamic compromise of static keys after the session keys have been established, if the ephemeral key of the same party is not compromised. This assumes that the parties delete the sessions' ephemeral and chaining keys after key derivation.

In variant 2 of our model, the query also shows secrecy provided the pre-shared key is not compromised, even if all other keys (static and ephemeral) are compromised. Our models do not consider the dynamic compromise of the pre-shared key, due to a limitation of CryptoVerif. We can still obtain forward secrecy with respect to the compromise of the pre-shared key using the following manual argument. As mentioned above, variant 2 of our model shows authentication when the pre-shared key is not compromised (all other keys are compromised in this model). This authentication property is preserved when the pre-shared key is compromised after the rcvd$*$ event, because the later compromise cannot alter the fact that the sent$*$ event has been executed. Furthermore, authentication guarantees that the ephemeral public key received by the initiator was generated by the responder and conversely. Variant 1 of our model then guarantees secrecy in this case, because the session is clean when the ephemeral received by the initiator was generated by the responder and conversely. Hence, we get the desired forward secrecy property: we have message secrecy when the pre-shared key is compromised after the session, and neither both S_i^{priv} and E_i^{priv} nor both S_r^{priv} and E_r^{priv} are compromised.

We cannot prove key secrecy for the session keys in the full protocol, because the session keys are used for encrypting transport data messages, and this allows an adversary to distinguish them from fresh random keys. Instead, we prove key secrecy for a model in which all transport data messages, including key confirmation, are removed. To prove this result, we need to strengthen the session cleanliness condition. Indeed, the first message is subject to a KCI attack, as mentioned above. Therefore, when the private static key of the responder is compromised, we additionally require that the ephemeral received by the responder is equivalent to one generated by the initiator. With this stronger cleanliness condition, we show that the session keys are secret, that is, the keys for various clean sessions are indistinguishable from independent random keys. We do not need this stronger cleanliness condition when we study the full protocol, since the key confirmation message protects the responder against KCI attacks.

Resistance against replay for the first message. We prove that the first message cannot be replayed but only if no static key is compromised when it is received. If S_i^{priv} were compromised, the adversary can impersonate the initiator as the sender of this message. If S_r^{priv} is compromised, we have a KCI attack, as described above. So we prove the following injective correspondence in a model where the static keys cannot be compromised but the ephemeral keys may be compromised, so we rely on the static-static Diffie-Hellman shared secret:

$$\begin{aligned}&\mathsf{inj\text{-}event}(\mathsf{rcvd1}(true, S_r^{pub}, E_i^{pub}, S_{i\,\text{🔒}}^{pub}, S_i^{pub}, ts_{\text{🔒}}, ts))\\ \Rightarrow\ &\mathsf{inj\text{-}event}(\mathsf{sent1}(\quad S_r^{pub}, E_i^{pub}, S_{i\,\text{🔒}}^{pub}, S_i^{pub}, ts_{\text{🔒}}, ts))\,.\end{aligned}$$

The first parameter of rcvd1 is $true$ if the public static key received by the responder with the first message is the public static key of the honest initiator: we prove this property only for sessions between honest peers. Replay protection is guaranteed by each timestamp being accepted only once. With this check removed, the first message can be replayed, but we still prove a non-injective correspondence between the two events, replacing inj-event by event in the query. This is a weaker property, meaning that, if an event rcvd1 has been executed, then at least one event sent1 with matching parameters has been executed before. Thus, even with the replay protection removed, we can prove that the origin of the first message cannot be forged in a model without static key compromise.

