Cryptographic Reverse Firewalls

Ilya Mironov[1] and Noah Stephens-Davidowitz[2]

[1] mironov@gmail.com
[2] Dept. of Computer Science, New York University. noahsd@gmail.com

Abstract. Recent revelations by Edward Snowden [PLS13, BBG13, Gre14] show that a user's own hardware and software can be used against her in various ways (e.g., to leak her private information). And, a series of recent announcements has shown that widespread implementations of cryptographic software often contain serious bugs that cripple security (e.g., [LHA+12, CVE14b, CVE14a, CVE14c]). This motivates us to consider the following (seemingly absurd) question: *How can we guarantee a user's security when she may be using a malfunctioning or arbitrarily compromised machine?* To that end, we introduce the notion of a *cryptographic reverse firewall* (RF). Such a machine sits between the user's computer and the outside world, potentially modifying the messages that she sends and receives as she engages in a cryptographic protocol.

A good reverse firewall accomplishes three things: (1) it *maintains functionality*, so that if the user's computer is working correctly, the RF will not break the functionality of the underlying protocol; (2) it *preserves security*, so that regardless of how the user's machine behaves, the presence of the RF will provide the same security guarantees as the properly implemented protocol; and (3) it *resists exfiltration*, so that regardless of how the user's machine behaves, the presence of the RF will prevent the machine from leaking any information to the outside world. Importantly, we do not model the firewall as a trusted party. It does not share any secrets with the user, and the protocol should be both secure and functional without the firewall (when it is implemented correctly).

Our security definition for reverse firewalls depends on the security notion(s) of the underlying protocol. As such, our model generalizes much prior work (e.g., [OO90, YY96, BBS98, BPR14a]) and provides a general framework for building cryptographic schemes that remain secure when run on compromised machine. It is also a modern take on a line of work that received considerable attention in the 80s and 90s (e.g., [Sim84, Sim85, BD91, Des90, Des94, BDI+96, BBS98]).

We show that our definition is achievable by constructing a private function evaluation protocol with a secure reverse firewall for each party. Along the way, we design an oblivious transfer protocol that also has a secure RF for each party, and a rerandomizable garbled circuit that is both more efficient and more secure than previous constructions. Finally, we show how to convert *any* protocol into a protocol with an exfiltration-resistant reverse firewall for all parties. (In other words, we provide a generic way to prevent a tampered machine from leaking information to an eavesdropper via *any* protocol.)

1 Introduction

Recent revelations of Edward Snowden show that powerful actors will go to remarkable lengths to obtain secret information. In particular, the National Security Agency has engineered a backdoor into a public cryptographic standard [PLS13, BBG13] and intercepted hardware as it was being delivered to customers in order to tamper with it [Gre14]. Meanwhile, multiple serious flaws have been uncovered in widely used implementations of cryptographic protocols, leaving many users vulnerable to simple but devastating attacks (e.g., [LHA+12, CVE14b, CVE14a, CVE14c]). The extreme complexity of modern cryptographic implementations makes it extremely difficult for experts (let alone the typical user) to detect such vulnerabilities, even when they are introduced innocently. Attackers that deliberately insert such vulnerabilities into hardware and software can make this even harder by using cryptographic methods to cover their tracks.

So, facing the disturbing (and quite real) possibility of a compromise that reaches inside one's communication platform, we consider the following seemingly paradoxical question: *Can we design cryptographic protocols that achieve meaningful security when the adversary may arbitrarily tamper with the victim's computer?*

To resolve this question, we present a strong and general notion of security in the presence of an active tampering adversary and show how to instantiate powerful cryptographic primitives in this model. Of

course, if Alice's computer simply chooses to replace her first message to Bob in some protocol with, for example, her secret business plans, we cannot hope to guarantee her security without some sort of help. Inverting the metaphor from network security, we propose and investigate the power of a (cryptographic) reverse firewall—an entity whose role is to protect cryptographic schemes and protocols from insider attacks. Informally, a cryptographic reverse firewall (RF) is a machine run by a third party (e.g., a security contractor hired by Alice's employer) that sits somewhere between Alice and the outside world and prevents Alice's computer from compromising her security by potentially modifying the messages that it sends and receives. In contrast to the standard firewall, the focus of a reverse firewall is on the inside of the perimeter. In particular, one important goal of reverse firewall is prevention of exfiltration attacks. Our primary contribution is the definition of reverse firewalls and the additional level of security that they bring to cryptographic protocols.

More specifically, we define three desirable properties of reverse firewalls. First, a reverse firewall should *maintain functionality*. I.e., if Alice's computer is behaving as it should, then the RF should not break the underlying functionality of the protocol. Second, a reverse firewall should *preserve security*. I.e., if the protocol without the RF present provides some security guarantee when Alice's computer behaves as it should, then the protocol *with* the RF present should provide this same security guarantee *regardless of how Alice's computer behaves*. Finally, a reverse firewall should *resist exfiltration*. Intuitively, an RF is exfiltration-resistant if Alice's tampered implementation cannot leak any information to the outside world through the firewall.

We defer much of the discussion of our definition to Section 2, where we introduce it formally. We emphasize, however, that the reverse firewall is *not* a trusted third party, and we do not rely solely on it for security. If Alice's implementation of the protocol is correct, then the protocol should be secure and functional without the firewall. In other words, we ask that the firewall *preserves* security, not that it *provides* it. In addition, the RF only has access to Alice's incoming and outgoing messages and any public parameters—not to Alice's state or input or any shared secrets. In effect, we place no more trust in the reverse firewall than we do in the communication medium. (We additionally require that firewalls be "stackable," so that one party may have arbitrarily many firewalls. Security is then guaranteed if just one of the firewalls is implemented correctly—or if Alice's own implementation is correct.)

Note that our security definition is quite strong, as it imagines the adversary "living inside of our computer." Consider, for example, a secure coin-flipping protocol in which Alice wishes to agree on a fair coin toss with Bob. Informally, the protocol is secure for Alice in the standard setting (i.e., without reverse firewalls) if Bob cannot bias the resulting coin toss alone. In our setting, we imagine both parties *working together* to bias the coin toss in Bob's favor. (Bob is adversarial as always, and in our setting, Bob may have also tampered with Alice's computer so that it is effectively "on Bob's side.") The only defense against this attack is a reverse firewall that can modify the messages that Alice sends and receives but must do so in a way that does not break the protocol when Alice and Bob are honest. (And, again, it must do so without access to any privileged information.)

In spite of this strength, we show that security in this model is achievable for very strong primitives. Indeed, we construct a two-round private function evaluation protocol that is secure in this model (Section 4). In particular, *each* party in this protocol has a corresponding secure reverse firewall. In other words, we show a relatively simple protocol that allows Alice and Bob to jointly and securely compute any circuit with the remarkable property that a reverse firewall can guarantee Alice's security even when Bob has tampered with her computer, and vice versa. This immediately shows that a very large class of two-party primitives can be realized securely in this model. The main ingredients for this protocol are an oblivious transfer scheme that itself has a secure reverse firewall for each party (Section 3) and a rerandomizable version of Yao's garbled circuit (Section 4.1). Our oblivious transfer protocol is a modified version of the Naor-Pinkas/Aiello-Ishai-Reingold protocol [NP01, AIR01]. Our rerandomizable garbled circuit is significantly more efficient than the construction of Gentry et al. [GHV10], and it achieves a stronger notion of rerandomizability. (See Section 1.1 for further comparison.)

Finally, in Section 5, we show a generic construction that can convert *any* protocol into a protocol with the same functionality that has an exfiltration-resistant reverse firewall. In other words, we provide a generic way to prevent a tampered machine from leaking information to an eavesdropper via any protocol. So, for the important special case in which Alice is primarily concerned with passive eavesdroppers, we show that *any* multiparty functionality can be implemented in our model.

Our protocols are described in full in terms of basic group operations, and we avoid using "heavy machinery" like non-interactive zero-knowledge proofs in our constructions. In particular, this means that our protocols are relatively simple and efficient and that the security of our constructions follows from relatively weak complexity-theoretic assumptions (namely, the slight variants of the decisional Diffie-Hellman assumption presented in Appendix B).

1.1 Related work

In this section we give a summary of related prior work, starting with the most directly comparable and recent literature. Given the size and the scope of existing work dealing with various models of insider attacks and mitigation strategies, our focus is on the similarities and differences between our work and prior art rather than a comprehensive review of all previous approaches.

Algorithm-substitution attacks. Motivated by the potential threat of powerful adversaries subverting implementations of cryptographic algorithms, Bellare, Paterson, and Rogaway recently proposed a formalization of the notion of resilience of symmetric encryption schemes to algorithm-substitution attacks (ASA) [BPR14a]. They observe that modern standards for symmetric encryption crucially rely on sender-chosen randomness to attain acceptable security levels. Since these standards do not include any mechanisms for ensuring that randomness used in the encryption stage is unbiased, they effectively enable a communication channel, which a corrupt implementation may use to leak information to an external party.

Bellare et al. define a general framework for ASA security, identifying two adversarial goals—avoiding detection and conducting surveillance. They cast several algorithm-substitution attacks against symmetric-key encryption in this framework, showing that widely deployed secure communication protocols, such as SSL/TLS, IPsec, and SSH, are vulnerable to these attacks. Furthermore, they present a universal, essentially undetectable attack effective against any stateless, randomized symmetric-key encryption scheme.

On the positive (defensive) side, Bellare et al. advocate using stateful, deterministic encryption schemes with unique ciphertexts as a counter-ASA measure. They construct a provably ASA-resilient encryption scheme based on the encode-then-encipher paradigm, and prove that all nonce-based schemes satisfying a natural non-degeneracy condition can be converted into stateful schemes with unique ciphertexts by choosing their nonces sequentially.

Our work extends Bellare et al. in several directions. First, we include in our treatment arbitrary two- and multi-party protocols, as opposed to just symmetric-key encryption. . Second, we shift our objective from developing primitives that are ASA-resilient by design to constructing protocols that are reverse-firewall-ready. Bellare et al. only achieve security against adversaries that do not break the functionality of the encryption scheme ("functionality-maintaining adversaries" in our terminology). By making a stronger assumption—availability of an uncorrupted reverse firewall—we are able to achieve stronger security guarantees, such as security against tampered implementations that break functionality.

Our results and techniques can be viewed as complementary. Whereas Bellare et al. make a strong case for suppressing "freedom of choice" in cryptographic primitives, we demonstrate that additional randomness can be injected by an intermediary in some protocols to achieve stronger security guarantees for a much wider range of primitives.

Collusion-free protocols and mediated collusion-free protocols. Informally, Lepinski, Micali, and shelat say that a multi-party protocol is collusion-free if the parties cannot communicate information about their private inputs to each other via the protocol [LMs05]. For example, a collusion-free protocol

for the game of poker allows parties to play a hand of poker, but it does not allow them to communicate information about their cards to other players during the hand.

This notion resembles our definition of exfiltration resistance in that it disallows subliminal communication via the protocol, but the two notions are incomparable. On one hand, the definition of Lepinski et al. is much stronger than ours because it does not allow the use of a third-party reverse firewall to prevent subliminal communication. On the other hand, it is much weaker because it specifies what specific information parties are not allowed to communicate. Indeed, their constructions involve a setup phase that is conducted before the parties are given their inputs, and the authors observe that this setup phase can be used as a subliminal channel. So, in our model, their protocols are completely insecure. Their constructions also require strong physical assumptions to ensure verifiable determinism.

To avoid the need for the setup phase and physical assumptions, Alwen, shelat, and Visconti introduce the mediated model for collusion-free protocols [AsV08]. In this model, all communication between the parties is routed through a mediator. Intuitively, the mediator rerandomizes the parties' messages in much the same way that our reverse firewalls do. However, the mediator is much more powerful than a reverse firewall in that (1) it intercepts all parties' messages and (2) it may exchange messages with the parties in any order. In contrast, our firewall modifies the messages sent and received by a single party in an online fashion, and we require our protocols to work without the firewall present. Because Alwen et al. give the mediator this additional power, they must explicitly model security against the mediator as a separate property of the protocol. In contrast, we get security "against the firewall" for free, as a natural consequence of the security of the underlying protocol. Their security definition is also stronger in the sense that it includes a strong notion of secure multi-party computation. While our notion of security preservation allows for such security, we intentionally do not require it in general.

Subliminal channels and divertible protocols. A long series of works explored the idea of subliminal channels in various cryptosystems (e.g., [Sim84, Sim85, BD91, Des90, Des94, BDI+96, BBS98]). Simmons [Sim84] introduced the notion by showing subliminal channels in various signature and authentication schemes. The underlying theme of this work is a story in which two prisoners, Alice and Bob, wish to communicate in some sanctioned way through the prison's warden (e.g., Alice wishes to tell Bob in some authenticated manner that she has not been harmed). The warden wishes to remove any subliminal messages from this communication (e.g., to prevent Alice from communicating escape plans to Bob).. The warden in this story is quite similar to our reverse firewall, and the notion of a subliminal-free channel is closely related to our notion of exfiltration resistance. Because of the wide body of work with a variety of definitions, results, and applications, we focus on a small portion that is most related to our work—divertible protocols.

Intuitively, a protocol is divertible if a warden sitting between Alice and Bob can rerandomize the messages of both parties so that (1) neither party is aware of the warden's existence and (2) neither party can distinguish between an interaction with a dishonest party with the warden in the middle and an interaction with an honest party. Okamoto and Ohta provided the first definition of divertibility for zero-knowledge proofs [OO90] (based on earlier definitions of subliminal-free zero-knowledge proofs), and Burmester et al. showed that all languages in NP have a divertible zero-knowledge proof [BD91, BDI+99]. These simple and elegant constructions immediately provide zero-knowledge proofs with reverse firewalls for all languages in NP.

Blaze, Bleumer, and Strauss showed how to generalize and strengthen the definition of divertibility to apply to any two-party cryptographic protocol [BBS98]. Indeed, their prescient definition comes close to our notion of a protocol with an exfiltration-resistant reverse firewall. We highlight three primary differences between their work and ours.

1. In our terminology, Blaze et al. consider only exfiltration resistance and not security preservation. In some applications (e.g., zero-knowledge proofs), the two properties are equivalent, but in many important applications (such as those that we consider in the sequel), the two properties are very different. (See Section 2.3 for further discussion of the distinction between these two properties.)

2. Blaze et al. implicitly assume that any dishonest version of the prover still provides valid proofs. (In our language, tampered provers must "maintain functionality.") This assumption is necessary for the prover of zero-knowledge proofs, but in general we can and should do better.
3. They consider only synchronous protocols with two parties and one warden. We consider asynchronous multi-party protocols in which each party may have its own firewall. By separating the warden into multiple firewalls and moving away from the synchronous model, our definition becomes much stronger, as our firewalls do not have the benefit of seeing all messages from all parties sent during a round before deciding how to modify them. (Indeed, Blaze et al. provide an example of a simple divertible key-agreement protocol. However, this protocol is *not* secure in our model because it crucially relies on the synchronous model of communication for its security.) We also find our more modular model to be more natural in our modern context, in which different parties may have different security needs.

Divertible protocols also differ from protocols with reverse firewalls in a number of more subtle ways. For example, Blaze et al. require that the warden is undetectable to either party. The protocols presented in the sequel achieve this notion of "transparency", but we intentionally do not require it as part of our definition.

In short, divertible protocols and subliminal-free channels were founded on a story that predates the concerns that motivate our work. Our more modern story, in which Alice and Bob (who need not be prisoners!) are concerned that their computers have been corrupted, leads naturally to our more general definition.

Kleptography. Young and Yung identified an important subclass of insider threats against cryptographic schemes, which they called *kleptographic attacks* [YY96]. The goal of a kleptographic attack is to leak a secret to an adversary who planted a malicious implementation of a cryptographic system on a victim's computer. The attack is asymmetric—the compromised implementation may carry the attacker's public key, but a private key is necessary in order to read from the subliminal channel. A secure kleptographic attack is undetectable as long as the system is accessed as a black box, and while it may be identified if one reverse engineers the implementation, this will only expose the attacker's public key. In particular, if multiple systems run the same compromised software stack, a successful reverse engineering effort of one such system will not help in breaching the security of others.

The (now withdrawn) NIST-standardized Dual Elliptic Curve Determnistic Random Bit Generator (Dual_EC_DRBG) is an example of a mechanism with a potential kleptographic backdoor [BV07, SF07].

Our adversarial model is a relaxation of the kleptographic attacker. We consider the possibility that the adversary may not worry about detection and is not concerned about a split-key solution.

Rerandomizable garbled circuits. Gentry, Halevi, and Vaikuntanathan construct a rerandomizable version of Yao's garbled circuit in order to build an "i-hop" homomorphic encryption scheme [GHV10]. Their construction is quite elegant, and its security is based on a slightly weaker assumption than ours (pure decisional Diffie-Hellman, as opposed to the slight variants presented in Appendix B). But, it does not work in our context. Informally, their circuit is rerandomizable when constructed honestly, but the rerandomization of a dishonestly constructed circuit can easily be distinguished from a freshly garbled circuit. (With negligible prboability, even the honest implementation can create circuits that in some sense "cannot be rerandomized.") In our context, in which we consider the possibility that the garbled circuit was constructed by a corrupted algorithm, this is a fatal flaw. We thus construct a new garbling scheme that can be rerandomized in a much stronger sense.

