ZQL: A Compiler for Privacy-Preserving Data Processing
Cédric Fournet, Markuld Kohlweiss, and George Danezis, Microsoft Research; Zhengqin Luo, MSR-INRIA Joint Centre

This paper is included in the Proceedings of the 22nd USENIX Security Symposium.
August 14–16, 2013 • Washington, D.C., USA
ISBN 978-1-931971-03-4

Open access to the Proceedings of the 22nd USENIX Security Symposium is sponsored by USENIX
ZQL: A Compiler for Privacy-Preserving Data Processing

Cédric Fournet  
Microsoft Research

Markulf Kohlweiss  
Microsoft Research

George Danezis  
Microsoft Research

Zhengqin Luo  
MSR-INRIA Joint Centre

Abstract

ZQL is a query language for expressing simple computations on private data. Its compiler produces code to certify data, perform client-side computations, and verify the correctness of their results. Under the hood, it synthesizes zero-knowledge protocols that guarantee both integrity of the query results and privacy for all other data.

We present the ZQL language, its compilation scheme down to concrete cryptography, and the security guarantees it provides. We report on a prototype compiler that produces F# and C++. We evaluate its performance on queries for smart-meter billing, for pay-as-you-drive insurance policies, and for location-based services.

1 Introduction

A variety of private user data is used to tailor modern services, and some go as far as billing based on fine grained customer readings. For example, smart meters are used to charge a different tariff depending on the time of electricity usage; pay-as-you-drive insurance premiums depend on detailed driving pattern of drivers. Such schemes are currently implemented by collecting fine-grained information, and processing it on the service side—an architecture that has led to serious privacy concerns.

This paper supports an alternative approach: clients could perform sensitive computations on their own data certified by meters or car on-board units [55, 60], and upload only the results, together with a proof of correctness to ensure their integrity. We propose ZQL, a simple query language to express at a high level such computations, without any cryptographic details. Queries are compiled to code for the data sources, the clients, and the verifiers by synthesizing zero-knowledge protocols.

The most popular language for querying and performing computations on user data is SQL [29] based on relational algebra. The ZQL feature set was chosen to support a subset of SQL. Data is organized into tables of rows, with private and public columns. Queries accept tables as inputs, and can iterate over them to produce other tables, or aggregate values. Simple arithmetic operations on rows are supported natively, and so is a limited form of SQL joins through table lookups.

ZQL offers advantages over hand-crafted protocols, in that computations are flexible and can be expressed at a high level by application programmers. The computations can also be modified and recompiled, without the need to involve cryptography experts.

The ZQL compiler is free to synthesize custom zero-knowledge protocols behind the scene, and we currently support two main branches, for RSA and Elliptic Curve primitives. We also support a symbolic execution backend to derive estimates of the cost of evaluating and verifying queries. Synthesized protocols themselves are internally represented and optimized as fragments of an extended ZQL language until the final code is emitted. Intermediate ZQL is strongly typed, and precise refinement types can be used to verify security properties on the final compiled code, using F7 [16] or F* [58].

Informally, for a given source query, the desired security properties on the resulting ZQL-compiled code are:

- **Correctness.** For any given source inputs, the sequential composition of the cryptographic queries for the data sources, the user, and the verifier yields the same result as the source query.
- **Integrity.** An adversary given the capabilities of the user cannot get the verifier to accept any other result—except with a negligible probability.
- **Privacy.** An adversary given the capabilities of the verifier, able to choose any two collections of inputs such that the source query yields the same result, and given the result of the user’s cryptographic query, cannot tell which of the two inputs was used.

This corresponds to the source query being executed by a fictional trusted third party sitting between the data sources, the user, and the verifier.
Contents The rest of the paper is organized as follows. §2 introduces our query language using a series of privacy-preserving data processing examples. §3 specifies our target privacy and integrity goals. §4 reviews the main cryptographic mechanisms used by our compiler. §5 describes the compilation process. §6 gives our main security theorems. §7 discusses applications and §8 provides experimental results and discusses future work.

This short version of the paper omits many details and discussions; an extended version with auxiliary definitions, proofs, and examples is available at http://research.microsoft.com/zql.