Correctness. Correctness means that, if the adversary does not modify the first two messages, then the initiator and responder share the same session keys and transcript hash H_7. Actually, it suffices that the adversary does not modify the ephemerals and ciphertexts of the first two messages. We

prove it with the following query:

$$\begin{aligned}&\mathsf{event}(\mathsf{responder\_corr}(E_i^{pub}, S_{i\,\text{🔒}}^{pub}, ts_{\text{🔒}}, E_r^{pub}, empty_{\text{🔒}},\\ &\quad T_r^{\rightarrow}, T_r^{\leftarrow}, H_{r7}))\\ &\wedge\ \mathsf{event}(\mathsf{initiator\_corr}(E_i^{pub}, S_{i\,\text{🔒}}^{pub}, ts_{\text{🔒}}, E_r^{pub}, empty_{\text{🔒}},\\ &\quad T_i^{\rightarrow}, T_i^{\leftarrow}, H_{i7})) \Rightarrow T_i^{\rightarrow} = T_r^{\rightarrow} \wedge T_i^{\leftarrow} = T_r^{\leftarrow} \wedge H_{i7} = H_{r7}\,.\end{aligned}$$

The events initiator_$*$ and responder_$*$ used in this query and in the following ones are issued after key derivation, in the initiator and responder respectively. Here, the two events given as assumptions guarantee that the adversary did not modify the ephemerals and ciphertexts of the first two messages, and the query concludes that the session keys and transcript hash must be equal. However, in our main models, CryptoVerif is currently unable to prove that the ciphertexts have not been created by the adversary, although this is true in the sessions considered by the correctness query. Thus, we created a separate model to prove correctness, in which the assumption is hard-coded by interleaving the initiator and responder in a single sequential process. In this model, we prove correctness even if all keys are compromised.

Session Uniqueness. First, we prove that there is a single initiator and a single responder session with a given $T^{\rightarrow}$ or $T^{\leftarrow}$. The query below shows that there cannot be two distinct initiator sessions with the same $T^{\rightarrow}$:

$$\begin{aligned}&\mathsf{event}(\mathsf{initiator\_uniq\_T}^{\rightarrow}(i_i, T^{\rightarrow}))\\ &\wedge\ \mathsf{event}(\mathsf{initiator\_uniq\_T}^{\rightarrow}(i_i', T^{\rightarrow})) \Rightarrow i_i = i_i'\,,\end{aligned}$$

where i_i, i_i' are replication indices: CryptoVerif assigns each execution of the initiator (or responder) process a unique replication index, so the query means that if we execute two events initiator_uniq_T$^{\rightarrow}$ with the same $T^{\rightarrow}$, then they have the same replication index $i_i = i_i'$, hence they belong to the same session. This query is proved in variant 3 of Section V-A, so the property holds even if all keys are compromised. (It relies on the choice of a fresh ephemeral at each session.) The queries for the other cases are similar.

Second, we show similarly that there is a single initiator and a single responder session for a given set of publicly transmitted protocol values.

Channel Binding. We prove channel binding with the query:

$$\begin{aligned}&\mathsf{event}(\mathsf{initiator\_H7}(params, H_7))\\ &\wedge\ \mathsf{event}(\mathsf{responder\_H7}(params', H_7)) \Rightarrow params = params'\end{aligned}$$

This query shows that if the initiator and responder have the same value of the session transcript H_7, then they share the same value of all session parameters $params$ (static and ephemeral public keys, timestamp, pre-shared key, session keys). This query is also proved in variant 3 of Section V-A, so the property holds even if all keys are compromised. (It relies on the collision resistance of hash.)

Identity Mis-Binding. For this property, we need to show that if an initiator and a responder session share the same session keys $T^{\rightarrow}$ and $T^{\leftarrow}$, then they share the same view on the ephemeral and static keys used in that session. This is formalised by the following query:

$$\begin{aligned}&\mathsf{event}(\mathsf{responder\_imb}(T^{\rightarrow}, T^{\leftarrow}, E_{i,rcvd}^{pub}, E_r^{pub}, S_{i,rcvd}^{pub}, S_r^{pub}))\\ &\wedge \mathsf{event}(\mathsf{initiator\_imb}(T^{\rightarrow}, T^{\leftarrow}, E_i^{pub}, E_{r,rcvd}^{pub}, S_i^{pub}, S_{r,rcvd}^{pub}))\\ &\Rightarrow E_i^{pub} = E_{i,rcvd}^{pub} \wedge E_r^{pub} = E_{r,rcvd}^{pub}\\ &\wedge S_i^{pub} = S_{i,rcvd}^{pub} \wedge S_r^{pub} = S_{r,rcvd}^{pub}\,.\end{aligned}$$