Our scheme (presented in Section 4) is also substantially more efficient than that of Gentry et al. The size of a single gate in their circuit is $O(\lambda^2)$ group elements, where λ is the security parameter, whereas our gates require only a constant number of group elements. As a consequence, our rerandomizable garbling scheme (which uses a trick inspired by Prabhakaran and Rosulek [PR07]) also implies a significantly more efficient implementation of i-hop homomorphic encryption.

Combiners. An alternative defense against untrusted implementations of a cryptographic primitive is to *combine* multiple implementations of the same primitive in some way so that the combined primitive

will be secure if a suitably large subset of the initial primitives are secure. This idea is quite common in the literature, and it was formalized by Harnik et al., who show that many primitives have elegant robust combiners [HKN+05].

Combiners solve a slightly different problem than reverse firewalls. Firewalls guarantee security when a user's system has been arbitrarily compromised, while combiners provide security only when the user already has access to at least one secure implementation of a primitive (and a secure implementation of the combiner itself!). Intuitively, combiners are applicable when multiple implementations of the same primitive exist that either (1) may have bugs in them or (2) rely on different unproven assumptions. In contrast, reverse firewalls work even when our implementations have been intentionally compromised.

2 Cryptographic reverse firewalls

We now present our general definition of a cryptographic reverse firewall that can be applied to a large class of primitives. This requires us first to define a cryptographic protocol in a (very general) way that suits our purposes. We note, however, that we describe the concrete schemes presented in the sequel in simpler terms. So, this level of generality is not necessary to understand the rest of the paper.

2.1 Cryptographic protocols

Definition 1 (Cryptographic protocol). *A* cryptographic protocol $\mathcal{P}$ *defines an interaction between stateful* parties $(P_1, \ldots, P_\ell)$*. First, a setup procedure* $\mathsf{setup}(1^\lambda)$ *is run, where* λ *is the security parameter. It returns a starting state for each party* $(\sigma_{P_i})_{i=1}^{\ell}$*, which we call their respective* input*; public parameters* ρ*; and a schedule of messages.*[1] *The parties proceed to send messages to each other according to the schedule. Each party has an associated* next message algorithm $\mathsf{next}_{P_i}(\sigma_{P_i})$ *that is called when it must output a message and a* message receipt algorithm $\mathsf{receive}_{P_i}(\sigma_{P_i}, m)$ *that is called upon receipt of a message to update the party's state. After the protocol is finished, each party runs its* output algorithm $\mathsf{output}_{P_i}(\sigma_{P_i})$ *and returns the result.*

We identify the protocol with its parties and setup procedure, $\mathcal{P} = (\mathsf{setup}, (P_i)_{i=1}^{\ell})$*, and we identify the parties with the algorithms that define them,* $P_i = (\mathsf{receive}_{P_i}, \mathsf{next}_{P_i}, \mathsf{output}_{P_i})$*. A complete record of all messages sent during a run of the protocol is a* transcript $\mathcal{T}$*.*

We call a run of a protocol a run with input I *if the parties' respective input and the public parameters are set to the values represented by* I*. We assume implicitly that the input* I *satisfies certain* validity requirements*.*

A protocol must satisfy functionality requirements $\mathcal{F}$*, which place constraints on the output of the parties for particular input* I*, and* security requirements $\mathcal{S}$*, which place constraints on the message distribution conditioned on specific input* I*. For our purposes, it will often be convenient to assign to each security requirement* $\mathcal{S}$ *a specific party who "is concerned with* $\mathcal{S}$*." For a party* P*, we say that a protocol is* secure for P *if all of* P*'s security requirements are met.*

For example, a one-out-of-two oblivious transfer (OT) protocol is a protocol between a sender, Alice, and a receiver, Bob. Alice's input is a pair of messages m_0, m_1, and Bob's input is a bit b. The protocol is functional if Bob's output is m_b. It is secure for Bob if for any valid messages m_0, m_1, no efficient algorithm playing the role of Alice can predict b with non-negligible advantage after the protocol is complete. (I.e., Alice is *oblivious* to Bob's bit b.) Intuitively, the protocol is secure for Alice if no efficient algorithm playing the role of Bob can "learn any information" other than m_0 or m_1 (but not both!). (See Appendix A.1 for a more formal definition.)

Below, we list some terminology and notation that will be useful in the next section.

[1] Note that we only consider protocols in which the message schedule is fixed by setup. Formally, this schedule determines the number of messages that a party must receive from each other party before sending each message. (E.g., Alice will send her second message after she has received three messages from Bob, two from Carol, and one from David.) We omit explicit reference to this schedule in the sequel as it will always be clear.

Definition 2 (Protocols and parties). *For a protocol $\mathcal{P} = (\mathsf{setup}, (P_i)_{i=1}^{\ell})$ satisfying functionality $\mathcal{F}$, input I, party P, index j, and index set $J \subseteq \{1, \ldots, \ell\}$,*

1. *$\mathcal{T} \leftarrow \mathcal{P}(I)$ denotes setting the variable $\mathcal{T}$ to the transcript obtained by a run of $\mathcal{P}$ with input I;*
2. *$\mathcal{P}_{P_j \Rightarrow P}$ is the protocol obtained by replacing party P_j with P in the protocol $\mathcal{P}$;*
3. *$\mathcal{P}_{J \Rightarrow P}$ is the protocol obtained by replacing all of the parties $\{P_j\}_{j \in J}$ with a single implementation of P in $\mathcal{P}$ (i.e., the parties $\{P_j\}_{j \in J}$ "collapse" into a single party P that has a single state σ_P);*
4. *if any party sends the special symbol $\perp$ as a message at any time, then the protocol immediately ends and, by definition, functionality has been violated; and*
5. *P* maintains *$\mathcal{F}$* for *P_j* in *$\mathcal{P}$ if $\mathcal{P}_{P_j \Rightarrow P}$ satisfies $\mathcal{F}$ with all but negligible probability over the random coins of the parties and* setup *procedure of P for any fixed input.*

When $\mathcal{F}$, P_j, and $\mathcal{P}$ are clear, we simply say that P maintains functionality*.*

2.2 Cryptographic reverse firewalls

Definition 3 (Cryptographic reverse firewall). *A* cryptographic reverse firewall *(RF) is a stateful algorithm $\mathcal{W}$ that takes as input its state and a message and outputs an updated state and message. For simplicity, we do not write the state of $\mathcal{W}$ explicitly.*

For a party $P = (\mathsf{receive}, \mathsf{next}, \mathsf{output})$ and reverse firewall $\mathcal{W}$, the composed party *is defined as*

$$\begin{aligned} \mathcal{W} \circ P := (&\mathsf{receive}_{\mathcal{W} \circ P}(\sigma, m) = \mathsf{receive}_P(\sigma, \mathcal{W}(m)), \\ &\mathsf{next}_{\mathcal{W} \circ P}(\sigma) = \mathcal{W}(\mathsf{next}_P(\sigma)), \\ &\mathsf{output}_{\mathcal{W} \circ P}(\sigma) = \mathsf{output}_P(\sigma)) \ . \end{aligned}$$

When the composed party engages in a protocol, the state of $\mathcal{W}$ is initialized to the public parameters ρ. If $\mathcal{W}$ is meant to be composed with a party P, we call it a reverse firewall for *P.*

Intuitively, an RF simply modifies Alice's incoming and outgoing messages. Alice of course does not want a reverse firewall to ruin her protocol's functionality when her internal implementation is correct. Indeed, we want something more than this. Alice's employer may wish to deploy multiple reverse firewalls (one internal firewall, one provided by a security contractor, etc.), and we do not want such "stacking" of firewalls to break functionality. The definition below captures this.

Definition 4 (Functionality-maintaining RFs). *For any reverse firewall $\mathcal{W}$ and any party P, let $\mathcal{W}^1 \circ P = \mathcal{W} \circ P$, and for $k \geq 2$, let $\mathcal{W}^k \circ P = \mathcal{W} \circ (\mathcal{W}^{k-1} \circ P)$.*

For a protocol $\mathcal{P}$ that satisfies some functionality requirements $\mathcal{F}$, we say that a reverse firewall $\mathcal{W}$ maintains *$\mathcal{F}$* for *P_j* in *$\mathcal{P}$ if $\mathcal{W}^k \circ P_j$ maintains $\mathcal{F}$ for P_j in $\mathcal{P}$ for any polynomially bounded $k \geq 1$. When $\mathcal{F}$, P_j, and $\mathcal{P}$ are clear, we simply say that $\mathcal{W}$* maintains functionality*.*

We emphasize that we are interested in reverse firewalls that *maintain* the functionality of an already functional protocol—protocols that do not function without the firewall are not nearly as interesting. We also note that the reverse firewalls described in the sequel actually achieve much stronger properties. In particular, they are all "transparent", so that the behavior of $\mathcal{W} \circ P$ is identical to the behavior of P if P is the honest implementation. And, $\mathcal{W} \circ P$ is functionality maintaining whenever P is (and not just when P is an honest implementation). While these properties seem desirable for many applications, we do not wish to exclude from our definitions firewalls that, for example, append a signature to each message that they send.

More interestingly, we would like a reverse firewall to protect Alice from an adversary that may have tampered with her computer. To that end, we ask that the firewall *preserves* the security properties of the underlying protocol. So, we are only interested in protocols that are already secure without the firewall present. Since this definition depends on the security properties of the underlying protocol, it provides a

general framework for the study of arbitrary cryptographic primitives in this model. Our strongest notion of security imagines a completely adversarial algorithm replacing Alice's implementation of the protocol and requires that security is still preserved even in this setting. Our weaker notion only considers tampered implementations that maintain functionality.

Definition 5 (Security-preserving RFs). *For a protocol* $\mathcal{P} = (\mathsf{setup}, (\mathsf{next}_{P_i}, \mathsf{receive}_{P_i}, \mathsf{output}_{P_i})_{i=1}^{\ell})$ *that satisfies some security requirements* $\mathcal{S}$ *and functionality* $\mathcal{F}$ *and a reverse firewall* $\mathcal{W}$*,*

1. $\mathcal{W}$ strongly preserves $\mathcal{S}$ for P_j in $\mathcal{P}$ *if the protocol* $\mathcal{P}_{P_j \Rightarrow \mathcal{W} \circ P^*_{\mathcal{A}}}$ *satisfies* $\mathcal{S}$ *for any probabilistic polynomial-time* $P^*_{\mathcal{A}}$*; and*
2. $\mathcal{W}$ weakly preserves $\mathcal{S}$ for P_j in $\mathcal{P}$ against $\mathcal{F}$-maintaining adversaries *if the protocol* $\mathcal{P}_{P_j \Rightarrow \mathcal{W} \circ P^*_{\mathcal{A}}}$ *satisfies* $\mathcal{S}$ *for any probabilistic polynomial-time* $P^*_{\mathcal{A}}$ *that maintains functionality* $\mathcal{F}$*.*

When $\mathcal{S}$*,* P_j*,* $\mathcal{P}$ *and* $\mathcal{F}$ *are clear, we simply say that* $\mathcal{W}$ strongly preserves security *or* weakly preserves security *respectively.*

One type of attack that particularly concerns us is *exfilitration*, in which Alice's corrupted computer attempts to leak some private information (e.g., secret business plans) to an adversary who has control over some (possibly empty) list of other parties J. We call security against such an attack *exfiltration resistance*, and we define it in terms of the game $\mathsf{LEAK}(\mathcal{P}, P_j, J, \lambda)$, presented in Figure 1. Intuitively, the game LEAK asks the adversary to distinguish between a tampered implementation of party P_j and an honest implementation. An exfiltration-resistant reverse firewall therefore prevents an adversary from even learning whether Alice's computer has been compromised—let alone her secret business plans.

proc. $\mathsf{LEAK}(\mathcal{P}, P_j, J, \mathcal{W}, \lambda)$
$(\sigma_{\mathcal{A}}, P^*_{\mathcal{A}}, P^*_{\mathcal{B}}, I) \leftarrow \mathcal{A}(1^{\lambda})$
$b \xleftarrow{\$} \{0,1\}$
`IF` $b = 1$, $P^* \leftarrow \mathcal{W} \circ P^*_{\mathcal{A}}$
`ELSE`, $P^* \leftarrow \mathcal{W} \circ P_j$
$\mathcal{T}^* \leftarrow \mathcal{P}_{P_j \Rightarrow P^*, J \Rightarrow P_{\mathcal{B}^*}}(I)$
$b^* \leftarrow \mathcal{A}(\sigma_{\mathcal{A}}, \mathcal{T}^*, \sigma_{P^*_{\mathcal{B}}})$
`OUTPUT` $(b = b^*)$

Fig. 1: $\mathsf{LEAK}(\mathcal{P}, P_j, J, \lambda)$, the exfiltration resistance security game for a reverse firewall $\mathcal{W}$ for party P_j in protocol $\mathcal{P}$ with corrupted parties J and security parameter λ. (Formally, the adversary represents a party by a collection of three (possibly randomized) circuits that implement the relevant functions $\mathsf{receive}$, next, and output.)

Definition 6 (Exfiltration-resistant RFs). *For a protocol* $\mathcal{P}$ *satisfying functionality* $\mathcal{F}$ *and a reverse firewall* $\mathcal{W}$*,*

1. $\mathcal{W}$ *is* $(\mathcal{P}, P_j, J)$-strongly exfiltration-resistant *if no PPT adversary* $\mathcal{A}$ *achieves advantage that is non-negligible in the security parameter* λ *in the game* $\mathsf{LEAK}(\mathcal{P}, P_j, J, \mathcal{W}, \lambda)$*; and*
2. $\mathcal{W}$ *is* $(\mathcal{P}, P_j, J)$-weakly exfiltration-resistant against $\mathcal{F}$-maintaining adversaries *if no PPT adversary* $\mathcal{A}$ *achieves advantage that is non-negligible in the security parameter* λ *in the game* $\mathsf{LEAK}(\mathcal{P}, P_j, J, \mathcal{W}, \lambda)$ *provided that the adversary's output* $P^*_{\mathcal{A}}$ *maintains* $\mathcal{F}$ *for* P_j*.*

When P_j*,* $\mathcal{P}$*, and* $\mathcal{F}$ *are clear, we simply say that* $\mathcal{W}$ *is* strongly exfiltration-resistant against J *or* weakly exfiltration-resistant against J *respectively. In the special case when* J *is empty, we say that* $\mathcal{W}$ *is* exfiltration-resistant against eavesdroppers.

This brings us to our strongest security notion.

Definition 7 (Robust RFs). *A cryptographic reverse firewall $\mathcal{W}$ is* robust for a party P_j in $\mathcal{P}$ with functionality requirements $\mathcal{F}$ and security requirements $\mathcal{S}$ *if it is $\mathcal{F}$-maintaining, strongly $\mathcal{S}$-preserving, and strongly exfiltration-resistant against the collection of all parties other than P_j in $\mathcal{P}$. We often simply say that $\mathcal{W}$ is robust when P_j, $\mathcal{P}$, $\mathcal{F}$, and $\mathcal{S}$ are clear.*

2.3 Discussion of the definitions

The relationship between exfiltration-resistant and security-preserving firewalls. For many natural notions of security, exfiltration resistance and security preservation are equivalent. For example, a reverse firewall preserves the semantic security of an encryption scheme if and only if it is exfiltration-resistant against an eavesdropper. However, for notions of security that do not promise privacy, a security-preserving firewall is not necessarily exfiltration-resistant. For example, a reverse firewall may preserve the binding property of a commitment scheme, but it may still allow information to leak out of a compromised machine. Even when a security requirement does imply some type of privacy, a firewall that preserves it may not be exfiltration-resistant. For example, the hiding property of a commitment scheme guarantees privacy during the commitment phase, but it certainly does not prevent information from leaking during the opening phase. In fact, it is relatively easy to construct reverse firewalls that strongly preserve the hiding property of commitment schemes (just use a rerandomizable commitment scheme), but it is provably impossible to construct a strongly exfiltration-resistant reverse firewall for the sender against the receiver in any commitment scheme! (Intuitively, the functionality of a commitment scheme allows the sender to communicate a message to the receiver. So, a reverse firewall cannot hope to simultaneously maintain functionality and prevent the sender from leaking information to the receiver.)

On the other hand, it may seem at first that an exfiltration-resistant reverse firewall always preserves security, since interaction with such an RF composed with an adversarially chosen circuit is, by definition, indistinguishable from interaction with an honest implementation. (Technically, we ask that the RF composed with an adversarially chosen circuit is indistinguishable from *the RF composed with* an honest implementation.) However, this is not always the case. For example, if security requirements are simulation-based or consider adversaries who have access to oracles or are computationally unbounded, then an exfiltration-resistant firewall may not preserve security.

Functionality-maintaining adversaries. Intuitively, our weaker security notions exclude the "more conspicuous" adversaries whose tampered circuit would be noticed by honest parties participating in the protocol with non-negligible probability. However, even our weakest adversaries may behave arbitrarily some negligible fraction of the time against honest parties. This distinction can be quite important in the context of security definitions that allow for the corruption of other players in the protocol. For example, consider an oblivious transfer protocol in which Bob's first message is uniformly random over some large set (as is the case in Section 3). A tampered implementation of Alice in this protocol may respond to one specific such message by, say, encoding the XOR of both of Alice's inputs into its response to Bob. Such an implementation can still be functionality-maintaining because this event rarely happens when Bob behaves honestly. But, the security definition of oblivious transfer requires that an adversary playing the role of Bob should not be able to learn the XOR of the inputs.