Related work The ZQL language provides private data processing. The zero-knowledge protocols synthesized are standard \( \Sigma \)-protocols [32, 30, 33], but in ZQL they are used for proving the correctness of general computations rather than for cryptographic protocol design.

Arguably, previous works on zero-knowledge compilers focused on the latter as the primary use-case [19, 51, 2, 1]. The use of zero-knowledge for authentication and authorization as in credential and e-cash technologies [23, 51, 4] received particular attention, but, to our knowledge, no-one considered the use of \( \Sigma \)-protocols to prove the execution of general programs.

More specifically, a long line of work [19, 7, 2] culminating in the CACE compiler tackles the problem of automatically translating proof goals specified in the Camenisch-Stadler notation [22] into efficient \( \Sigma \)-protocols. Intermediate translations steps of ZQL (the shared translation) are at a similar level of abstraction to the Camenisch-Stadler notation but ZQL also synthesizes those representations from source code, and then proceeds to compile them to low level operations. ZKPDL [51], an alternative compiler for \( \Sigma \)-protocols, uses a natural language inspired specification of zero-knowledge proof goals. This specification language may be even closer in spirit to our intermediary notation, as it allows for the possibility to specify the generation of the protocol inputs. The authors of the CACE compiler discuss the difference and similarity between these two approaches in a Usenix poster [9]. The cryptographic prototyping language Charm [1] also includes a zero-knowledge proof compiler for Camenisch-Stadler statements which is currently primarily a proof of concept and thus less sophisticated than CACE and ZKPDL. We are also aware of an embedding of a zero-knowledge language in C++ [45].

ZQL differs from standard multi-party computation compilers [49], in that it assumes the client knows all private data. This assumption allows for single round protocols, and the efficient non-interactive implementation of non-linear operations including joins.

Figure 1: ZQL in a privacy friendly computation system.

2 A Language for Private Data Processing

Why ZQL? We design ZQL to support privacy protocols that rely on client-side computation while requiring high integrity [35]. In this setting, a number of (possibly independent) data sources provide signed personal data items to a user. The signed data is then used as an input to some computation performed on a user device on which the operation of a service relies (for example billing for a utility, determining the proximity of to a specific path, or profiling the shopping habits of a user). The results of the computation are then sent to the relying service, while private input data is kept secret. The ZQL compiler takes a high level description of the computation and is responsible for producing the code executed by the data sources to sign personal data, the computation prover and the computation verifier, as illustrated in Figure 1.

It is assumed that communications take place over private authenticated channels; the data sources are trusted by all to maintain the privacy of the raw personal data they produce, and to securely sign them. Given this, our protocols guarantee integrity through cryptographic proofs that establish the authenticity of the personal data inputs and the correctness of a particular computation. Thus, a malicious client cannot manipulate the result of the computation. On the other hand, the private inputs to the computations are kept secret by the user, and the proofs do not leak anything about them. Thus, privacy is preserved, and only the result of the computations (and any inferences that can be drawn from them) become known to the relying service. Compiling allows us to statically verify the security of the resulting protocols using refinement types (see the full paper). Hence, both the prover and the verifier, or anyone they trust, can separately review the source query, compile the protocol, and verify its security by typing.

There are advantages in de-coupling data sources from specific computations. It allows for meters, or services providing personal data, to remain simple, cheap, and generic. In turn, the computations, such as billing, can be
The ZQL language

At its core, ZQL is a pure expression language, with built-in operators that act on integers and tables. Figure 2 gives its abstract syntax. A query \( \theta \rightarrow e \) consists of the declaration of input variables (\( \theta \)) that can be either public or private, and of an expression body (\( e \)).

Expressions consist of variables, operators applied to sub-expressions \( \bar{e} \) (including constants as a special case when \( \bar{e} \) is empty), and let bindings for sequential composition. Expressions evaluate to tuples of values: for example, the expression \( \text{let } x : \text{int}, y : \text{int} = e \text{ in } e_0 \) first evaluates the sub-expression \( e \) to a pair of integers \( v_x, v_y \), then evaluates \( e_0 \) after substituting \( v_x \) and \( v_y \) for \( x \) and \( y \).