CryptoVerif proves it in variant 1 of our model, so it holds when neither both S_i^{priv} and E_i^{priv} nor both S_r^{priv} and E_r^{priv} are compromised. However, the proof fails when all static and ephemeral keys are compromised (variant 3 of our model): CryptoVerif can prove only the weaker property that $pow\_k\left(S_i^{pub}\right) = pow\_k\left(S_{i,rcvd}^{pub}\right)$ and $pow\_k\left(S_r^{pub}\right) = pow\_k\left(S_{r,rcvd}^{pub}\right)$. An adversary can indeed break the equality of public static keys in this case:

- The adversary instructs A to initiate a session to a public static key $S_r^{pub\prime}$ equivalent to our model's honest responder public static key: $pow\_k(S_r^{pub}) = pow\_k(S_r^{pub\prime})$ but $S_r^{pub} \neq S_r^{pub\prime}$. This is possible because S_r^{priv} is compromised. In this session, the adversary acts as responder, and because the ephemeral is also compromised, gets A's E_i^{priv}.
- The adversary now acts as initiator to start a session with B using a public static key $S_i^{pub\prime}$ equivalent to the honest initiator public static key: $pow\_k(S_i^{pub}) = pow\_k(S_i^{pub\prime})$ but $S_i^{pub} \neq S_i^{pub\prime}$. This is possible because S_i^{priv} is compromised. The adversary uses E_i^{priv} as ephemeral. The ephemeral of this session is also compromised, so the adversary gets E_r^{priv}.
- The adversary continues the session with A using the ephemeral E_r^{priv}.

If a pre-shared key is used, we assume that the adversary has the same pre-shared key with A (presenting itself with key $S_r^{pub\prime}$) and with B (presenting itself with $S_i^{pub\prime}$). The session keys $T^{\rightarrow}$ and $T^{\leftarrow}$ for these two sessions are computed as hashes of E_i^{pub}, $\mathsf{dh}(E_i^{priv}, S_r^{pub})$, $\mathsf{dh}(S_i^{priv}, S_r^{pub})$, E_r^{pub}, $\mathsf{dh}(E_i^{priv}, E_r^{pub})$, $\mathsf{dh}(S_i^{priv}, E_r^{pub})$, and psk. They are the same in both sessions, so the session keys are also the same.

This scenario, with a session between A and B' and one between B and A' that share the same session keys, is an instance of a bilateral unknown key-share attack [33] and of a key synchronisation attack [5]. It appears only when all static and ephemeral Diffie-Hellman keys are compromised, and hence should be considered a corner-case. However, we note that this scenario does not require the psk shared by A and B to be compromised, since this psk does not get used in the execution above. We suggest a possible fix of this identity mis-binding issue in Section VII.

Resistance against DoS. As described in Section II, WireGuard provides a cookie mechanism that a peer under load can use to enforce a round trip per sender address, and thus to bind a handshake message to a sender address; this permits per-address rate limiting. We model this mechanism in a separate

model in which a responder generates R_r, replies with a cookie $\tau = \mathsf{mac}(R_r, A_i)$ upon receipt of messages 1 from A_i with zero mac_2, and verifies mac_2 upon receipt of messages 1 with non-zero mac_2. The rest of the protocol is run by the adversary, which has the long-term static keys. In particular, we do not model the encryption of the cookie τ, but send it in the clear, assuming that the adversary carries out the encryption and decryption, which depend only on values it knows.

In this model, we prove that, if a responder under load accepts a handshake message from a sender with address A_i, then this sender passed through a round trip, that is, the responder did indeed previously generate a cookie for the address A_i. This formalised by the following query:

$$\begin{aligned} &\mathsf{event}(\mathsf{accepted\_cookie}(A_i, i_r, \tau, msg_\beta, mac_2)) \\ \Rightarrow\ &\mathsf{event}(\mathsf{generated\_cookie}(A_i, i_r, \tau))\,, \end{aligned}$$

where i_r is an index that uniquely identifies the key R_r used for generating the cookie. This query is proved under the assumption that mac is a pseudo-random function (PRF).