So, any reverse firewall that even *weakly* preserves Alice's security in such a model must somehow address this issue. In Section 3, we address this issue by composing a firewall for Alice that only works against tampered implementations that *always* maintain functionality with a firewall that is exfiltration-resistant *for Bob against Alice*. We expect this approach to be useful in future work.

Timing and scheduling issues. Our model does not explicitly account for the timing of messages. In practice, message timing is a natural channel, and a tampered implementation could of course use this to leak information and compromise Alice's security. So, any reverse firewall in the real world must account for this (e.g., by fixing the time between when it receives a message and when it forwards Alice's response). As the above discussion shows, in some cases, it might be necessary for the firewall to control the timing of both outgoing *and* incoming messages. In a protocol with more than two parties, this issue

naturally becomes more complicated. In such cases, protocol designers should consider the relative timing of messages from multiple parties' perspectives should and the order in which Alice receives messages from various parties.

3 Oblivious transfer

Naor and Pinkas and Aiello, Ishai, and Reingold independently developed very similar OT protocols whose security reduces immediately to DDH [NP01, AIR01]. We present a version of this protocol that is suitable for our setting. In particular, Alice's input is a pair of elements (m_0, m_1) in some group G of order p in which DDH is hard, and Bob's input is a bit b. Alice and Bob then engage in the protocol shown in Figure 2.

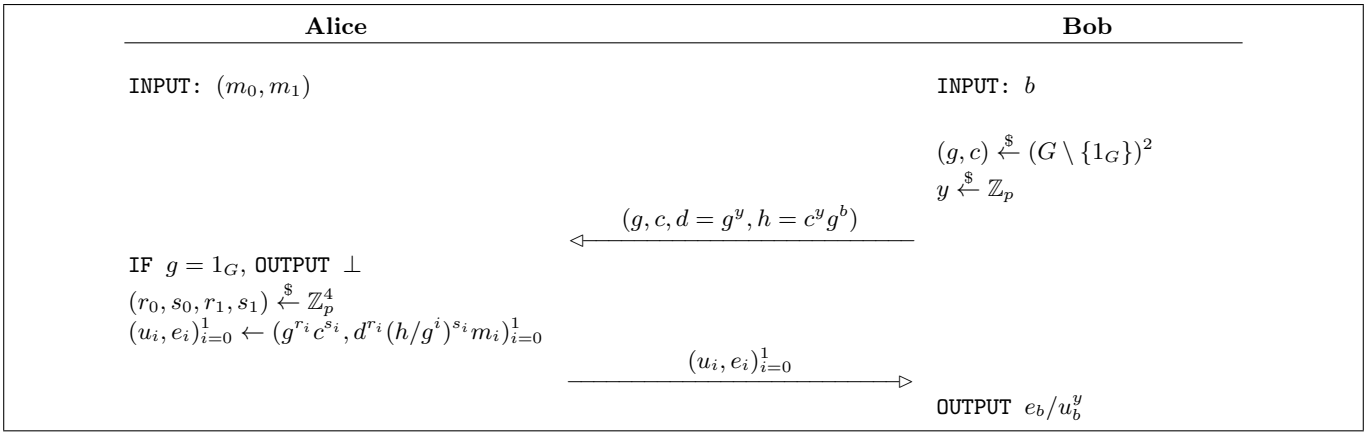

Fig. 2: A version of Naor-Pinkas/Aiello-Ishai-Reingold protocol for oblivious transfer.

Proposition 1. *The protocol shown in Figure 2 is correct and secure for both parties if DDH is hard in G.*

Proof. Correctness follows from the fact that $u_b^y = g^{r_b xy + s_b y} = (h/g^b)^{r_b} d^{s_b}$. Bob's security follows immediately from the DDH assumption in G.

To prove security for Alice, we note that if $(g, c, d, h) \neq (g, g^x, g^y, g^{xy+b})$ for some $x, y \in \mathbb{Z}_p$, then (u_b, e_b) is uniformly random. Indeed, note that

$$\log_g e_b = s_b(\log_g h - b) + r_b y = y \cdot \Big(\log_g u_b + s_b \Big(\frac{\log_g h - b}{y} - x \Big) \Big) .$$

It follows that u_b and e_b are distributed uniformly and independently unless $\log_g h - b = xy$. This allows us to construct, for any (not necessarily efficient) adversary $\mathcal{B}$ playing the role of Bob, an (inefficient) simulator $S_{\mathcal{B}}$ with access to the ideal functionality $\mathcal{F}$ that behaves as follows on input b.

1. $(\sigma, g, c, d, h) \leftarrow \mathcal{B}()$.
2. $(m_0, m_1) \xleftarrow{\$} G^2$.
3. If $(g, c, d, h) = (g, g^x, g^y, g^{xy+b})$ for $b \in \{0, 1\}$, set $m_b \leftarrow \mathcal{F}(b)$.
4. $(r_0, r_1, s_0, s_1) \xleftarrow{\$} \mathbb{Z}_p$.
5. $(u_i, e_i)_{i=0}^1 \leftarrow (g^{r_i} c^{s_i}, d^{r_i} (h/g^i)^{s_i} m_i)_{i=0}^1$.
6. Output $\mathcal{B}(\sigma, u_0, e_0, u_1, e_1)$.

It should be clear that the simulator's "message" $(u_i, e_i)_{i=0}^1$ is distributed identically to the message that $\mathcal{B}$ receives from Alice in the real protocol, and the result follows. □

We present reverse firewalls for both parties in our variant of the Naor-Pinkas/Aiello-Ishai-Reingold protocol and show that they are secure. Bob's reverse firewall is shown in Figure 3, and Alice's is shown in Figure 4. Alice's firewall by itself strongly prevents leaks against eavesdroppers. In order for it to weakly maintain security, it must be composed with Bob's firewall.

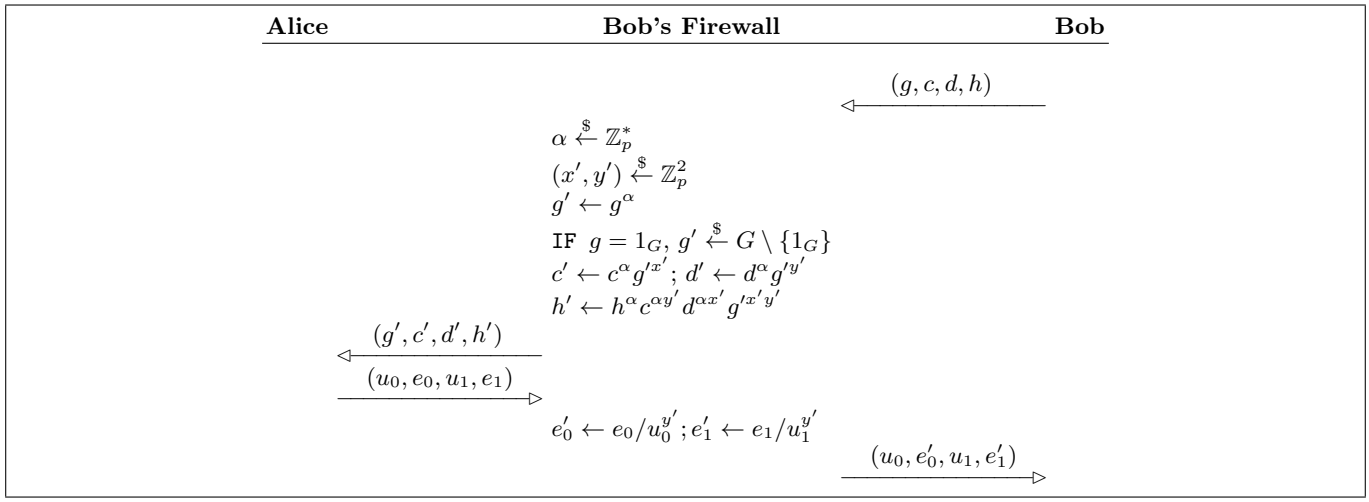

Fig. 3: Bob's reverse firewall for the protocol shown in Figure 2.

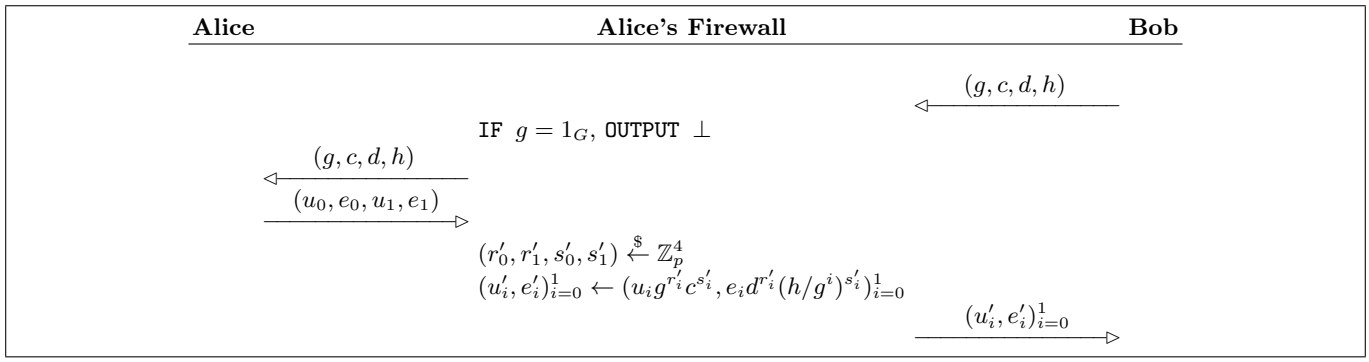

Fig. 4: Alice's reverse firewall for the protocol shown in Figure 2. It is strongly exfiltration-resistant, and it weakly preserves security when it is composed with the firewall shown in Figure 3

In Appendix C, we prove the following theorem.

Theorem 2. *Bob's reverse firewall $\mathcal{W}_\mathsf{B}$ shown in Figure 3 maintains correctness and is robust if the chosen-base DDH with a hint game is hard in G.*

Alice's reverse firewall $\mathcal{W}_\mathsf{A}$ shown in Figure 4 maintains correctness and is strongly exfiltration-resistant against an eavesdropper if DDH is hard in G. The composed firewall $\mathcal{W}_\mathsf{B} \circ \mathcal{W}_\mathsf{A}$ also weakly preserves security against Bob.

4 Private function evaluation

We now construct a private function evaluation scheme based on the oblivious transfer protocol from Section 3 and a version of Yao's garbled circuit. We assume that the reader is familiar with garbled circuits and this type of construction (see, for example, [BHR12]). Our key technical tool is a rerandomizable garbled circuit based on ElGamal encryption [ElG85], which may be of independent interest.

4.1 A rerandomizable garbled circuit

We wish to use the homomorphic properties and rerandomizability of ElGamal encryption to make a rerandomizable garbled circuit. But, a subtlety immediately arises: Yao's garbled circuit construction makes heavy use of encryptions of private keys (which can be used to decrypt more encryptions of private keys, etc.). However, in ElGamal encryption, private keys are elements in $\mathbb{Z}_p$ but messages are elements of a group G of order p in which DDH is hard. Our construction requires an efficient injective homomorphism from the key space to the message space. But, since DDH is easy in $\mathbb{Z}_p$, such a map cannot exist.

To get around this issue, we use a technique inspired by Prabhakaran and Rosulek [PR07]. In particular, for circuits of depth D, we need groups $G_1, \ldots, G_D$ of prime order $|G_d| = p_d$ such that G_d is a subgroup of $\mathbb{Z}^*_{p_{d+1}}$ and p_{d+1}/p_d is polynomially bounded[2] for $d < D$. In particular, this means that, given $g \in G_d$ and $h \in G_{d+1}$, the operation h^g is well-defined, and elements from G_d can therefore serve as private keys for ElGamal encryption over G_{d+1}.

Formally we say that a vertex z is at depth d in a circuit layout $\mathcal{L}$ if the longest path from an input vertex to z has length $d-1$, and we write $\mathsf{depth}(z) = d$. For ease of presentation, we assume that all edges in the circuit layout $\mathcal{L}$ are between vertices of adjacent depths (i.e., edges do not "skip levels") and that all output vertices have maximal depth D. With this simplification, we can use the group G_d to garble vertices at depth d. (Note that this restriction is not necessary, and the garbling scheme generalizes naturally to handle arbitrary circuits.)

Our garbling scheme for a circuit $\mathcal{C}$ is shown in Figure 5, and a schematic illustration of gate evaluation is provided in Figure 6. Alice can use the function Garble to garble a circuit $\mathcal{C}$ (represented by a collection of gate functions (f_z)), yielding a collection of ciphertexts (A_z) and *input tags* $(T_z^{(b)})_{z\in\mathcal{I},b\in\{0,1\}}$. Given a collection of ciphertexts (A_z) and input tags $(T_z^{(x_z)})_{z\in\mathcal{I}}$ corresponding to some input x, Bob can use the function Eval to compute $\mathcal{C}(x)$.

In particular, Garble assigns two tags $T_z^{(0)}$ and $T_z^{(1)}$ to each vertex z, which represent the vertex taking the value 0 and 1 respectively. $T_z^{(b)}$ is a uniformly random group element for all vertices that are not output gates, while the tags of output gates are simply encodings g_D^b of output bits. Intuitively, we want Bob to "only be able to learn" the tag corresponding to the value that each gate takes when $\mathcal{C}$ is evaluated on his input.

The function Garble represents each non-input gate z by A_z, a collection of ElGamal ciphertexts. The ciphertexts are encrypted under one of four private keys, each of which is the *product* of a tag from the gate's left parent $T_{L(z)}^{(b_L)}$ and a tag from the gate's right parent $T_{R(z)}^{(b_R)}$. The ciphertexts contain encryptions under the private key $T_{L(z)}^{(b_L)} \cdot T_{R(z)}^{(b_R)}$ of the tag $T_z^{f_z(b_L,b_R)}$ corresponding to the gate's output on some input (b_L, b_R). The ciphertexts are arranged randomly in the collection so that their order does not reveal any information about the circuit. So that Bob can know *which* ciphertext he should decrypt at each gate, together with each encrypted tag we also include a second ciphertext that encrypts a location bit τ (under the same key). Bob can then use the two location bits from a gate's left and right parent to know which ciphertext C_{τ_L,τ_R} to decrypt at the current gate.

[2] In practice, such chains of primes can be found efficiently. Indeed, the sequence defined by starting at $q_1 = 2$ and setting q_{i+1} to be the minimal prime with $q_{i+1} \equiv 1 \mod q_i$ is suitable. There are 497 primes in this sequence between 2^{1024} and 2^{6144}. The ratios between successive primes in this specific chain are conjectured to remain polynomially bounded in the length of the primes, and other chains with this property are conjectured to be abundant. See, for example, [FKL10].