A variety of operators support arithmetic (0, 1, +, *), booleans (=, \&\&), and operations on tables (map, fold, lookup). The iterators map and fold are parameterized by a ZQL expression, conceptually acting as the body of the corresponding loop. (We write these expressions as functions, but they can only specialize the iterator; they cannot be assigned to variables.)

Query inputs and expression results are specified using tuples of typed variables (\( \theta \) for query inputs and \( \rho \) for sub-expressions). Each base type can be marked as public, and is otherwise treated as private. Types also include tables, where \( \rho \) indicates the type of each row in the table. Tables can contain mixtures of public and private columns; for example, \((\text{time: int pub, reading: int})\text{table}\) is the type of tables of private readings indexed by public times. On the other hand, the current ZQL compiler does not attempt to hide the query definition itself, or the number of rows in tables.

Intermediate expressions are automatically classified as public or private, depending on the types of their variables, following a standard information flow discipline: public inputs can flow to private results, but not the converse. Alternatively, a ZQL expression can be explicitly declassified, using the special operator \( \downarrow \) which specifies that the result of \( e \) can be released to the verifier, and marks it as public.

A ZQL query itself defines the privacy goals of the synthesized zero-knowledge protocols. For example, a query \( \theta \rightarrow \downarrow e \), where \( e \) does not contain any declassification, states that only the final result of the query is released, and that the protocol should not leak any side information on inputs marked as private in \( \theta \). A key feature of the language is that the underlying cryptographic mechanisms are totally hidden in the definition of the ZQL query. Since the ZQL query defines what results are declassified, it is important that users, or their proxies, review it to ensure no more than the necessary information leaks from it. Additional privacy mechanisms, such as differential privacy [37], could be used to measure or minimize any leakage resulting from the query declassification.

The ZQL language is strongly typed, with a type system simple enough to allow for automated type checking and type inference, which means that the programmer only has to write the input types of the query. We write \( \Gamma \vdash e : \rho \) to state that expression \( e \) has type \( \rho \) in environment \( \Gamma \). The type system ensures both runtime safety: \( e \) returns only to values of types \( \rho \), and non-interference: in the absence of declassification, \( e \) does not leak inputs typed as private in \( \Gamma \) to results typed as public in \( \rho \). The type system can also be used to track the maximal length of private variables to statically prevent arithmetic overflows. We omit the formal definition of the language semantics and type system, which are standard. Internally, ZQL relies on a richer type system with refinements types [14, 42] to keep track of various properties and to structure our security proofs—see the full paper.

ZQL by example

We present the ZQL language and semantics through simple concrete examples, building to fuller queries that address problems in the literature in \$7. The first example query computes the discriminant of the polynomial \( xk^2 + zk + y \), for public \( x \) and private \( y \) and \( z \).

\[
\text{let discriminant } (x:\text{int pub}) (y:\text{int}) (z:\text{int}) = \downarrow (z^2 - 4*x*y)
\]
Anticipating on its compilation, the part of the expression that is linear in the secrets, namely \( y - 4 \times x \times y \), can be proved efficiently through homomorphisms of Pedersen commitments, while the non-linear \( z \times z \) requires a \( \Sigma \)-protocol to prove the correctness of the private multiplication. The ZQL compiler will choose to synthesize the right proof mechanisms for each case.

The query declassifies its result, which leaks some information about \( y \) and \( z \). For instance, given \( x = 30 \) and discriminant \( x \times y = 1000 \), if the verifier knows a priori that \( 0 < y < 200 \) and \( 0 < z < 200 \), then it can infer that \( (y, z) \) is one of the pairs \((5, 40), (45, 80), (75, 100)\) or \((155, 140)\), but our privacy theorem ensures that its does not learn which pair was actually used.