Identity Hiding. When the adversary has a candidate public key S_Y^{pub}, it can determine whether this public key is involved in WireGuard sessions, as already mentioned in the WireGuard specification [8]. In the first message, it can test whether $mac_1 = \mathsf{mac}(\mathsf{hash}(\mathsf{label}_{\mathsf{mac1}} \| S_Y^{pub}), msg_\alpha)$ and that reveals whether $S_Y^{pub} = S_r^{pub}$. A similar test on message 2 reveals whether $S_Y^{pub} = S_i^{pub}$. When an entity with public key S_m^{pub} sends a cookie reply, the adversary can try to decrypt the encrypted cookie $\tau_{\text{🔒}}$ with the key $\mathsf{hash}(\mathsf{label}_{\mathsf{cookie}} \| S_Y^{pub})$, the nonce $nonce$ (obtained from the cookie reply), and the associated data mac_1 (obtained from a previous message). If decryption succeeds, then the adversary knows that $S_Y^{pub} = S_m^{pub}$. In practice, the public keys of VPN endpoints may be easy to obtain: they are often published to subscribers on a web page. In such scenarios, WireGuard does not provide identity hiding.

If we consider the protocol without MACs and cookie reply, that is, basically the Noise protocol IKpsk2, we can obtain stronger identity protection guarantees, however with the additional assumption that the AEAD scheme also preserves the secrecy of the associated data. Indeed, if the AEAD scheme is only IND-CPA and INT-CTXT, then the adversary may obtain the associated data of the first ciphertext $S_i^{pub}{}_{\text{🔒}}$, that is, $\mathsf{hash}(\mathsf{hash}(H_0 \| S_r^{pub}) \| E_i^{pub})$. It can compare this value with $\mathsf{hash}(\mathsf{hash}(H_0 \| S_Y^{pub}) \| E_i^{pub})$ since E_i^{pub} is sent in the first message and H_0 is a constant. Thus, it can determine whether $S_r^{pub} = S_Y^{pub}$.

However, assuming that the AEAD scheme also preserves the secrecy of the associated data, we prove using CryptoVerif that the protocol without MACs and cookie reply satisfies the following identity hiding property: an adversary that has S_{A1}^{pub}, S_{A2}^{pub}, S_{B1}^{pub}, S_{B2}^{pub} cannot distinguish a configuration in which the entity with public key S_{A1}^{pub} initiates sessions with S_{B1}^{pub} from one in which the entity with public key S_{A2}^{pub} initiates sessions with S_{B2}^{pub}. ChaCha20Poly1305 indeed preserves the secrecy of the associated data, because it satisfies the stronger IND$-CPA property, which requires the ciphertext to be indistinguishable from random bits, as shown in [24].

We discuss possible solutions to strengthen the identity hiding for the protocol with MACs in Section VII.

Proof Guidance and Metrics. CryptoVerif needs to be manually guided to perform these proofs. We sketch the main instructions given to CryptoVerif for proving authentication and message secrecy in variant 1 of our model, with dynamic compromise of the private static keys. The guidance we give for other proofs follows similar ideas.

The first step is to distinguish cases. In the initiator A, we add a test to distinguish whether the partner public key is equivalent to B's static public key, and whether the received ephemeral is equivalent to an ephemeral generated by B. Similarly, in the responder B, we distinguish whether the partner public key is equivalent to A's static public key, and whether the received ephemeral is equivalent to an ephemeral generated by A. These case distinctions allow us to isolate Diffie-Hellman shared secrets that the adversary will be unable to compute because both shares come from honest participants. Then, we apply the random oracle assumption for the 3 random oracles chain'_6, chain'_2, chain'_1. For the arguments of these oracles that are Diffie-Hellman shared secrets in the protocol (and thus are in G_{sub}), we distinguish whether the argument received by the random oracle from the adversary is in G_{sub} before applying the random oracle assumption. (When it is not in G_{sub}, it cannot collide with a call coming from the protocol.) Next, we apply the gap Diffie-Hellman assumption; we split the keys generated by chain'_6 into 4 keys, and apply ciphertext integrity of the AEAD scheme. (For keys that the adversary may have after compromising the static keys, we apply a variant of the ciphertext integrity transformation that allows corruption.) This suffices to obtain authentication. Then we apply the IND-CPA property of the AEAD scheme to prove message secrecy.