The output of Garble is a collection of tags and location bits $(T_z^{(b)}, \tau_z^{(b)})_{z\in\mathcal{I},b\in\{0,1\}}$ for the input vertices together with the ciphertexts $(A_z)_{z\in V\setminus\mathcal{I}}$. The function Eval takes as input the ciphertexts (A_z) and the tags and location bits corresponding to some input x, $(T_z^{(x_z)}, \tau_z^{(x_z)})$, and it outputs $\mathcal{C}(x)$.

proc. Garble($\mathcal{C} = (f_z), g_1$)
FOR $d = 2, \ldots, D$,
  $g_d \xleftarrow{\$} G_d \setminus \{1_{G_d}\}$
FOR z in V,
  $b_z^* \xleftarrow{\$} \{0,1\}$
  IF $z \in \mathcal{O}$,
    $(T_z^{(0)}, T_z^{(1)}) \leftarrow (1_{G_D}, g_D)$
  ELSE,
    $d \leftarrow \mathsf{depth}(z)$
    $(T_z^{(0)}, T_z^{(1)}) \xleftarrow{\$} G_d^2$
FOR z in $V \setminus \mathcal{I}$,
  $A_z \xleftarrow{\$} \mathsf{GarbleGate}(z)$
FOR z in $\mathcal{I}$,
  $\tau_z^{(0)} \leftarrow b_z^*; \tau_z^{(1)} \leftarrow 1 \oplus b_z^*$
OUTPUT $((T_z^{(b)}, \tau_z^{(b)}), (A_z), (g_d))$

proc. GarbleGate(z)
$d \leftarrow \mathsf{depth}(z)$
FOR (b_L, b_R) in $\{0,1\}^2$,
  $k \leftarrow T_{L(z)}^{(b_L)} \cdot T_{R(z)}^{(b_R)}$
  $\tau_L = b_L \oplus b_{L(z)}^*$
  $\tau_R = b_R \oplus b_{R(z)}^*$
  $\boldsymbol{\eta} \leftarrow (\tau_L, \tau_R)$
  $h_{\boldsymbol{\eta}} \leftarrow g_d^k$
  $b \leftarrow f_z(b_L, b_R)$
  $(r, s) \xleftarrow{\$} \mathbb{Z}_{p_d}^2$
  $(u_{\boldsymbol{\eta}}, e_{\boldsymbol{\eta}}) \leftarrow (g_d^r, h_{\boldsymbol{\eta}}^r T_z^{(b)})$
  $\tau \leftarrow b \oplus b_z^*$
  $(v_{\boldsymbol{\eta}}, w_{\boldsymbol{\eta}}) \leftarrow (g_d^s, h_{\boldsymbol{\eta}}^s g_d^\tau)$
OUTPUT $(h_{\boldsymbol{\eta}}, u_{\boldsymbol{\eta}}, e_{\boldsymbol{\eta}}, v_{\boldsymbol{\eta}}, w_{\boldsymbol{\eta}})$

proc. Eval($(T_z, \tau_z), (A_z), (g_d)_{d=1}^D$)
FOR $d = 1, \ldots, D$,
  IF $g_d = 1_{G_D}$, OUTPUT $\perp$
FOR z in $V \setminus \mathcal{I}$,
  $(T_z, \tau_z) \leftarrow \mathsf{EvalGate}(z)$
FOR z in $\mathcal{O}$,
  IF $T_z \notin \{1_{G_D}, g_D\}$, OUTPUT $\perp$
OUTPUT $(\log_{g_D} T_z)_{z\in\mathcal{O}}$

proc. EvalGate(z)
$k \leftarrow T_{L(z)} \cdot T_{R(z)}$
$d \leftarrow \mathsf{depth}(z)$
PARSE $(h_{\boldsymbol{\eta}}, u_{\boldsymbol{\eta}}, e_{\boldsymbol{\eta}}, v_{\boldsymbol{\eta}}, w_{\boldsymbol{\eta}}) \leftarrow A_z$
$\boldsymbol{\eta} \leftarrow (\tau_{L(z)}, \tau_{R(z)})$
IF $h_{\boldsymbol{\eta}} \neq g_d^k$, OUTPUT $\perp$
$(T_z, B_z) \leftarrow (e_{\boldsymbol{\eta}}/u_{\boldsymbol{\eta}}^k, w_{\boldsymbol{\eta}}/v_{\boldsymbol{\eta}}^k)$
IF $B_z \notin \{1_{G_d}, g_d\}$, OUTPUT $\perp$
OUTPUT $(T_z, \log_{g_d} B_z)$

Fig. 5: Our garbled circuit scheme with input circuit $\mathcal{C}$ of depth D and a publicly known layout. (We assume for simplicity that all edges in the circuit are from vertices of depth d to vertices of depth $d+1$ and that all output vertices have maximal depth.)

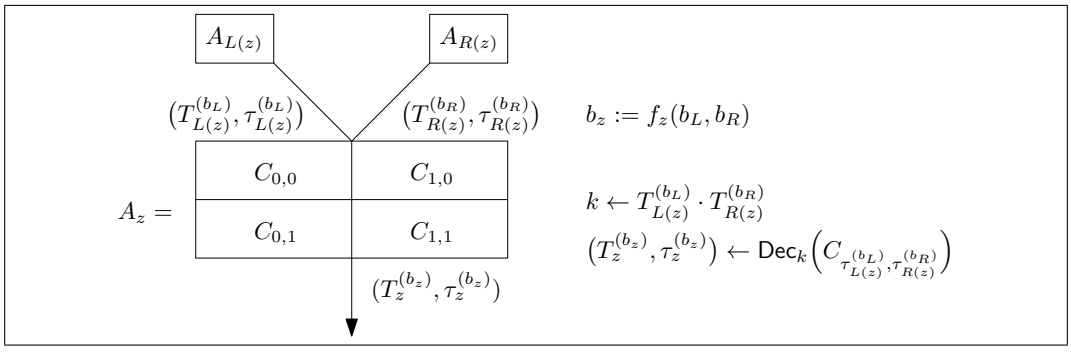

Fig. 6: A schematic representation of the evaluation of a single gate. The bits b_L, b_R, and b_z are not known to the evaluation algorithm.

4.2 PFE from garbled circuits and OT

With the garbling scheme from Figure 5 and the oblivious transfer scheme from Section 3, we can build a private function evaluation protocol, which we present in Figure 7. We note that the protocol can be optimized so that Bob sends his oblivious transfer messages in one batch. With this optimization, the protocol requires only two messages. (Bob sends (g_1, c) and his oblivious transfer requests in a single message. Alice then sends her responses to the oblivious transfer requests and the garbled circuit, also in

a single message.) The proof of security as well as the reverse firewalls and their corresponding proofs of security can be naturally modified to accommodate this change.

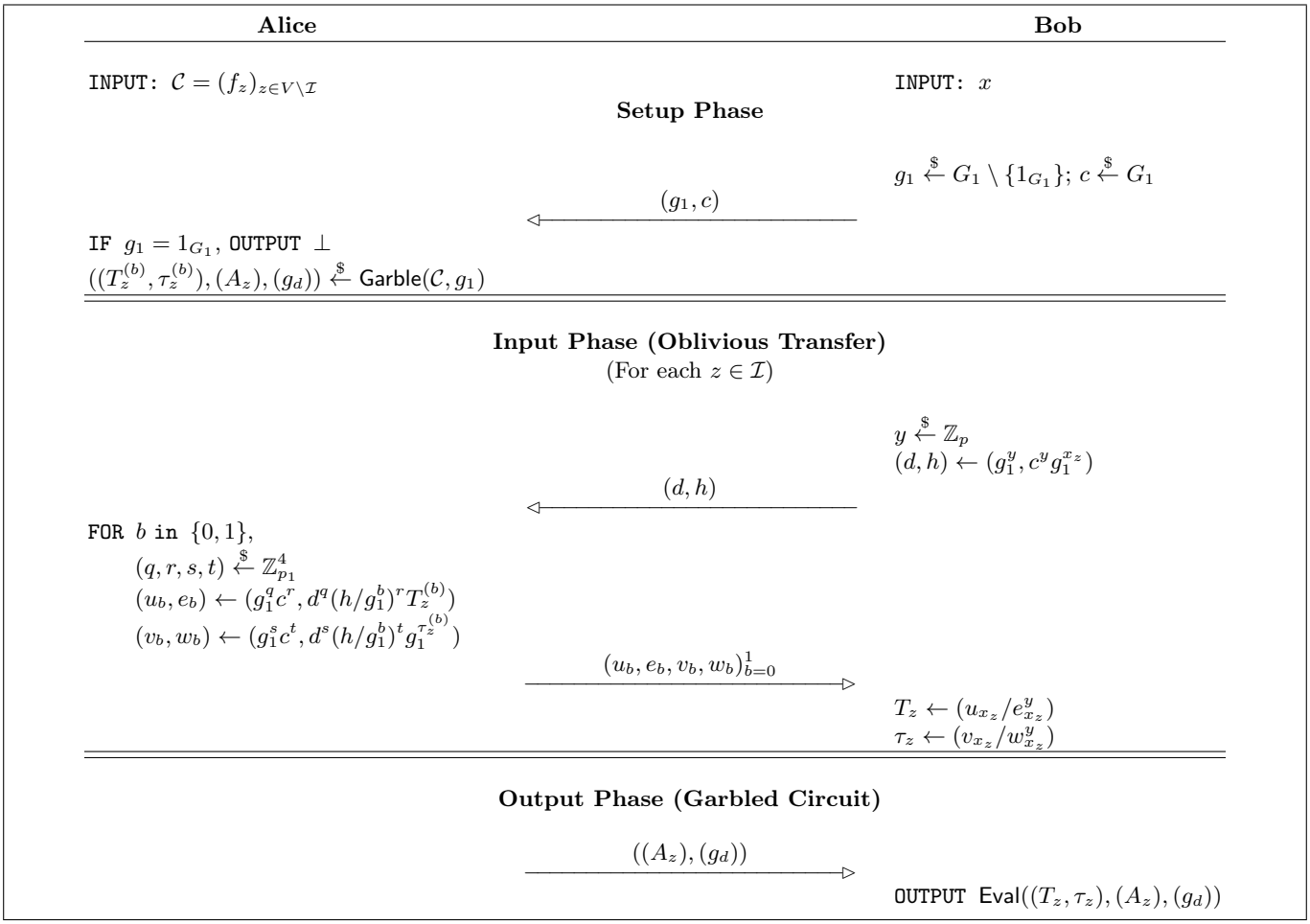

Fig. 7: A private function evaluation protocol using our oblivious transfer protocol from Section 3 and our garbled circuit scheme shown in Figure 5. (See Figure 5 for the functions Garble and Eval.)

We prove the following proposition in Appendix E.

Proposition 3. *The private function evaluation protocol shown in Figure 7 is correct and secure for both Alice and Bob if DDH is hard in the* (G_i).

4.3 Reverse firewalls for PFE

Bob's reverse firewall is very similar to his reverse firewall in Section 3. For completeness, we present it in Figure 11 in Appendix D. Alice's reverse firewall is shown in Figure 8. It makes use of a function $\mathsf{Rerand}_{\mathsf{Garble}}$ that rerandomizes garbled circuits. This procedure is rather complicated because our garbled circuits necessarily have many moving parts: location bits τ; the ordering of the ciphertexts, which is determined by the location bits; tags T; the keys, which are products of the tags; and the randomness used to encrypt the tags and location bits. Our task is to rerandomize all of this without breaking functionality. Below, we describe the intuition behind the rerandomization procedure. We provide the (rather complex) pseudocode in Figure 10 in Appendix D.

Ideally, in order to rerandomize the tags in the garbled circuit, we would simply use the malleability of ElGamal to multiply each tag $T_z^{(b)}$ by a uniformly random mask $\mathcal{R}_z^{(b)}$. But (as Gentry et al. observe in a similar context [GHV10]), the firewall cannot know which tags are used to generate which keys—so maintaining consistency between tags and keys would not be possible with this approach. We can, however, multiply *both* $T_z^{(0)}$ and $T_z^{(1)}$ by a single uniformly random mask $\mathcal{R}_z$ (i.e., we can multiply all the corresponding ciphertexts by $\mathcal{R}_z$). Since we apply this operation to both tags, we can easily maintain consistency (after noting that an ElGamal encryption under a private key k can be easily converted into an encryption under $k \cdot k'$ without knowing k). But, we need a second degree of freedom per gate (otherwise, $T_z^{(0)}/T_z^{(1)}$ would remain unchanged).

We find our solution in the location bits. In particular, for each ciphertext $(h, u, e, v, w) = (h, g^r, h^r T, g^s, h^s g^\tau)$, we use the homomorphic property of ElGamal encryption to multiply the tag T by $g^{\beta_z \tau}$ for uniformly random β_z. Since the location bit τ encodes which ciphertexts will be encrypted under a key generated from T, we do not need to know which tags correspond to which keys to maintain consistency between the tags and keys—we just need to know that whichever tag had a corresponding location bit $\tau = 1$ was multiplied by g^{β_z}. The complete mask is therefore $\mathcal{R}_z g^{\beta_z \tau}$ for the appropriate value of τ.[3]

We also need a way to rerandomize the location bits themselves. Recall that the location bits τ are encrypted as $(v, w) = (g^s, h^s g^\tau)$. In order to rerandomize them, we note that $(v^{-1}, w^{-1} g) = (g^{-s}, h^{-s} g^{1-\tau})$ is an encryption of $\tau \oplus 1$. We can therefore flip the location bits without knowing their values.

To maintain consistency with the rerandomization of the oblivious transfer rounds, $\mathsf{Rerand}_{\mathsf{Garble}}$ takes as input the masks that should be used to rerandomize the input tags. In particular, the procedure takes as input a collection of garbled gates $(A_z)_{z \in V \setminus \mathcal{I}}$, group elements $(g_d)_{d=1}^D$, and masks for the input vertices $(\mathcal{R}_z, \beta_z, b_z^*)_{z \in \mathcal{I}}$, and it outputs new ciphertexts (A'_z) and new group elements (g'_d). The masks $\mathcal{R}_z$ and β_z are used to mask tags as described above, and the bit b_z^* determines whether the location bits $\tau_z^{(b)}$ should be flipped. (The masks for non-input vertices are selected uniformly at random by the rerandomization procedure.)

In Appendix F, we prove the following theorem.

Theorem 4. *The reverse firewall for Bob shown in Figure 11 is robust if the DDH with a hint game is hard in G_1.*

The reverse firewall for Alice shown in Figure 8 maintains correctness, weakly preserves Alice's security, and is strongly exfiltration-resistant against an eavesdropper if non-uniform DDH is hard in the (G_d).

5 A generic construction for strong exfiltration resistance against eavesdroppers

We now show that *any* protocol can be converted into a protocol that has a reverse firewall for each party that is strongly exfiltration-resistant against eavesdroppers. The resulting protocol will have at most one additional (broadcast) message per party (or fewer than two additional messages per party in the non-broadcast model). For all of the primitives that we consider in this paper, the resulting protocol will also satisfy the same security requirements as the original protocol. We cannot say that the resulting protocol will always satisfy the same security requirements for arbitrary primitives because security requirements are quite a general notion. For example, a security requirement could specifically ask that a protocol does *not* have an exfiltration-resistant reverse firewall.

In order to achieve this, the key idea is to use a public-key encryption scheme that is rerandomizable and has a *rerandomizable key*. I.e., a reverse firewall should be able to convert any public key into a uniformly random public key in such a way that it can also convert messages encrypted under the resulting key

[3] Of course, a tampered implementation playing the role of Alice may not produce ciphertexts of the correct form $(h, g^r, h^r T, g^s, h^s g^\tau)$ where τ is a bit. But, our rerandomization algorithm will still multiply each key of the children of the node z by the mask $\mathcal{R}_z g^{\beta_z \tau}$ for the appropriate value of τ. This rerandomization of *keys* is what makes the scheme secure.

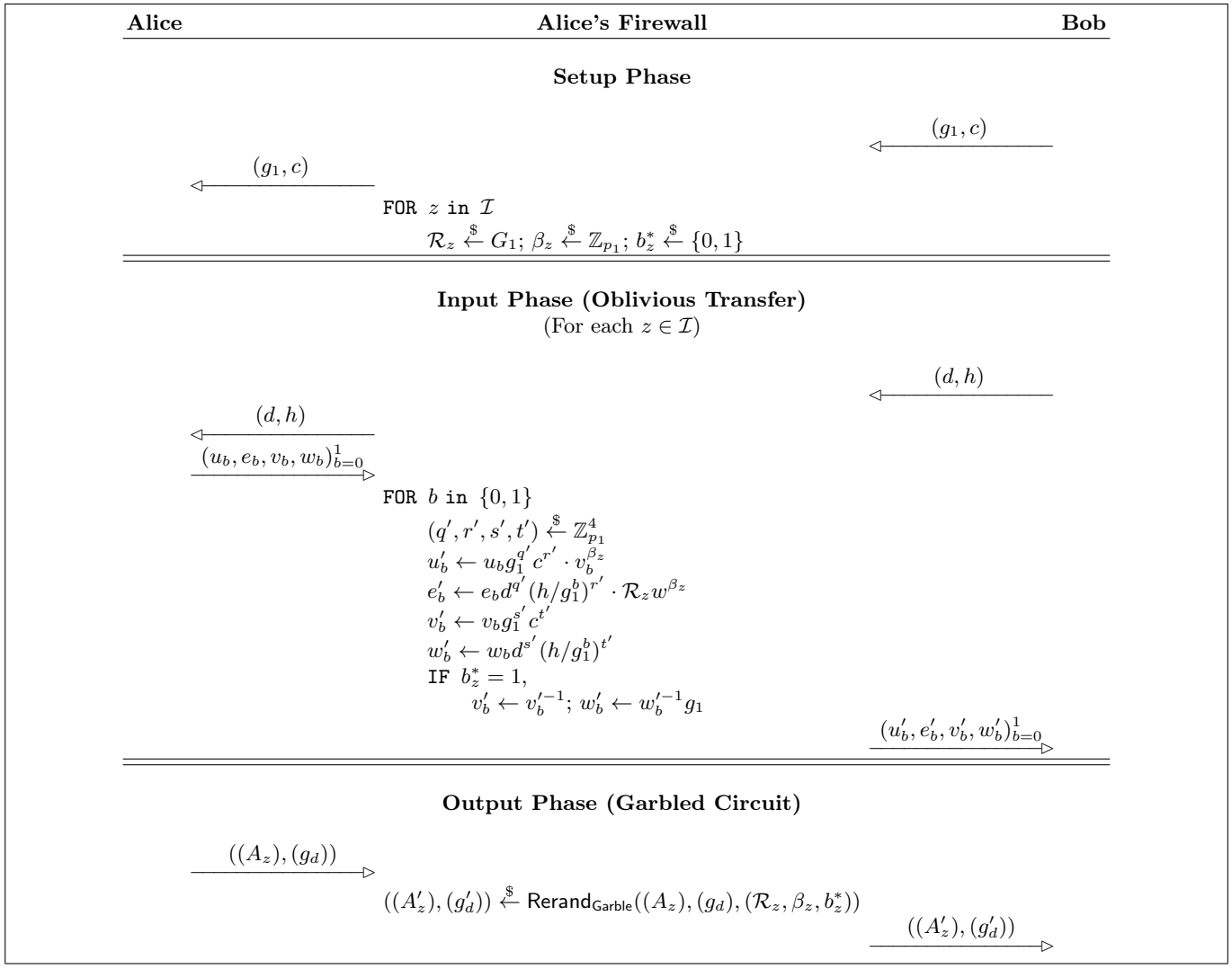

Fig. 8: Alice's firewall for the private function evaluation protocol shown in Figure 7. See Figure 10 in Appendix D for the formal definition of $\mathsf{Rerand}_{\mathsf{Garble}}$.

into messages encrypted under the original key. ElGamal encryption has this property (as we observe in Section 4), so we describe the scheme using ElGamal.