In total, we give 36 instructions to CryptoVerif to perform this proof (not counting the instruction to display the current game), and CryptoVerif generates a sequence of 168 games. This proof takes 14 min, the proof of key secrecy with dynamic compromise of private static keys takes 16 min, and the one for identity hiding 18 min on one core of an Intel Xeon 3.6 GHz; these are our longest proofs.

VII. Discussion

WireGuard is a promising new VPN protocol that aims to replace IPsec and OpenVPN, and is being considered for adoption within the Linux kernel. We presented a mechanised cryptographic proof for a detailed model of WireGuard using the CryptoVerif prover. Our model accounts for the full Noise IKpsk2 secure channel protocol as well as WireGuard's extensions for stealthy operation and DoS resistance. We consider an arbitrary number of parallel sessions, with an arbitrary number of transport data messages. Furthermore, we base our proof on a precise model of the Curve25519 group.

We proved correctness, message and key secrecy, forward secrecy, mutual authentication, session uniqueness, channel

binding, and resistance against replay, key compromise impersonation, and denial of service attacks. In some cases, our analysis pointed out potential improvements in the protocol (which we did not prove secure using CryptoVerif):

Adding Public Keys to the Chaining Key Derivation. When analysing WireGuard for Identity Mis-Binding attacks, our analysis uncovered a corner case. Suppose all the Diffie-Hellman keys in a session between two hosts A and B were compromised, but the pre-shared key between them is still secret. Then the adversary can set up a man-in-the-middle attack where A thinks it is connected to B', B thinks it is connected to A', but in fact they are both connected to each other, in the sense that the two connections have the same traffic keys, even though they have different static keys.

In particular, once it has set up the session, the adversary can step away and let A and B directly communicate with each other, while retaining the ability to read and modify messages at will. Interestingly, this vulnerability only appears in our precise model of Curve25519; it cannot be detected under a classic Diffie-Hellman assumption.

Although this attack scenario may be quite unrealistic, it points to a theoretical weakness in the protocol that is easy to prevent with a simple modification. Noise IKpsk2 already adds ephemeral public keys to the chaining key derivation; we recommend that the static public keys be added as well. Alternatively, adding the full transcript hash to the traffic key derivation would also prevent this corner case.

Separately, it is also worth noting that adding public keys to the key derivation significantly helps with the cryptographic proof. For example, consider the Noise IK protocol, which is similar to IKpsk2 except that it does not use PSKs. IK does not mix the ephemeral keys into the chaining key, and it turns out that it is much harder for CryptoVerif to verify than IKpsk2, since we now have to reason about mis-matched ephemeral keys. In particular, even if we use a public PSK key of all-zeroes, the IKpsk2 protocol is easier to prove secure than IK. In fact, our recommendation is to add further contextual information to the key derivation. It would not only prevent theoretical attacks, but also make proofs easier.

Balancing Stealth and Identity Hiding. Our analysis also points out that the use of static public keys in mac_1 and mac_2 in WireGuard negatively affects the identity hiding guarantees provided by IKpsk2. This is a conscious trade-off that WireGuard makes to achieve stealthy operation [8]. However, in deployment scenarios where identity hiding is more important than stealth, we recommend that the protocol use a constant (say all-zeroes) instead of the static public keys to compute the MACs and cookies.

While it is difficult to preserve stealth while hiding the responder's identity, a modification to the protocol can still hide the initiator's identity. We recommend that the initiator should send a MAC key (along with the timestamp) in the first handshake message, and the responder should use this MAC key to compute mac_1 in the second handshake message. The initiator can verify this MAC to get DoS protection, but its static public key is kept hidden from a network adversary.

Table I: Security models (upper part) and properties analysed (lower part) in different works on WireGuard or Noise IKpsk2.

	Noise Explorer [36]	Suter-Dörig [37]	Girol [38]	Donenfeld, Milner [11]	Dowling, Paterson [12]	this work
verified protocol	Noise IKpsk2			WireGuard		
tool set	PV	T	T	T	m	CV
computational model	x	x	x	x	✓	✓
Curve25519 w/ eq. keys	x	x	x	x	x	✓
compromise static keys	✓	✓	✓	✓	✓	✓
compromise eph. keys	x	x	✓	✓	✓	✓
dishonest eph. keys	x	x	✓	x	x	x
compromise psk	✓	✓	✓	✓	✓	✓
compromise all keys	x	x	✓	✓	x	✓
both roles per static key	x	✓	✓	✓	✓	✓
mutual authentication	✓	✓	✓	✓	✓	✓
KCI	✓	✓	✓	✓	✓	✓
1st message replay	—	—	—	x	x	✓
transport data replay	x	✓	✓	x	x	✓
session uniqueness	x	✓	x	✓	✓	✓
channel binding	x	✓	x	x	x	✓
DoS resistance	—	—	—	x	x	✓
forward key secrecy	✓	✓	✓	✓	✓	✓
forward message secrecy	✓	✓	✓	x	x	✓
identity hiding	x	x	✓	✓[2]	x	✓
identity mis-binding	x	x	x	✓[1]	x	✓

Definitions differ between models.
T = Tamarin, PV = ProVerif, CV = CryptoVerif, m = manual.
✓= included, x = not included, — = not applicable.
1) The identity mis-binding issue we found was *not* found.
2) Weaker identity hiding property using a surrogate term.

Essentially, the MAC key acts as an in-session cookie.

Related Work. The use of formal verification tools to analyse real-world cryptographic protocols is now a well-established research area with hundreds of case studies (see e.g. [34]). CryptoVerif itself has been used to analyse modern protocols like Signal [29] and TLS 1.3 [35]. We conclude this paper by comparing our results with closely related work; Table I provides a condensed, high-level overview.

WireGuard itself has been formally analysed before. Donenfeld et al. [11] symbolically analyse the IKpsk2 key exchange protocol used by WireGuard for a number of security goals, including identity mis-binding and identity hiding. However, they do not model the MACs or the cookie mechanism, and hence they do not prove DoS resistance. Interestingly, their analysis concludes the absence of identity mis-binding attacks even if all keys are compromised, because their model does not include equivalent public keys. We disprove this property by considering a precise model of Curve25519.

Dowling et al. [12] present a manual cryptographic analysis of WireGuard. In particular, they prove key indistinguishability for the WireGuard handshake based on the PRF-ODH assumption in an extension of the eCK-PFS key exchange model. (Because of this difference in the used assumption, our mechanization cannot be used directly to find issues in proof steps; it is a different proof.) Key indistinguishability no longer holds once the key is used, so they prove security for a slightly modified variant of the IKpsk2 protocol that includes a

key confirmation message independent of the session keys. In contrast, our proof requires no changes to the protocol, since we use an ACCE-style model. Furthermore, [12] focuses only on the key exchange, and does not consider other properties like identity hiding or DoS resistance. Their analysis also does not find the identity mis-binding issue since they do not consider a scenario where all Diffie-Hellman keys are compromised.

Finally, the Noise Explorer tool [36] has been used to perform a comprehensive symbolic analysis of numerous Noise protocols using the ProVerif analyser. Noise Explorer can be used to find violations of secrecy and authentication properties for any protocol expressed in the language defined by Noise, using per-message authentication and confidentiality grades. It includes a symbolic analysis of Noise IKpsk2. A similar work has been done in Tamarin [37], [38].

Acknowledgements

We thank Jason A. Donenfeld, the author of WireGuard, and Nadim Kobeissi, and our anonymous reviewers for their helpful feedback on our work. This research was partly funded by the European Union's Horizon 2020 NEXTLEAP Project (grant agreement nº 688722), ERC CIRCUS (grant agreement nº 683032), ANR AnaStaSec (decision number ANR-14-CE28-0014-01), and ANR TECAP (decision number ANR-17-CE39-0004-03).

References

[1] S. Kent and K. Yao, "Security Architecture for the Internet Protocol," 2005, IETF RFC 4301.

[2] E. Rescorla, "The Transport Layer Security (TLS) Protocol Version 1.3," 2018, IETF RFC 8446.