In particular, we interpret all messages as elements in some group G of order p in which DDH is hard. Each party computes $g \xleftarrow{\$} G \setminus \{1_G\}$ and $x \xleftarrow{\$} \mathbb{Z}_p$ and sends the message $(g, h = g^x)$ to all other parties. All future messages m sent to a party are then replaced by ciphertexts encrypted under her public key $(u = g^r, e = h^r m)$. Each time any party receives an encrypted message (u, e), she decrypts it $m = e/u^x$ and then proceeds with the protocol as normal. In addition, to prevent leakage due to early termination of the protocol, the parties never output $\perp$ until the end of the protocol; they instead send encryptions of a special message $m_\perp$ and wait until the end of the protocol to output $\perp$. A party's reverse firewall simply rerandomizes her keys and ciphertexts. If the party ever sends a message that is not of the right form, the firewall simply sends two uniformly random group elements in place of an encryption.

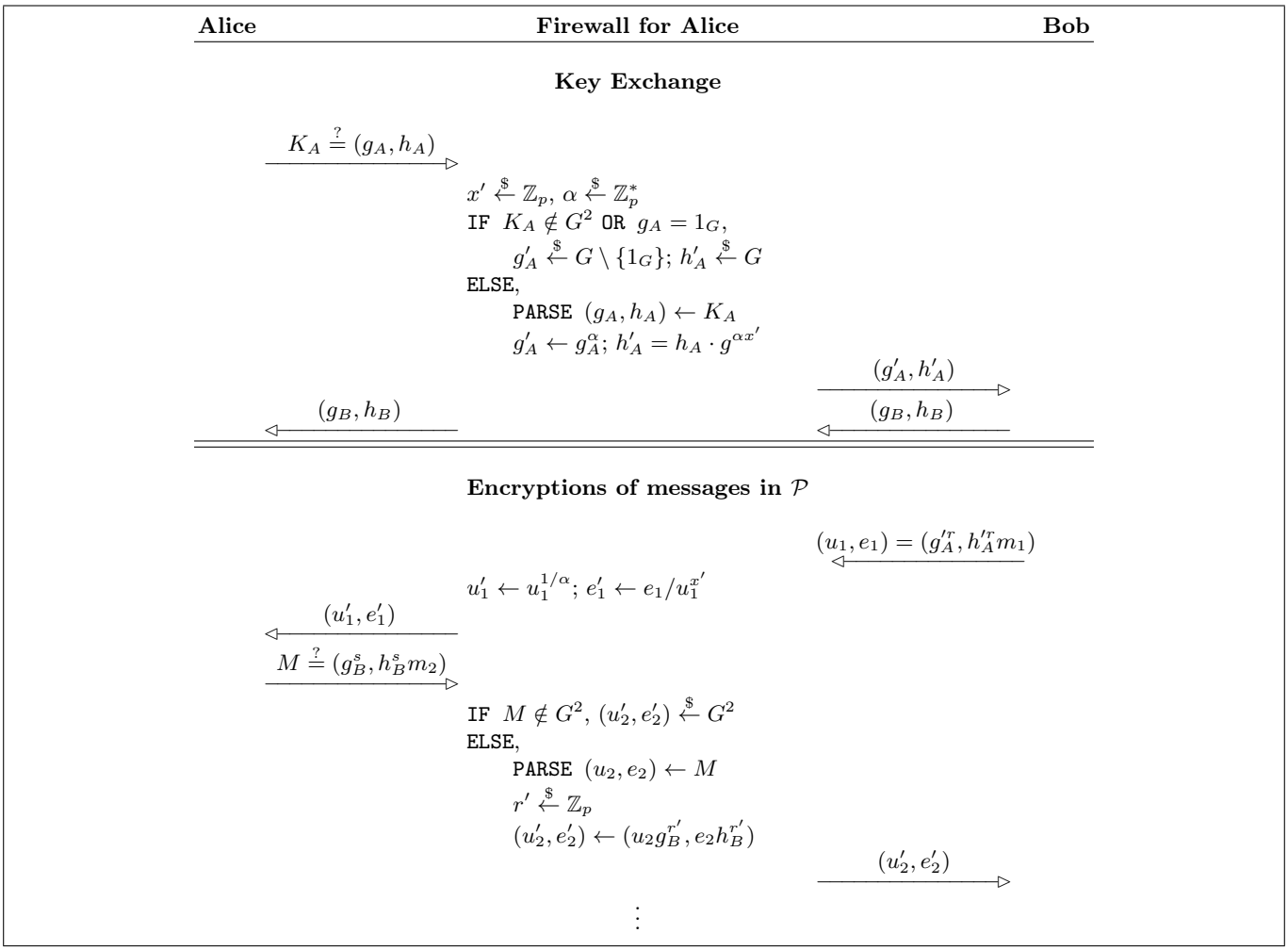

Fig. 9: A reverse firewall for Alice in a modified arbitrary two-party protocol $\mathcal{P}$. Two messages are added to the protocol in which the parties exchange public keys. They then follow the specification of $\mathcal{P}$, replacing messages m_i with ciphertexts $(g^r, h^r m_i)$. Bob has a similar reverse firewall.

We show such a firewall for Alice in the two-party case in Figure 9. Note that Bob can implement essentially the same firewall. The fact that this firewall is strongly exfiltration-resistant against eavesdroppers follows immediately from the assumption that DDH is hard in G.

We note that, in general, the construction in Figure 9 should not be expected to "compose well" with other reverse firewalls. I.e., if some protocol has a reverse firewall that preserves Alice's security but is not exfiltration-resistant, we cannot necessarily apply the above transformation and obtain a protocol with a reverse firewall that *both* preserves Alice's security and is exfiltration-resistant, as it will not be possible for an efficient firewall to compute arbitrary functions on the messages if they are encrypted. Even a very simple operation like equality testing (e.g., testing whether a message is some specific element) cannot be done efficiently if the message is encrypted under a semantically secure scheme. So, in general, one may need to choose between strongly exfiltration-resistant firewalls and firewalls that preserve security.

6 Conclusion and directions for future work

The revelations of Edward Snowden [PLS13, BBG13, Gre14] highlight a different kind of threat posed by sophisticated adversaries—the potential hijacking of a user's own software or hardware for subversive purposes. A compromised machine engaged in a cryptographic protocol may (perhaps selectively) fail to protect security or enable a covert communication channel through which the attacker can leak sensitive information or coordinate its activities. Standard solutions such as testing, auditing, or monitoring cannot in general ensure security since the attacker may use cryptographic methods to cover its tracks (aided by the complexity of modern protocols and the ubiquitous use of randomness in communications).

To counter the threat of insider attacks, we propose the concept of a (cryptographic) reverse firewall, whose role is to backstop the security of some underlying cryptographic scheme. We discuss several desirable properties of reverse firewalls (maintaining functionality, preserving security, and protecting against exfiltration attacks) and two types of tampering (arbitrary tampering and functionality-maintaining tampering). The generality of our definition provides a framework for studying insider attacks and countermeasures across a wide range of primitives.

Our main technical contribution is a protocol for private function evaluation based on Yao's garbled circuits and oblivious transfer that admits a reverse firewall for both parties. The instantiation of this remarkably strong primitive in a way that remains secure even when the user's computer has been compromised shows the power of reverse firewalls as a tool for protecting against insider attacks. In addition, our rerandomizable garbling scheme is more efficient and is secure against a stronger adversary than the scheme proposed by Gentry et al. [GHV10] (though we rely on slightly stronger number-theoretic assumptions).

We also show that *any* protocol can be easily converted into a protocol with an exfiltration-resistant reverse firewall for each party (and the same functionality). This provides a generic way to prevent a tampered machine from leaking information to an eavesdropper via any protocol.

We conclude with a (non-exhaustive!) list of exciting directions for future work in the newly emphasized study of defense against insider attacks:

1. The most obvious direction for future work is simply the instantiation of more primitives in this framework. While this work includes an instantiation of private function evaluation (which can be used to instantiate many more primitives), there is still much more to study. For example, can we achieve stronger notions of security for two party computation? (We prove a relatively restricted notion of security for private function evaluation.) How efficiently (and under what assumptions) can we instantiate simpler primitives in this model? What can we achieve in the multi-party case? What about other primitives that are not implied by PFE (such as authenticated key agreement)?
2. We hope that future work on reverse firewalls develops a comprehensive collection of composable, efficient, modular protocols with secure reverse firewalls. The "holy grail" would be a full characterization of functionalities and security properties for which reverse firewalls exist.
3. More generally, we hope to see a systematic study of defensive mechanisms against deliberate insider attacks. The legitimate targets of these attacks include software libraries, hardware platforms, communication channels, standards, protocols, sources of entropy, system parameters, and the choice of constants.

References

[AIR01] William Aiello, Yuval Ishai, and Omer Reingold. Priced oblivious transfer: How to sell digital goods. In *Proceedings of the International Conference on the Theory and Application of Cryptographic Techniques: Advances in Cryptology*, EUROCRYPT '01, pages 119–135, London, UK, UK, 2001. Springer-Verlag.

[AsV08] Joël Alwen, abhi shelat, and Ivan Visconti. Collusion-free protocols in the mediated model. In David Wagner, editor, *Advances in Cryptology — CRYPTO 2008*, volume 5157 of *Lecture Notes in Computer Science*, pages 497–514. Springer Berlin Heidelberg, 2008.

[BBG13] James Ball, Julian Borger, and Glenn Greenwald. Revealed: how US and UK spy agencies defeat internet privacy and security. *Guardian Weekly*, September 2013.

[BBS98] Matt Blaze, Gerrit Bleumer, and Martin Strauss. Divertible protocols and atomic proxy cryptography. In Kaisa Nyberg, editor, *Advances in Cryptology — EUROCRYPT'98*, volume 1403 of *Lecture Notes in Computer Science*, pages 127–144. Springer Berlin Heidelberg, 1998.

[BD91] Mike Burmester and Yvo Desmedt. All languages in NP have divertible zero-knowledge proofs and arguments under cryptographic assumptions. In Ivan Damgård, editor, *Advances in Cryptology — EUROCRYPT '90*, volume 473 of *Lecture Notes in Computer Science*, pages 1–10. Springer Berlin Heidelberg, 1991.

[BDI+96] Mike Burmester, Yvo Desmedt, Toshiya Itoh, Kouichi Sakurai, Hiroki Shizuya, and Moti Yung. A progress report on subliminal-free channels. In Ross Anderson, editor, *Information Hiding*, volume 1174 of *Lecture Notes in Computer Science*, pages 157–168. Springer Berlin Heidelberg, 1996.

[BDI+99] Mike Burmester, Yvo Desmedt, Toshiya Itoh, Kouichi Sakurai, and Hiroki Shizuya. Divertible and subliminal-free zero-knowledge proofs for languages. *J. Cryptology*, 12:197–223, 1999.

[BHR12] Mihir Bellare, Viet Tung Hoang, and Phillip Rogaway. Foundations of garbled circuits. In *Proceedings of the 2012 ACM Conference on Computer and Communications Security*, CCS '12, pages 784–796, New York, NY, USA, 2012. ACM.

[BPR14a] Mihir Bellare, Kenneth G. Paterson, and Phillip Rogaway. Security of symmetric encryption against mass surveillance. In *CRYPTO*, volume 8616 of *Lecture Notes in Computer Science*, pages 1–19. Springer, 2014. Full version: [BPR14b].

[BPR14b] Mihir Bellare, Kenneth G. Paterson, and Phillip Rogaway. Security of symmetric encryption against mass surveillance. Cryptology ePrint Archive, Report 2014/438, 2014. `http://eprint.iacr.org/`.

[BV07] D. Brown and S. Vanstone. Elliptic curve random number generation, August 16 2007. US Patent App. 11/336,814.

[CVE14a] Vulnerability summary for CVE-2014-1260 ('Heartbleed'). `http://cve.mitre.org/cgi-bin/cvename.cgi?name=CVE-2014-1260`, April 2014.

[CVE14b] Vulnerability summary for CVE-2014-1266 ('goto fail'). `http://cve.mitre.org/cgi-bin/cvename.cgi?name=CVE-2014-1266`, February 2014.

[CVE14c] Vulnerability summary for CVE-2014-6271 ('Shellshock'). `http://cve.mitre.org/cgi-bin/cvename.cgi?name=CVE-2014-6271`, September 2014.

[Des90] Yvo Desmedt. Abuses in cryptography and how to fight them. In Shafi Goldwasser, editor, *Advances in Cryptology — CRYPTO '88*, volume 403 of *Lecture Notes in Computer Science*, pages 375–389. Springer New York, 1990.

[Des94] Y. Desmedt. Subliminal-free sharing schemes. In *Information Theory, 1994. Proceedings., 1994 IEEE International Symposium on*, pages 490–, Jun 1994.

[ElG85] Taher ElGamal. A public key cryptosystem and a signature scheme based on discrete logarithms. In *Proceedings of CRYPTO 84 on Advances in Cryptology*, pages 10–18, New York, NY, USA, 1985. Springer-Verlag New York, Inc.

[FKL10] Kevin Ford, Sergei V. Konyagin, and Florian Luca. Prime chains and Pratt trees. *Geometric and Functional Analysis*, 20(5):1231–1258, 2010.

[GHV10] Craig Gentry, Shai Halevi, and Vinod Vaikuntanathan. *i*-Hop homomorphic encryption and rerandomizable Yao circuits. In Tal Rabin, editor, *Advances in Cryptology — CRYPTO 2010*, volume 6223 of *Lecture Notes in Computer Science*, pages 155–172. Springer Berlin Heidelberg, 2010.

[Gre14] Glenn Greenwald. *No Place to Hide: Edward Snowden, the NSA, and the U.S. Surveillance State*. Metropolitan Books, May 2014.

[HKN+05] Danny Harnik, Joe Kilian, Moni Naor, Omer Reingold, and Alon Rosen. On robust combiners for oblivious transfer and other primitives. In Ronald Cramer, editor, *Advances in Cryptology — EUROCRYPT 2005*, volume 3494 of *Lecture Notes in Computer Science*, pages 96–113. Springer Berlin Heidelberg, 2005.

[LHA+12] Arjen K. Lenstra, James P. Hughes, Maxime Augier, Joppe W. Bos, Thorsten Kleinjung, and Christophe Wachter. Public keys. In Reihaneh Safavi-Naini and Ran Canetti, editors, *Advances in Cryptology — CRYPTO 2012*, volume 7417 of *Lecture Notes in Computer Science*, pages 626–642. Springer Berlin Heidelberg, 2012.

[LMs05] Matt Lepinksi, Silvio Micali, and abhi shelat. Collusion-free protocols. In *Proceedings of the Thirty-seventh Annual ACM Symposium on Theory of Computing*, STOC '05, pages 543–552, New York, NY, USA, 2005. ACM.

[NP01] Moni Naor and Benny Pinkas. Efficient oblivious transfer protocols. In *Proceedings of the Twelfth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '01, pages 448–457, Philadelphia, PA, USA, 2001. Society for Industrial and Applied Mathematics.

[OO90] Tatsuaki Okamoto and Kazuo Ohta. Divertible zero knowledge interactive proofs and commutative random self-reducibility. In Jean-Jacques Quisquater and Joos Vandewalle, editors, *Advances in Cryptology — EUROCRYPT '89*, volume 434 of *Lecture Notes in Computer Science*, pages 134–149. Springer Berlin Heidelberg, 1990.

[PLS13] Nicole Perlroth, Jeff Larson, and Scott Shane. N.S.A. able to foil basic safeguards of privacy on Web. *The New York Times*, September 2013.

[PR07] Manoj Prabhakaran and Mike Rosulek. Rerandomizable RCCA encryption. In Alfred Menezes, editor, *CRYPTO*, volume 4622 of *Lecture Notes in Computer Science*, pages 517–534. Springer, 2007.

[SF07] Dan Shumow and Niels Ferguson. On the possibility of a back door in the NIST SP800-90 Dual Ec Prng. CRYPTO Rump Session, 2007.

[Sim84] Gustavus Simmons. The prisoners' problem and the subliminal channel. In David Chaum, editor, *Advances in Cryptology*, pages 51–67. Springer US, 1984.

[Sim85] Gustavus Simmons. The subliminal channel and digital signatures. In Thomas Beth, Norbert Cot, and Ingemar Ingemarsson, editors, *Advances in Cryptology*, volume 209 of *Lecture Notes in Computer Science*, pages 364–378. Springer Berlin Heidelberg, 1985.

[YY96] Adam L. Young and Moti Yung. The dark side of "black-box" cryptography, or: Should we trust Capstone? In Neal Koblitz, editor, *Advances in Cryptology - CRYPTO '96, 16th Annual International Cryptology Conference, Santa Barbara, California, USA, August 18-22, 1996, Proceedings*, volume 1109 of *Lecture Notes in Computer Science*, pages 89–103. Springer, 1996.