[3] K. Bhargavan and G. Leurent, "On the practical (in-)security of 64-bit block ciphers: Collision attacks on HTTP over TLS and OpenVPN," in *ACM CCS'16*, 2016, pp. 456–467.

[4] B. Beurdouche, K. Bhargavan, A. Delignat-Lavaud, C. Fournet, M. Kohlweiss, A. Pironti, P. Strub, and J. K. Zinzindohoue, "A messy state of the union: Taming the composite state machines of TLS," in *IEEE S&P (Oakland)*, 2015, pp. 535–552.

[5] K. Bhargavan, A. Delignat-Lavaud, C. Fournet, A. Pironti, and P. Strub, "Triple handshakes and cookie cutters: Breaking and fixing authentication over TLS," in *IEEE S&P (Oakland)*, 2014, pp. 98–113.

[6] M. Marlinspike and T. Perrin, "The X3DH key agreement protocol," Nov. 2016, available at https://signal.org/docs/specifications/x3dh/.

[7] T. Perrin, "The Noise protocol framework," Jul. 2018, https://noiseprotocol.org/noise.html.

[8] J. A. Donenfeld, "WireGuard: Next generation kernel network tunnel," in *Network and Distributed System Security Symposium, NDSS*, 2017, we use the up-to-date whitepaper version for our analysis, which differs in how the MACs are defined: https://www.wireguard.com/papers/wireguard.pdf, Nov. 2nd, 2017, draft revision ceb3a49.

[9] B. Blanchet, "Modeling and verifying security protocols with the applied pi calculus and ProVerif," *Foundations and Trends in Privacy and Security*, vol. 1, no. 1-2, pp. 1–135, Oct. 2016.

[10] S. Meier, B. Schmidt, C. Cremers, and D. A. Basin, "The TAMARIN prover for the symbolic analysis of security protocols," in *Computer Aided Verification, CAV'13*, ser. LNCS, vol. 8044. Springer, 2013, pp. 696–701.

[11] J. A. Donenfeld and K. Milner, "Formal verification of the WireGuard protocol," 2018, https://www.wireguard.com/papers/wireguard-formal-verification.pdf.

[12] B. Dowling and K. G. Paterson, "A cryptographic analysis of the WireGuard protocol," in *Applied Cryptography and Network Security, ACNS 2018*, ser. LNCS, vol. 10892. Springer, 2018, pp. 3–21.

[13] B. Blanchet, "A computationally sound mechanized prover for security protocols," *IEEE Transactions on Dependable and Secure Computing*, vol. 5, no. 4, pp. 193–207, Oct.–Dec. 2008.

[14] ——, "Computationally sound mechanized proofs of correspondence assertions," in *IEEE CSF'07*, Jul. 2007, pp. 97–111, extended version available at http://eprint.iacr.org/2007/128.

[15] B. Lipp, B. Blanchet, and K. Bhargavan, "A mechanised cryptographic proof of the WireGuard virtual private network protocol," Inria, Research report 9269, Apr. 2019, https://hal.inria.fr/hal-02100345.

[16] A. Langley, M. Hamburg, and S. Turner, "Elliptic curves for security," Jan. 2016, IETF RFC 7748.

[17] M.-J. Saarinen and J.-P. Aumasson, "The BLAKE2 cryptographic hash and message authentication code (MAC)," 2015, IETF RFC 7693.

[18] Nir, Yoav and Langley, Adam, "ChaCha20 and Poly1305 for IETF Protocols," Jun. 2018, IETF RFC 8439.

[19] D. J. Bernstein, "Extending the Salsa20 nonce," 2011, https://cr.yp.to/snuffle/xsalsa-20110204.pdf.

[20] H. Krawczyk and P. Eronen, "HMAC-based extract-and-expand key derivation function (HKDF)," 2010, IETF RFC 5869.

[21] M. Bellare and P. Rogaway, "Random oracles are practical: a paradigm for designing efficient protocols," in *ACM CCS'93*. ACM Press, 1993, pp. 62–73.

[22] A. Luykx, B. Mennink, and S. Neves, "Security analysis of BLAKE2's modes of operation," *IACR Transactions on Symmetric Cryptology*, vol. 2016, no. 1, pp. 158–176, Dec. 2016.

[23] M. Bellare and C. Namprempre, "Authenticated encryption: Relations among notions and analysis of the generic composition paradigm," in *ASIACRYPT'00*, ser. LNCS, vol. 1976. Springer, Dec. 2000, pp. 531–545.

[24] G. Procter, "A security analysis of the composition of ChaCha20 and Poly1305," Cryptology ePrint Archive, Report 2014/613, 2014, https://eprint.iacr.org/2014/613.

[25] D. J. Bernstein, "The Poly1305-AES message-authentication code," in *FSE 2005*, ser. LNCS, vol. 3557. Springer, 2005, pp. 32–49.

[26] T. Okamoto and D. Pointcheval, "The gap-problems: a new class of problems for the security of cryptographic schemes," in *PKC 2001*, ser. LNCS, vol. 1992. Springer, Feb. 2001, pp. 104–118.

[27] J. Brendel, M. Fischlin, F. Günther, and C. Janson, "PRF-ODH: Relations, instantiations, and impossibility results," in *CRYPTO 2017*, ser. LNCS, vol. 10403. Springer, Aug. 2017, pp. 651–681.

[28] J.-S. Coron, Y. Dodis, C. Malinaud, and P. Puniya, "Merkle-Damgård revisited: How to construct a hash function," in *CRYPTO 2005*, ser. LNCS, vol. 3621. Springer, 2005, pp. 430–448.

[29] N. Kobeissi, K. Bhargavan, and B. Blanchet, "Automated verification for secure messaging protocols and their implementations: A symbolic and computational approach," in *IEEE EuroS&P'17*, Apr. 2017, pp. 435–450.

[30] J.-P. Aumasson, S. Neves, Z. Wilcox-O'Hearn, and C. Winnerlein, "BLAKE2: Simpler, smaller, fast as MD5," in *Applied Cryptography and Network Security*, ser. LNCS, vol. 7954. Springer, 2013, pp. 119–135.

[31] Y. Dodis, T. Ristenpart, J. Steinberger, and S. Tessaro, "To hash or not to hash again? (in)differentiability results for H^2 and HMAC," in *CRYPTO 2012*, ser. LNCS, vol. 7417. Springer, 2012, pp. 348–366, full version at https://eprint.iacr.org/2013/382.

[32] T. Jager, F. Kohlar, S. Schäge, and J. Schwenk, "On the security of TLS-DHE in the standard model," in *CRYPTO 2012*, ser. LNCS, vol. 7417. Springer, 2012, pp. 273–293.

[33] L. Chen and Q. Tang, "Bilateral unknown key-share attacks in key agreement protocols," *Journal of Universal Computer Science*, vol. 14, no. 3, pp. 416–440, Feb. 2008.

[34] B. Blanchet, "Security protocol verification: Symbolic and computational models," in *Principles of Security and Trust, POST'12*, ser. LNCS, vol. 7215. Springer, 2012, pp. 3–29.

[35] K. Bhargavan, B. Blanchet, and N. Kobeissi, "Verified models and reference implementations for the TLS 1.3 standard candidate," in *IEEE S&P (Oakland)*, 2017, pp. 483–502.

[36] N. Kobeissi, G. Nicolas, and K. Bhargavan, "Noise Explorer: Fully automated modeling and verification for arbitrary Noise protocols," in *IEEE EuroS&P 2019*, Jun. 2019, in this volume. The tool is available at https://noiseexplorer.com/.

[37] A. Suter-Dörig, "Formalizing and verifying the security protocols from the Noise framework," Bachelor's thesis, ETH Zürich, Nov. 2018, available at https://www.ethz.ch/content/dam/ethz/special-interest/infk/inst-infsec/information-security-group-dam/research/software/noise_suter-doerig.pdf.

[38] G. Girol, "Formalizing and verifying the security protocols from the Noise framework," Master's thesis, ETH Zürich, Mar. 2019, available at https://doi.org/10.3929/ethz-b-000332859.