A Definitions of primitives

A.1 Oblivious transfer

We provide a (relatively weak) security definition for oblivious transfer as in [NP01].

Definition 8. *A* one-out-of-two oblivious transfer protocol *is a two-party protocol with a sender, Alice, and a receiver, Bob. Alice's input is a pair of messages (m_0, m_1), and Bob's is a bit b. The protocol is* correct *if Bob's output is m_b.*

The ideal functionality *of an oblivious transfer protocol is a protocol that Alice and Bob play with a trusted third party. Alice simply sends her input messages (m_0, m_1) to the trusted third party, and Bob provides a bit b (not necessarily his input). The third party then sends m_b to Bob.*

The protocol is secure for Alice *if for every (not necessarily efficient) adversary playing the role of Bob $\mathcal{B}$, there exists a (not necessarily efficient) simulator $S_{\mathcal{B}}$ with access to the ideal functionality such that for any input (m_0, m_1, b), the distribution of the output of $S_{\mathcal{B}}$ has negligible statistical distance from the output produced by $\mathcal{B}$ playing the role of Bob in the real protocol.*

The protocol is secure for Bob *if no probabilistic polynomial-time adversary playing the role of Alice can distinguish between the transcript generated when Bob's input is* 0 *and the transcript generated when his input is* 1.

A.2 Private function evaluation

(See [BHR12] for a more formal definition of circuits and circuit layouts.)

Definition 9 (Circuits and circuit layouts). *A* circuit layout *is a directed acyclic graph such that all vertices except for* input vertices *have in-degree* 2 *and all vertices except for* output vertices *have positive out degree. Input vertices have in-degree* 0 *and cannot be output vertices. Here, non-input vertices represent gates (and we often refer to them as such) and edges represent wires. We require that the vertices are numbered topologically (i.e. edges go only from lower-numbered vertices to higher-numbered vertices), and input vertices come first such that the ith bit of input corresponds to vertex i. We write V for the set of all vertices, $\mathcal{I}$ for the set of input vertices, and $\mathcal{O}$ for the set of output vertices. For each gate z, we define $L(z)$ and $R(z)$ as the two vertices preceding z.*

A circuit *$\mathcal{C}$ is a circuit layout together with a collection of functions $f_z : \{0,1\}^2 \to \{0,1\}$, one for each non-input gate z. We view the individual functions as instantiations of the gates and the input (b_L, b_R) as the bits from the left and right wire respectively.*

Given an input x and a circuit $\mathcal{C}$, we define a bit b_z for each vertex z as follows. For each input vertex z, we simply set $b_z \leftarrow x_z$. For each gate z (from the lowest-numbered gate to the highest), we set $b_z \leftarrow f(b_{L(z)}, L_{R(z)})$. We say that gate z takes the value *b_z. Finally, we define $\mathcal{C}(x)$ as the values taken by the output gates in order, $(b_z)_{z \in \mathcal{O}}$.*

Definition 10 (Private function evaluation). *A* private function evaluation protocol *is a two-party protocol between Alice and Bob. A circuit layout $\mathcal{L}$ is a public parameter. Alice's input is a circuit $\mathcal{C}$ with layout $\mathcal{L}$. Bob's input is a bit string x with $|x| = |\mathcal{I}|$.*

The ideal functionality *of a private function evaluation protocol is played with a trusted third party. Alice provides the circuit $\mathcal{C}$ to the third party, and Bob provides his input x. The third party then sends $\mathcal{C}(x)$ to Bob, and Bob outputs the result.*

A protocol is correct *if Bob's output in the real protocol is identical to his output in the ideal protocol.*

The protocol is secure for Alice *if there exists a probabilistic polynomial-time simulator S with access to the ideal functionality and the input x (but not the circuit!) such that for any adversarially chosen input $(\mathcal{C}, x)$, the output of S is computationally indistinguishable from the view of Bob during an execution of the protocol.*

The protocol is secure for Bob *if no probabilistic polynomial-time adversary playing the role of Alice can provide two inputs (x_0, x_1) and then distinguish between the view of Alice when Bob's input is x_0 and the view of Alice when his input is x_1.*

B Groups and hardness assumptions

Definition 11 (Family of groups). *We say that $\mathbb{G} = (G_i)_{i=1}^{\infty}$ is an* efficiently computable family of groups *if there is some probabilistic polynomial-time algorithm* setup *such that* $\mathsf{setup}(1^\lambda)$ *outputs a representation of a group G_i with all group elements represented by* $\mathrm{poly}(\lambda)$ *bits, a polynomial-size circuit that outputs a uniformly random group element on random input, the order of the group, and a polynomial-size circuit that computes the group operation over G_i.*

Throughout this paper, whenever we refer to a group G with certain properties, we implicitly define a family of groups $\mathbb{G}$ with these properties and assume that $G \xleftarrow{\$} \mathsf{setup}(1^\lambda)$, where λ is the security parameter. We assume that all algorithms have access to the group description. We write 1_G to denote the identity element in G. When we speak of negligible probabilities, polynomial-time algorithms, etc., we mean probabilities that are negligible in the security parameter λ, algorithms whose running time is polynomial in the security parameter, etc. We sometimes need to work with more than one group at a time, so we extend these notions in the natural way to a collection of groups.

Definition 12 (Decisional Diffie-Hellman). *Let G be a group of order p. Then, we say that* decisional Diffie-Hellman (DDH) is hard in *G if no probabilistic polynomial-time algorithm $\mathcal{A}$ can distinguish between (g, g^x, g^y, g^{xy}), where $g \xleftarrow{\$} G$, $(x, y) \xleftarrow{\$} \mathbb{Z}_p^2$, and $(g_1, g_2, g_3, g_4) \xleftarrow{\$} G^4$.*

We will need a slight variant of the DDH assumption, which we call *DDH with a hint*.

Definition 13 (DDH with a hint). *We say that* DDH with a hint is hard in *G if no probabilistic polynomial-time adversary $\mathcal{A}$ has non-negligible advantage in the following game.*

1. *$(\sigma, g, c, d) \xleftarrow{\$} \mathcal{A}(1^\lambda)$, with $(g, c, d) \in G^3$.*
2. *Sample $b \xleftarrow{\$} \{0, 1\}$ and $(x, y) \xleftarrow{\$} \mathbb{Z}_p^2$.*
3. *If $b = 1$, set $z \leftarrow xy$. Otherwise, set $z \xleftarrow{\$} \mathbb{Z}_p$*
4. *$b^* \xleftarrow{\$} \mathcal{A}(\sigma, (g^x, g^y, g^z, c^x, d^y))$.*
5. *$\mathcal{A}$ wins if and only if $b = b^*$.*

It will also be convenient to define two hardness assumptions that are implied by DDH.

Definition 14 (Subgroup DDH). *Let G be a group of order p and $\hat{G}$ be a subgroup of $\mathbb{Z}_p^*$. We say that $\hat{G}$-*subgroup DDH *is hard in G if no probabilistic polynomial-time algorithm $\mathcal{A}$ can distinguish between (g, g^x, g^y, g^{xy}), where $g \xleftarrow{\$} G$, $(x, y) \xleftarrow{\$} \hat{G}^2$, and $(g_1, g_2, g_3, g_4) \xleftarrow{\$} G^4$.*

Lemma 1. *Let G be a group of order p and $\hat{G}$ be a subgroup of $\mathbb{Z}_p^*$. If DDH is hard in G and $|G|/|\hat{G}|$ is polynomially bounded, then $\hat{G}$-subgroup DDH is hard in G.*

Proof. Fix some probabilistic polynomial-time adversary $\mathcal{A}$ in the DDH game, and let $g \xleftarrow{\$} G$. Consider the cosets of $\hat{G}$ in $\mathbb{Z}_p^*$. In particular, we can associate to each pair of cosets (C_i, C_j) an advantage

$$P_{C_i,C_j} := \Pr_{(x,y)\xleftarrow{\$} C_i \times C_j}[\mathcal{A}(g, g^x, g^y, g^{xy}) = 1] - \Pr_{(x,y,z)\xleftarrow{\$} C_i \times C_j \times \mathbb{Z}_p^*}[\mathcal{A}(g, g^x, g^y, g^z) = 1] .$$

Suppose that $\mathcal{A}$ has non-negligible advantage in the $\hat{G}$-subgroup DDH game. Then, there is some constant k such that $P_{\hat{G},\hat{G}} > \lambda^{-k}$. Let $I = |G|/|\hat{G}|$ be the index of $\hat{G}$ over G, and recall that I is polynomially bounded. By the pigeonhole principle, there must be some interval between 0 and λ^{-k} of length $1/(2\lambda^k I^2)$ such that none of the values P_{C_i,C_j} is in this interval. So, by the Chernoff bound, for each (C_i, C_j), we can run $\mathcal{A}$, say, $100\lambda^{3k} I^5$ times to classify P_{C_i,C_j} as either greater than the midpoint of this interval or less than it, failing with only negligible probability. Indeed, given (g, g^x) for unknown x in coset $C(x)$ and any coset C_j, we can classify $P_{C(x),C_j}$ by running $\mathcal{A}$ a total of $100\lambda^{3k} I^5$ times on input of the form $(g, g^{\alpha x}, g^y, g^{\alpha xy})$ and input of the form $(g, g^{\alpha x}, g^y, g^z)$ for $z \xleftarrow{\$} \mathbb{Z}_p$, $\alpha \xleftarrow{\$} \hat{G}$, and $y \xleftarrow{\$} C_j$ and comparing the results. Similarly, given (g, g^y), we can classify $P_{C_i,C(y)}$.

Finally, we claim that given (g, g^x, g^y), we can classify $P_{C(x),C(y)}$. We do this by first classifying all of the P_{C_i,C_j}. We then divide the C_i into left equivalence classes such that for two elements C_i and C_k in the same equivalence class, P_{C_i,C_j} has the same classification as P_{C_k,C_j} for all C_j. We similarly divide the C_j into right equivalence classes. Finally, using the idea outlined above, we can identify the left equivalence class of $C(x)$ and the right equivalence class of $C(y)$. We can then categorize $P_{C(x),C(y)}$ by finding the unique category that "matches" these equivalence classes.

So, an adversary $\mathcal{A}'$ in the DDH game can, on input (g, g^x, g^y, g^z), first categorize $P_{C(x),C(y)}$. If $P_{C(x),C(y)}$ is greater than the midpoint of the interval, it outputs $\mathcal{A}(g, g^x, g^y, g^z)$. Otherwise, it flips a coin and outputs the result. Since $P_{\hat{G},\hat{G}} > \lambda^{-k}$, it follows that with probability at least $|\hat{G}|/|G| = 1/\text{poly}(\lambda)$, we have that $P_{C(x),C(y)}$ is larger than the midpoint. The result follows. □

Definition 15 (k-DDH and subgroup k-DDH). *Let G be a group of order p. For $k \geq 2$, we say that k-*DDH *is hard in G if no probabilistic polynomial-time algorithm $\mathcal{A}$ can distinguish between $(g, (g^{x_i})_{i=1}^k, (g^{x_i x_j})_{1\leq i<j\leq k})$ and $(g_i^*)_{i=1}^\ell$ with $\ell = k(k+1)/2 + 1$, where $g \xleftarrow{\$} G$, $(x_i) \xleftarrow{\$} \mathbb{Z}_p^k$, and $(g_i^*) \xleftarrow{\$} G^\ell$.*

*Let G be a group of order p, $\hat{G}$ be a subgroup of $\mathbb{Z}_p^*$, and $k \geq 2$. We say that $\hat{G}$-*subgroup k-DDH *is hard in G if no probabilistic polynomial-time algorithm $\mathcal{A}$ can distinguish between $(g, (g^{x_i})_{i=1}^k, (g^{x_i x_j})_{1\leq i<j\leq k})$ and $(g_i^*)_{i=1}^\ell$, where $g \xleftarrow{\$} G$, $(x_i) \xleftarrow{\$} \hat{G}^k$, and $(g_i^*) \xleftarrow{\$} G^\ell$.*

Lemma 2. *Let G be a group of order p. If DDH is hard in G, then k-DDH is hard in G for any polynomially bounded k.*

Let $\hat{G}$ be a subgroup of $\mathbb{Z}_p^*$. If DDH is hard in G and $|G|/|\hat{G}|$ is polynomially bounded, then $\hat{G}$-subgroup k-DDH is hard in G for any polynomially bounded k.

Proof. For $i = 0, \ldots, (k^2 - k)/2$, let **Game** i be the game of distinguishing between a uniformly random tuple $(g_1^*, \ldots, g_\ell^*) \xleftarrow{\$} G^\ell$ and a tuple of the form $(g, (g^{x_i})_{i=1}^k, (g^{x_i x_j})_{1\leq i<j\leq k})$ with the last i elements changed to a uniformly random element. It follows from the assumption that DDH is hard in G that no

adversary can have non-negligibly larger advantage in **Game** i than in **Game** $i+1$. The result follows by noting that **Game 0** is the k-DDH game and that no adversary can have any advantage in **Game** $(k^2-k)/2$. □

We will also require a non-uniform version of DDH.

Definition 16 (Non-uniform decisional Diffie-Hellman). *Let G be a group of order p. We say that* non-uniform decisional Diffie-Hellman is hard in G *if no probabilistic polynomial-time algorithm $\mathcal{A}$ with* auxiliary information $\mathsf{aux} = \mathsf{aux}(G)$ *can distinguish between (g, g^x, g^y, g^{xy}), where $g \xleftarrow{\$} G$, $(x,y) \xleftarrow{\$} \mathbb{Z}_p^2$, and $(g_1, g_2, g_3, g_4) \xleftarrow{\$} G^4$. Note that* aux *does not need to be efficiently computable.*

We similarly extend the definitions of $\hat{G}$-subgroup DDH, k-DDH, and $\hat{G}$-subgroup k-DDH to the non-uniform setting. Finally, one more definition will be useful.

Definition 17 (Chosen-bases $\hat{G}$-subgroup k-DDH). *Let G be a group of order p, $\hat{G}$ be a subgroup of $\mathbb{Z}_p^*$, and $k \geq 2$. We say that* chosen-bases $\hat{G}$-subgroup k-DDH is hard in G *if no probabilistic polynomial-time algorithm $\mathcal{A}$ has non-negligible advantage in the following game.*

1. $(\sigma, (g_i)_{i=1}^k, (h_{i,j})_{1 \leq i < j \leq k}) \xleftarrow{\$} \mathcal{A}(1^\lambda)$, *with* $g_i, h_{i,j} \in G \setminus \{1_G\}$.
2. *Sample* $b \xleftarrow{\$} \{0,1\}$, $(x_i)_{i=1}^k \xleftarrow{\$} \mathbb{Z}_p^k$, *and* $(g_i^*)_{i=1}^\ell \xleftarrow{\$} G^\ell$, *where* $\ell = (k^2+k)/2$.
3. *If* $b = 0$, $b^* \xleftarrow{\$} \mathcal{A}(\sigma, (g_i^{x_i})_{i=1}^k, (h_{i,j}^{x_i x_j})_{i<j})$. *Otherwise,* $b^* \xleftarrow{\$} \mathcal{A}(\sigma, (g_i^*))$.
4. $\mathcal{A}$ *wins if and only if* $b = b^*$.

Lemma 3. *Let G be a group of order p, and let $\hat{G}$ be a subgroup of $\mathbb{Z}_p^*$. If non-uniform DDH is hard in G and $|G|/|\hat{G}|$ is polynomially bounded, then chosen-bases $\hat{G}$-subgroup k-DDH is hard in G for any polynomially bounded k.*

Proof. We first note that the natural non-uniform analogue of Lemma 2 holds by an essentially identical proof. In particular, it suffices to show that chosen-bases $\hat{G}$-subgroup k-DDH is hard in G if non-uniform $\hat{G}$-subgroup k-DDH is hard in G.

Let $\mathcal{A}$ be an adversary in the chosen-bases $\hat{G}$-subgroup k-DDH game in G. Note that $\mathcal{A}$ may not be deterministic, but we can fix the output of $\mathcal{A}$, $(\sigma, (g_i)_{i=1}^k, (h_{i,j})_{1 \leq i < j \leq k}) \xleftarrow{\$} \mathcal{A}(1^\lambda)$, such that the advantage of $\mathcal{A}$ with this fixed output is maximal. Let $\mathsf{aux} = (\sigma, (\log_{g_1}(g_i))_{i=1}^k, (\log_{g_1}(h_{i,j}))_{1 \leq i < j \leq k})$.

We then build $\mathcal{A}'$, an adversary in the non-uniform $\hat{G}$-subgroup k-DDH in G as follows. $\mathcal{A}'$ receives auxiliary input $(\sigma, (\log_{g_1}(g_i))_{i=1}^k, (\log_{g_1}(h_{i,j}))_{1 \leq i < j \leq k})$ and challenge $((g_i^*)_{i=1}^k, (h_{i,j}^*))$. For each i, j, it sets $g_i' \leftarrow {g_i^*}^{\log_{g_1} g_i}$ and $h_{i,j}' \leftarrow {h_{i,j}^*}^{\log_{g_1}(h_{i,j})}$. It then returns $\mathcal{A}(\sigma, (g_i'), (h_{i,j}'))$.

It should be clear that the view of $\mathcal{A}$ is identical to its view in the $\hat{G}$-subgroup k-DDH game in G. □

C Proof of Theorem 2 (security of the OT firewalls)

Proof of Theorem 2. It should be clear that both firewalls maintain functionality.

Let $\mathcal{A}$ be some tampered implementation of Alice, (g, c, d, h) Bob's initial (possibly adversarial) message to Alice, (u_i, e_i) the response of $\mathcal{A}$, and (u_i', e_i') the same response after it passes through Alice's firewall $\mathcal{W}_\mathsf{A}$. Suppose Bob's message satisfies $(g, c, d, h) = (g, g^x, g^y, g^{xy+b})$. Then, it should be clear that, if (u_b, e_b) is a valid response, then (u_b', e_b') is a uniformly random valid response. Note that Bob's firewall $\mathcal{W}_\mathsf{B}$ replaces (g, g^x, g^y, g^{xy+b}) with a uniformly random message of the same form, $(g^\alpha, g^{\alpha(x+x')}, g^{\alpha(y+y')}, g^{\alpha((x+x')(y+y')+b)})$. It follows that if Alice maintains correctness, the response of $\mathcal{W}_\mathsf{B} \circ (\mathcal{W}_\mathsf{A} \circ \mathcal{A})$ to *any* valid message (g, g^x, g^y, g^{xy+b}) from Bob is a uniformly random valid response (u_b', e_b') except with negligible probability. And, an argument similar to the proof of Proposition 1 shows that if Bob's if $(g, c, d, h) \neq (g, g^x, g^y, g^{xy+i})$,

then (u'_i, e'_i) are uniformly random group elements. This immediately shows that the composed RF maintains correctness and weakly preserves Alice's security. To see that $\mathcal{W}_\mathsf{A}$ (as well as the composed firewall $\mathcal{W}_\mathsf{B} \circ \mathcal{W}_\mathsf{A}$) is strongly exfiltration-resistant against an eavesdropper, we note additionally that $(u'_b, e'_b) = (u_b g^r, e_b g^{ry})$ for $r \xleftarrow{\$} \mathbb{Z}_p$. The result then follows from the assumption that DDH is hard in G.

Turning to Bob's RF, in order to prove robustness, it suffices to show that no PPT adversary can distinguish between $\mathcal{W}_\mathsf{B}(\zeta_0)$ and $\mathcal{W}_\mathsf{B}(\zeta_1)$ with non-negligible advantage for adversarially chosen $\zeta_i = (g_i, c_i, d_i, h_i)$ with $g_i \neq 1_G$.

- **Game 1** is the game that asks an adversary to provide ζ_0 and ζ_1 and distinguish between $\mathcal{W}_\mathsf{B}(\zeta_0)$ and $\mathcal{W}_\mathsf{B}(\zeta_1)$. In particular, the "challenge message" takes the form

$$(g^*, c^*, d^*, h^*) = (g_b^\alpha, c_b^\alpha g_b^{\alpha x'}, d_b^\alpha g_b^{\alpha y'}, h_b^\alpha c_b^{\alpha y'} d_b^{\alpha x'} g_b^{\alpha x' y'}) \,,$$

  where $g \xleftarrow{\$} 1_G$, $(g_b, c_b, d_b, h_b) = \zeta_b$ are adversarially chosen with $g \neq 1_G$, and $(\alpha, x', y') \xleftarrow{\$} \mathbb{Z}_p^3$.
- **Game 2** is **Game 1** with the challenge message replaced by

$$(g^*, c^*, d^*, h^*) = (g_b^\alpha, c_b^\alpha g_b^{\alpha x'}, d_b^\alpha g_b^{\alpha y'}, h_b^\alpha c_b^{\alpha y'} d_b^{\alpha x'} g_b^{\alpha z'}) \,,$$

  where $g \xleftarrow{\$} 1_G$, $(g_b, c_b, d_b, h_b) = \zeta_b$ are adversarially chosen with $g \neq 1_G$, and $(\alpha, x', y', z') \xleftarrow{\$} \mathbb{Z}_p^4$.
- **Game 3** is **Game 2** with the challenge message replaced by four uniformly random group elements.

Claim 2.1. *For any PPT adversary $\mathcal{A}$, $|\mathsf{Adv}^{\textbf{Game 1}}(\mathcal{A}) - \mathsf{Adv}^{\textbf{Game 2}}(\mathcal{A})|$ is negligible if DDH with a hint is hard in G.*

Proof. Fix $\mathcal{A}$. We construct an adversary $\mathcal{A}'$ in the DDH with a hint game as follows.

1. $(\zeta_0, \zeta_1, \sigma) \xleftarrow{\$} \mathcal{A}()$.
2. $b \xleftarrow{\$} \{0,1\}$. Parse $(g_b, c_b, d_b, h_b) \leftarrow \zeta_b$.
3. Sample $\alpha \xleftarrow{\$} \mathbb{Z}_p$.
4. Send $(g_b^\alpha, d_b^\alpha, c_b^\alpha)$ to the DDH with a hint challenger, and receive response $(g_b^{\alpha x}, g_b^{\alpha y}, g_b^{\alpha z}, d_b^{\alpha x}, c_b^{\alpha y})$.
5. $b^* \xleftarrow{\$} \mathcal{A}(\sigma, (g_b^\alpha, c_b^\alpha g_b^{\alpha x}, d_b^\alpha g_b^{\alpha y}, h_b^\alpha c_b^{\alpha y} d_b^{\alpha x} g_b^{\alpha z}))$.
6. Output 1 if and only if $b = b^*$

Clearly, if $z = xy$, then the view of $\mathcal{A}$ in the above game is distributed identically to its view in **Game 1**, and if $z \xleftarrow{\$} \mathbb{Z}_p$, then the view of $\mathcal{A}$ in the above game is identically distributed to its view in **Game 2**. The claim follows from the hardness of DDH with a hint in G. (2.1) ■

The result follows by noting that the difference between **Game 2** and **Game 3** is merely syntactic and that no adversary can have any advantage in **Game 3**.

□

D Details of the firewalls from Section 5

In Figure 10, we present an algorithm $\mathsf{Rerand}_{\mathsf{Garble}}$ that rerandomizes the garbled circuits from Section 4. The rerandomization of the ciphertexts is accomplished by five subprocedures that work in concert. The procedure $\mathsf{Permute}$ reorders the ciphertexts corresponding to a given vertex according to the random bits b^*_z, and the procedure Flip_τ flips the location bits $\tau_z^{(b)}$ in the manner in Section 4 so that they remain consistent with this new ordering. Similarly, Change_T rerandomizes tags $T_z^{(b)}$ to new tags $T_z'^{(b)}$, and the procedure Change_{key} converts ciphertexts under the old secret key $T_L \cdot T_R$ to ciphertexts under the new key $T'_L \cdot T'_R$. Finally, $\mathsf{Rerand}_{\mathsf{Enc}}$ simply rerandomizes all of the ElGamal ciphertexts in the standard way, mapping $(g^r, h^r T)$ to $(g^{r+r'}, h^{r+r'} T)$. Intuitively, these functions together "reset all of the randomness used to garble the circuit."

In Figure 11, we present Bob's reverse firewall for the private function evaluation protocol from Section 4 for completeness. We note that it is nearly identical to Bob's firewall from Section 3.

proc. $\mathsf{Rerand}_{\mathsf{Garble}}((A_z),(g_d),(\mathcal{R}_z,\beta_z,b^*_z))$
$\alpha_1 \leftarrow 1$; $g'_1 \leftarrow g_1$
`FOR` $d = 2,\ldots,D$,
  $\alpha_d \xleftarrow{\$} \mathbb{Z}^*_{p_d}$; $g'_d \leftarrow g_d^{\alpha_d}$
`FOR` z `in` $V \setminus \mathcal{I}$,
  $(h'_{\boldsymbol{\eta}},u'_{\boldsymbol{\eta}},e'_{\boldsymbol{\eta}},v'_{\boldsymbol{\eta}},w'_{\boldsymbol{\eta}})_{\boldsymbol{\eta}\in\{0,1\}^2} \leftarrow \mathsf{Permute}(A_z)$
  $d \leftarrow \mathsf{depth}(z)$
  $\mathcal{R}_z \xleftarrow{\$} G_d$; $\beta_z \xleftarrow{\$} \mathbb{Z}_{p_d}$
  $b^*_z \xleftarrow{\$} \{0,1\}$
  `FOR` $\boldsymbol{\eta}$ `in` $\{0,1\}^2$
    `IF` $b^*_z = 1$, $(v'_{\boldsymbol{\eta}},w'_{\boldsymbol{\eta}}) \leftarrow \mathsf{Flip}_\tau(v'_{\boldsymbol{\eta}},w'_{\boldsymbol{\eta}},d)$
    $(u'_{\boldsymbol{\eta}},e'_{\boldsymbol{\eta}},v'_{\boldsymbol{\eta}},w'_{\boldsymbol{\eta}}) \leftarrow \mathsf{Change}_T(u'_{\boldsymbol{\eta}},e'_{\boldsymbol{\eta}},v'_{\boldsymbol{\eta}},w'_{\boldsymbol{\eta}},z)$
    $(h'_{\boldsymbol{\eta}},u'_{\boldsymbol{\eta}},v'_{\boldsymbol{\eta}}) \leftarrow \mathsf{Change}_{\mathsf{Key}}(h'_{\boldsymbol{\eta}},u'_{\boldsymbol{\eta}},v'_{\boldsymbol{\eta}},z,\eta_0,\eta_1)$
    $(u'_{\boldsymbol{\eta}},e'_{\boldsymbol{\eta}},v'_{\boldsymbol{\eta}},w'_{\boldsymbol{\eta}}) \xleftarrow{\$} \mathsf{Rerand}_{\mathsf{Enc}}(h'_{\boldsymbol{\eta}},u'_{\boldsymbol{\eta}},e'_{\boldsymbol{\eta}},v'_{\boldsymbol{\eta}},w'_{\boldsymbol{\eta}},d)$
  $A'_z \leftarrow (h'_{\boldsymbol{\eta}},u'_{\boldsymbol{\eta}},e'_{\boldsymbol{\eta}},v'_{\boldsymbol{\eta}},w'_{\boldsymbol{\eta}})_{\boldsymbol{\eta}\in\{0,1\}^2}$
`OUTPUT` $((A'_z),(g'_d))$

proc. $\mathsf{Rerand}_{\mathsf{Enc}}(h,u,e,v,w,d)$
$(r',s') \xleftarrow{\$} \mathbb{Z}_{p_d}$
`OUTPUT` $(ug_d'^{r'}, eh^{r'}, vg_d'^{s'}, wh^{s'})$

proc. $\mathsf{Permute}(z)$
`PARSE` $(C_i)_{\boldsymbol{\eta}\in\{0,1\}^2} \leftarrow A_z$
`FOR` $(\tau_L,\tau_R) \in \{0,1\}^2$
  $C'_{\tau_L\oplus b^*_{L(z)},\tau_R\oplus b^*_{R(z)}} \leftarrow C_{\tau_L,\tau_R}$
`OUTPUT` $(C'_{\boldsymbol{\eta}})_{\boldsymbol{\eta}\in\{0,1\}^2}$

proc. $\mathsf{Flip}_\tau(v,w,d)$
`OUTPUT` $(v^{-1}, w^{-1}g_d)$

proc. $\mathsf{Change}_T(u,e,v,w,z)$
$d \leftarrow \mathsf{depth}(z)$
`IF` $z \in \mathcal{O}$, `OUTPUT` $(u^{\alpha_d}, e^{\alpha_d}, v^{\alpha_d}, w^{\alpha_d})$
`OUTPUT` $(u \cdot v'^{\beta_z}, e \cdot \mathcal{R}_z \cdot w'^{\beta_z}, v^{\alpha_d}, w^{\alpha_d})$

proc. $\mathsf{Change}_{\mathsf{Key}}(h,u,v,z,\tau'_L,\tau'_R)$
$d \leftarrow \mathsf{depth}(z)$
$\mathcal{R} \leftarrow \mathcal{R}_{L(z)}\mathcal{R}_{R(z)}g_{d-1}'^{\tau'_L\beta_{L(z)}+\tau'_R\beta_{R(z)}}$
`IF` $h = 1_{G_d}$, $h' \xleftarrow{\$} G_d \setminus \{1_{G_d}\}$
`ELSE`, $h' \leftarrow h^{\alpha_d\mathcal{R}}$
$u' \leftarrow u^{1/\mathcal{R}}$; $v' \leftarrow v^{1/\mathcal{R}}$
`OUTPUT` (h',u',v')

Fig. 10: The algorithm $\mathsf{Rerand}_{\mathsf{Garble}}$ rerandomizes garbled circuits in the form of the output of the algorithm Garble shown in Figure 5. In addition to a garbled circuit $((A_z),(g_d))$, the algorithm also takes as input masks $(\mathcal{R}_z,\beta_z,b^*_z)$ for the input vertices. It outputs rerandomized ciphertexts (A'_z) and rerandomized group elements (g'_d). The location bits $\tau_z^{(b)}$ corresponding to input vertices are replaced by $\tau_z'^{(b)} = \tau_z^{(b)} \oplus b^*_z$. Tags $T_z^{(b)}$ corresponding to input vertices in the original circuit are replaced by tags $T_z^{(b)} \cdot \mathcal{R}_z \cdot g_d'^{\beta_z\cdot\tau_z'^{(b)}}$.

E Proof of Proposition 3 (security of the PFE protocol)

Proof of Proposition 3. Correctness follows immediately. Security for Bob follows from the security of the oblivious transfer protocol (see Section 3).

To prove security for Alice, we need to show that there exists a probabilistic polynomial-time simulator S such that for any adversarially chosen input $(\mathcal{C},x)$, the output of S on input x with access to the ideal functionality $\mathcal{F}$ is computationally indistinguishable from the view of Bob in the real protocol with input $(\mathcal{C},x)$. It suffices for S to query $y \leftarrow \mathcal{F}(x)$ and construct a constant circuit $\mathcal{C}_y$ whose output is y on all inputs. It can then simulate a run of the protocol with Bob on inputs $\mathcal{C}_y$ and x and output the view of Bob.

It will be convenient to define a *computation path* on a garbled circuit (A_z) with input tags and location bits (T_z,τ_z) as the collection of all ciphertexts (h,u,e,v,w) that are decrypted during a run of the Eval function. With that, we define the following sequence of games.

- **Game 1** is the indistinguishability game between the output of S on input x (with access to the ideal functionality computing $\mathcal{C}$) and the view of Bob in the real game on input $(\mathcal{C},x)$.
- **Game 2** is **Game 1** in which the group elements in Alice's input messages corresponding to tags and location bits of the form $(T_z^{(1-x_z)},\tau_z^{(1-x_z)})$ for input vertices z are replaced by uniformly random group elements (in both the simulated world and the real world).
- For $d = 2$ to D, **Game** $d+1$ is **Game** d in which all ciphertexts (h,u,e,v,w) at depth d that are not in the computation path are replaced by uniformly random group elements.

The difference between **Game 1** and **Game 2** is merely syntactic. (See the proof of Proposition 1.) In the final game, **Game** $D+1$, the views are identical. So, it suffices to show that no adversary's advantage

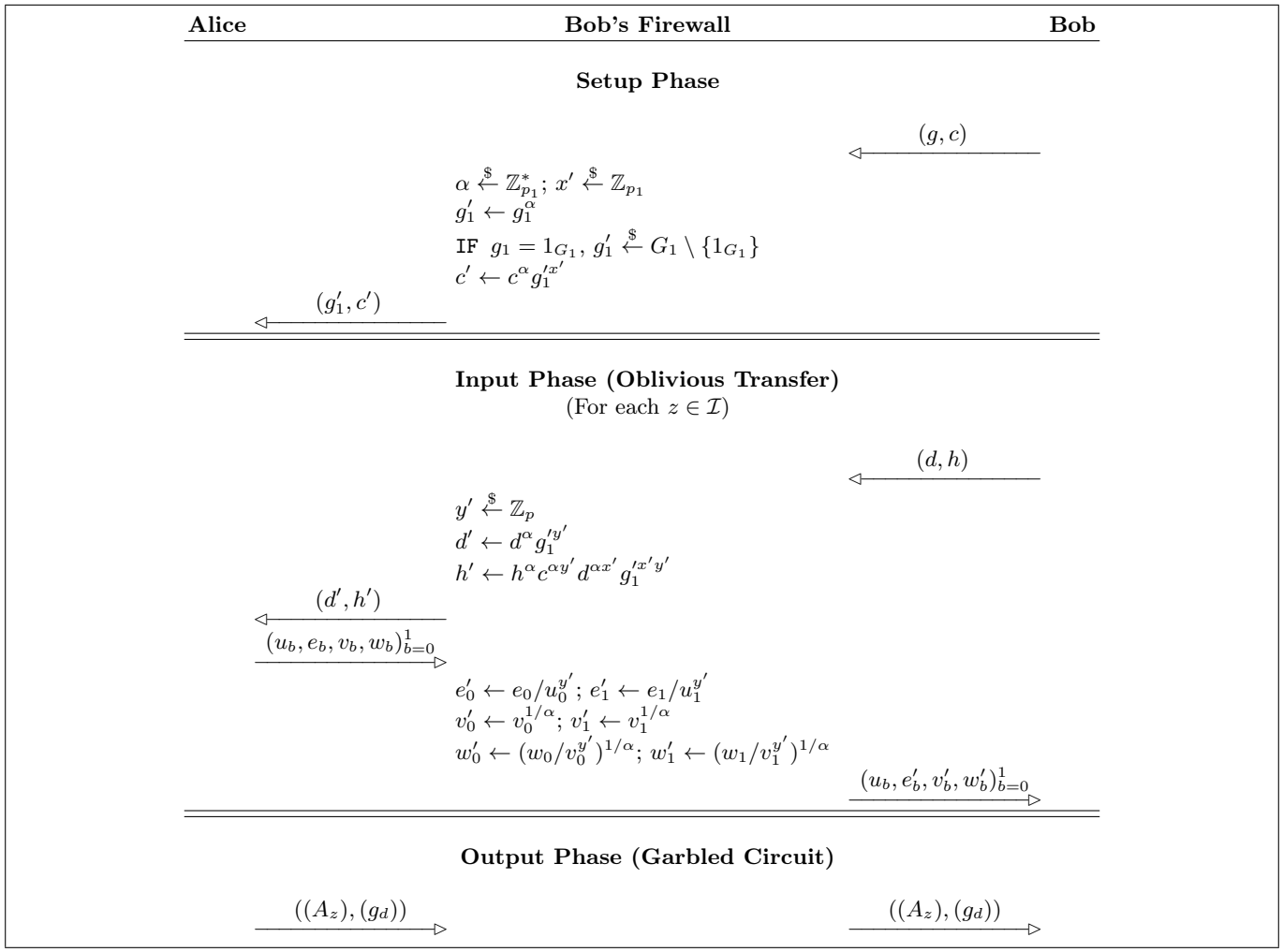

Fig. 11: Bob's firewall for the private function evaluation protocol shown in Figure 7.

in **Game** d is non-negligibly greater than its advantage in **Game** $d+1$. To that end, it will be convenient to define an intermediate game, **Game** $d+1/2$. Note that in **Game** d, the ciphertexts at depth d take the form $(h, g_d^r, h^r T, g_d^s, h^s g_d^\tau)$ where $h = g_d^{T_L T_R}$ for some tags $T_L, T_R \in G_{d-1}$. So, we define **Game** $d+1/2$ as **Game** d in which each public key h that is not in the computation path is independently replaced by $h' \xleftarrow{\$} G_d$, so that the full ciphertexts now take the form $(h', g_d^r, h'^r T, g_d^s, h'^s g_d^\tau)$.

Claim 3.1. *For any probabilistic polynomial-time adversary $\mathcal{B}$, $|\mathsf{Adv}^{\textbf{Game } d+1/2}(\mathcal{B}) - \mathsf{Adv}^{\textbf{Game } d}(\mathcal{B})|$ is negligible if DDH is hard in G_d.*

Proof. By Lemma 2, it suffices to show that the difference in advantages is small if G_{d-1}-subgroup k-DDH is hard in G_d. Indeed, note that for a vertex z of depth d, the public keys satisfy

$$h_{b_L, b_R} = g_d^{T_{L(z)}^{(b_L)} T_{R(z)}^{(b_R)}}$$

where the $T_{z'}^{(b)}$ are in G_{d-1}. The adversary can efficiently compute half of the tags $T_{z'}^{(b)}$—those corresponding to the computation path.

Let $\mathcal{B}$ be an adversary in **Game** d. Let k be the number of vertices of depth $d-1$. Then, we construct an adversary $\mathcal{B}'$ in the G_{d-1}-subgroup k-DDH game in G_d. $\mathcal{B}'$ receives input $(\gamma_i)_{i=1}^{k(k+1)/2+1} = (\gamma_1, (\gamma_1^{x_i})_{i=1}^k, (\gamma_1^{y_{i,j}})_{i<j})$ where the γ_i are elements of G, x_i are uniformly random elements from G_{d-1}, and either $y_{i,j} = x_i x_j$ or they are uniformly random elements from $\mathbb{Z}_{p_d}$. $\mathcal{B}'$ then behaves as follows.

1. $(\mathcal{C}, x) \xleftarrow{\$} \mathcal{B}()$. $y \leftarrow \mathcal{C}(x)$.
2. Sample $b \xleftarrow{\$} \{0,1\}$. If $b = 0$, let $\mathcal{C}^* = \mathcal{C}$. Otherwise, let $\mathcal{C}^* = \mathcal{C}_y$, the constant circuit that always outputs y.
3. Let $(z_i)_{i=1}^k$ be the vertices of depth $d-1$ in some ordering. For a run of the protocol, let T_i represent the tag corresponding to z_i that is in the computation path.
4. Simulate a run of the protocol as in **Game** d with input $(\mathcal{C}^*, x)$ and g_d fixed to γ_1, but replace each public key h by h' at depth d with parent vertices z_i and z_j as follows.
   (a) If h is in the computation path, then simply set $h' \leftarrow h$.
   (b) If one parent of h, z_j, is in the computation path and the other, z_i, is not, set $h' \leftarrow (\gamma_1^{x_i})^{T_j}$.
   (c) If neither parent of h is in the computation path, set $h' \leftarrow \gamma_1^{y_{i,j}}$.
5. Provide $\mathcal{B}$ with the view of Bob, receiving its output bit b^*.
6. Output 1 if and only if $b = b^*$.

It suffices to show that the view of $\mathcal{B}$ in the above game is identical to its view in **Game** d when $y_{i,j} = x_i x_j$ and that it is identical to its view in **Game** $d+1/2$ when $y_{i,j} \xleftarrow{\$} \mathbb{Z}_{p_d}$. To see this, recall that in both games, all elements off of the computation path of depth $d-1$ are uniformly and independently random. Therefore, the tags $T_z^{(b)}$ at depth $d-1$ that are off of the computation path are uniformly random and independent of all other elements, like the x_i. The claim follows. (3.1) ■

Claim 3.2. *For any probabilistic polynomial-time adversary $\mathcal{B}$, $|\mathsf{Adv}^{\textbf{Game } d+1}(\mathcal{B}) - \mathsf{Adv}^{\textbf{Game } d+1/2}(\mathcal{B})|$ is negligible if DDH is hard in G_d.*

Proof. Each ciphertext off of the computation path is of the form $(h', g_d^r, h'^r T, g_d^s, h'^s g_d^\tau)$ where h', r and s are uniformly random (and the other elements are all efficiently computable). The claim then follows immediately from the assumption that DDH is hard in G_d. (3.2) ■

The result follows from the claims.

□

F Proof of Theorem 4 (security of the PFE firewalls)

Proof of Theorem 4. The proof of the first statement is nearly identical to the proof of Theorem 2 (the corresponding theorem for the oblivious transfer protocol from Section 3).

It should be clear that Alice's reverse firewall maintains correctness. To prove security, fix some tampered implementation $\mathcal{A}$ that maintains functionality. Let S be the simulator from the proof of Proposition 3. For convenience, we also imagine that the simulator calls the $\mathsf{Rerand}_{\mathsf{Garble}}$ function on the garbled circuit that it produces before sending it to Bob. (This is simply a syntactic change.) As in the proof of Proposition 3, we define the *computation path* as the collection of ciphertexts in the garbled circuit that Bob decrypts in a given run. Note that this computation must be well-defined (with all but negligible probability) because $\mathcal{A}$ maintains functionality. We define the following sequence of games.

- **Game 1** is the indistinguishability game between the output of S on input x (with access to the ideal functionality computing $\mathcal{C}$) and the view of Bob in the real game on input $(\mathcal{C}, x)$.
- **Game 2** is **Game 1** in which the group elements in Alice's input messages corresponding to tags and location bits of the form $(T_z^{(1-x_z)}, \tau_z^{(1-x_z)})$ for input vertices z are replaced by uniformly random group elements (in both the simulated world and the real world).
- For $d = 2$ to D, **Game** $d+1$ is **Game** d in which, all ciphertexts (h, u, e, v, w) at depth d that are not in the computation path of Bob are replaced by uniformly random group elements.

The difference between **Game 1** and **Game 2** is simply syntactic. (See the proof of Proposition 1.) As in the proof of Proposition 3, we define an intermediate game, **Game** $d + 1/2$. Note that in **Game** d, the ciphertexts at depth d output by the reverse firewall take the form $(h', g_d'^{r'}u', h'^{r'}e', g_d'^{s'}v', h'^{s'}w')$ where r' and s' are uniformly random (see the function $\mathsf{Rerand}_{\mathsf{Enc}}$ in Figure 10). So, we define **Game** $d + 1/2$ as **Game** d in which each of each public key h' that is not in the computation path is independently replaced by $h'' \xleftarrow{\$} G_d$, so that the full ciphertexts now take the form $(h'', g_d'^{r'}u', h''^{r'}e', g_d'^{s'}v', h''^{s'}w')$.

Claim 4.1. *For any probabilistic polynomial-time adversary $\mathcal{B}$, $|\mathsf{Adv}^{\textbf{Game } d+1/2}(\mathcal{B}) - \mathsf{Adv}^{\textbf{Game } d}(\mathcal{B})|$ is negligible if non-uniform DDH is hard in G_d.*

Proof. We consider the effect of $\mathsf{Change}_{\mathsf{Key}}$ from Figure 10 on the ciphertexts (A_z) corresponding to vertices of depth d. First, consider the ciphertexts corresponding to one such vertex, $A_z = (h_{\boldsymbol{\eta}}, u_{\boldsymbol{\eta}}, e_{\boldsymbol{\eta}}, v_{\boldsymbol{\eta}}, w_{\boldsymbol{\eta}})_{\boldsymbol{\eta}\in\{0,1\}^2}$ with $h_{\boldsymbol{\eta}} \neq 1$. The function $\mathsf{Change}_{\mathsf{Key}}$ sets

$$\mathcal{R}_{\boldsymbol{\eta}} = \mathcal{R}_{L(z)}\mathcal{R}_{\mathcal{R}(z)}g_{d-1}'^{\eta_0\beta_{L(z)}+\eta_1\beta_{R(z)}}$$

and $h'_{\boldsymbol{\eta}} = h_{\boldsymbol{\eta}}^{\alpha_d\mathcal{R}_{\boldsymbol{\eta}}}$ where $\mathcal{R}_{L(z)}$, $\mathcal{R}_{R(z)}$, $\beta_{L(z)}$, $\beta_{R(z)}$ are uniformly random and $\boldsymbol{\eta} = (\eta_0, \eta_1)$. It therefore follows that the collection of *all* public keys h at depth d takes the form $(h_{i,j}^{\alpha_d x_i x_j})_{i<j}$ where the $h_{i,j} \in G_d$ are chosen by either the simulator or the tampered implementation $\mathcal{A}$ and each x_i is a mask $\mathcal{R}_{z'}g_{d-1}'^{\tau\beta_{z'}}$ for some τ and some z' with depth $d-1$. Note that the masks x_i corresponding to the computation path can be efficiently computed from the original circuit and the output of the rerandomization algorithm.

By Lemma 3, it suffices to show that the difference in advantages is small if chosen bases G_{d-1}-subgroup k-DDH is hard in G_d. Recall that in this game, $\mathcal{B}'$ must provide $(\gamma_i)_{i=1}^k$ and $(\delta_{i,j})_{i<j}$, all elements in G_d. It then receives $((\gamma_i^{x_i})_{i=1}^k, (\delta_{i,j}^{y_{i,j}})_{i<j})$ where the x_i are uniformly random elements from G_{d-1} and either $y_{i,j} = x_i x_j$ or they are uniformly random elements in $\mathbb{Z}_{p_d}$. So, let $\mathcal{B}$ be an adversary in **Game** d. Let k be the number of vertices of depth $d-1$. Then, we construct an adversary $\mathcal{B}'$ in the chosen bases G_{d-1}-subgroup k-DDH game in G_d, which behaves as follows.

1. $(\mathcal{C}, x) \xleftarrow{\$} \mathcal{B}()$
2. $y \leftarrow \mathcal{C}(x)$.

3. Sample $b \xleftarrow{\$} \{0,1\}$. If $b = 0$, let $\mathcal{C}^* = \mathcal{C}$. Otherwise, let $\mathcal{C}^* = \mathcal{C}_y$, the constant circuit that always outputs y.
4. Let (z_i) be the vertices of depth $d-1$ in some order. For a run of the $\mathsf{Rerand}_{\mathsf{Garble}}$ function, define $\mathcal{R}_i = \mathcal{R}_{z_i} \cdot g_{d-1}'^{\beta_{z_i} \cdot \tau_i}$ where τ_i is the bit corresponding to the computation path.
5. Simulate a run of the protocol with Alice replaced by the firewall composed with $\mathcal{A}$ if $b = 0$ or the simulator if $b = 1$ until the garbled circuit is produced but *before the* $\mathsf{Rerand}_{\mathsf{Garble}}$ *function is called.*
6. Run $\mathsf{Rerand}_{\mathsf{Garble}}$ on the garbled circuit, and replace all elements that are off of the computation path at depth less than d by uniformly random elements. Note that this defines the elements $(\mathcal{R}_i)$ and α_d.
7. For each public key h at depth d in the *original* garbled circuit with parents z_i and z_j, do the following.
   (a) If one parent of h, z_j is in the computation path and the other, z_i, is not, set $\gamma_i \leftarrow h^{\alpha_d \mathcal{R}_j}$.
   (b) If neither parent of h is in the computation path, set $\delta_{i,j} \leftarrow h^{\alpha_d}$.
8. If any values γ_i or $\delta_{i,j}$ are not set, set them to arbitrary values.
9. Send $((\gamma_i)_{i=1}^k, (\delta_{i,j})_{i<j})$ to the G_{d-1}-subgroup k-DDH challenger, receiving $((\gamma_i^{x_i})_{i=1}^k, (\delta_{i,j}^{y_{i,j}})_{i<j})$.
10. For each public key h' at depth d in the *rerandomized* garbled circuit with parents z_i and z_j, do the following.
    (a) If h' is in the computation path, set $h'' \leftarrow h'$.
    (b) If h' has one parent z_j in the computation path and the other z_i is not, set $h'' \leftarrow \gamma_i^{x_i}$
    (c) If neither parent of h'' is in the computation path, set $h'' \leftarrow \delta_{i,j}^{y_{i,j}}$.
11. Finish the run of the protocol by sending Bob the modified garbled circuit.
12. Provide $\mathcal{B}$ with the view of Bob, receiving result b^*.
13. Output 1 if and only if $b = b^*$.

It suffices to show that the view of $\mathcal{B}$ in the above game is identical to its view in **Game** d when $y_{i,j} = x_i x_j$ and that it is identical in **Game** $d + 1/2$ when $y_{i,j} \xleftarrow{\$} \mathbb{Z}_{p_d}$. To see this, recall that in **Game** d, all elements off of the computation path of depth $d-1$ are uniformly and independently random. Therefore, the random masks $\mathcal{R}_z \cdot g_{d-1}'^{\beta_z \cdot \tau}$ at depth $d-1$ that are off of the computation path are uniformly random and independent of all other elements. The claim follows. (4.1) ■

Claim 4.2. *For any probabilistic polynomial-time adversary* $\mathcal{B}$, $|\mathsf{Adv}^{\textbf{Game } d+1}(\mathcal{B}) - \mathsf{Adv}^{\textbf{Game } d+1/2}(\mathcal{B})|$ *is negligible if DDH is hard in* G_d.

Proof. Each ciphertext off of the computation path is now of the form $(h'', g_d'^{r'} u, h''^{r'} e, g_d'^{s'} v, h''^{s'} w)$ where h'', r', and s' are uniformly random and the other elements are all efficiently computable (see the procedure $\mathsf{Rerand}_{\mathsf{Enc}}$). The claim then follows immediately from the assumption that DDH is hard in G_d. (4.2) ■

Claim 4.3. *The two views in* **Game** $D + 1$ *are identically distributed.*

Proof. It should be clear that the input messages are identically distributed in both views. In both garbled circuits, all elements not in the computation path are uniformly and independently random. Along the computation path, the ciphertexts (h', u', e', v', w') take the form $(g_d'^{r'}, h'^{r'} T', g_d'^{s'}, h'^{s'} g_d'^{\tau'})$ where r' and s' are uniformly and independently random (see the procedure $\mathsf{Rerand}_{\mathsf{Enc}}$) and g_d' is uniformly random but fixed for each depth. Since the computation path is well-defined, τ' must be a uniformly random bit (see the procedure Flip). For ciphertexts that do not correspond to output vertices, T' is also uniformly and independently random (see the procedure Change_T). Because $\mathcal{A}$ maintains functionality, tags in the computation path corresponding to output vertices satisfy $T' = g_d'^b$ where b is the corresponding output bit of $\mathcal{C}(x)$ with all but negligible probability (again, see the procedure Change_T). Similarly, h' is uniquely determined by the elements before it in the path (or in the input messages)—in particular, it is $g_d'^{T'_L \cdot T'_R}$ where T'_L, T'_R are the tags corresponding to the parents of the current node (see the procedure $\mathsf{Change}_{\mathsf{Key}}$). The computation path itself is uniquely determined by the bits τ' (see the procedure $\mathsf{Permute}$).

The claim follows. (4.3) ■

To see that Alice's reverse firewall is strongly exfiltration-resistant, consider the above argument in which the computation path is empty.

The result follows. □