Our next examples illustrate the use of tables and iterators \texttt{map} and \texttt{fold}. The first query computes the sum of all integers in table \( X \), while the second returns the sum of their squares. The third query takes a table with a public column and two secret columns and returns a table with the same public column, and the element-wise sum of the secret columns. By design, the size of the tables is not hidden by ZQL. (Hiding table sizes naively would involve padding the computation to the maximum size of allowed tables, which would be very expensive.)

\[
\begin{align*}
\text{let } \text{sum} \_\text{of} \_\text{X} (X : \text{int table}) &= \\
& \downarrow (\text{fold} ((s, x) \rightarrow s + x) 0 X) \\
\text{let } \text{sum} \_\text{of} \_\text{square} (X : \text{int table}) &= \\
& \downarrow (\text{fold} ((s, x) \rightarrow s + x \times x) 0 X) \\
\text{let } \text{linear} (T : \text{int pub + int int table}) &= \\
& \downarrow (\text{map} ((a, x, y) \rightarrow a \times x + y) T)
\end{align*}
\]

In these queries, the iterators are parametrized by a subquery, which is applied to every row of the table, accumulating the sums in \( s \), or building another table of results. The equivalent SQL statements would be written:

\[
\begin{align*}
\text{select } \sum(x) \text{ from } X \\
\text{select } (x \times x) \text{ from } X \\
\text{select } a, x+y \text{ from } T
\end{align*}
\]

The equivalent SQL statement would be select \( F \_\text{country} \) from \( X, F \) where \( F \_\text{city} = X \_\text{city} \). The query implementation relies on a data source that issues a signed table from cities to countries.

3 Security

The next two sections provide rigorous security definitions for what the ZQL compiler achieves and the cryptographic building blocks it uses, necessary for formulating our security theorems in §6. The mere fact that we can give formal cryptographic definitions for a large class of cryptographic protocols relies on our simple expression language having a formal semantic for both source and compiled programs. Readers interested in compiler architecture can jump straight to §5, or those curious about applications can find them in §7.

Notations Consider a well-typed ZQL source query \( Q \equiv \theta \rightarrow e \), with \( \ell \) input variables \( \theta = (\zeta_i : r_i)_{i=0,..,\ell-1} \), that declassifies only its result. As explained in §2, the typed variables \( \theta \) specify the data sources and privacy policy. Let \( T \) range over values of type \( \theta \), and \( R = Q(T) \) be the corresponding query result. Given \( Q \), our compiler produces queries \( (S, (K_i, D_i))_{i=0,..,\ell-1}, P, V ) \) with formal parameters indicated in parentheses as follows. (We use primed variables for compiled values.)

- \( S \), the setup generator, generates global parameters \( \chi \) used by the commitment scheme;
- \( (K_i)_{i=0,..,\ell-1} \), the data sources key generation, generate key pairs \( sk_i, vk_i := K_i(\chi) \) used to sign data and verify their signatures;
- \( (D_i)_{i=0,..,\ell-1} \), the data sources, extend and sign each input: \( T'_i := D_i(\chi, sk_i, T_i) \);
- \( P \), the prover, produces an extended result from extended inputs: \( R' := P(\chi, vk, T') \);
- \( V \), the verifier, returns either some source result \( R := V(\chi, vk, R') \) or a verification error.

Main Properties We first define functional correctness when all participants comply with the protocol.
Definition 1 \((S, (K_i, D_i)) = (0..t-1, P, V)\) correctly implements the source query \(Q\) when, for any source inputs \(T : \theta\) and \(\chi := S, (k_i, v_k) := K_i(\chi)\) \(= 0..t-1\), we have

\[
V(\chi, v_k, P(\chi, v_k, D_i(\chi, k_i, T_i)) = Q(T).
\]

We define privacy as indistinguishability between two series of chosen inputs that yield the same query result.

Definition 2 Given a source query \(Q\) and an adversary \(\mathcal{A}\), let \(Adv_{ps}^{\mathcal{A}} = |2Pr[\mathcal{A} \text{ wins}] - 1|\) where the event ‘\(\mathcal{A}\) wins’ is defined by the following game:

1. The challenger runs \(S\) and \(K_i\) to generate setup \(\chi\) and keys \(sk, v_k\); it provides \(\chi\) and \(v_k\) to \(\mathcal{A}\).
2. The adversary \(\mathcal{A}\) provokes two vectors of input data \(\widehat{T}^0 : \theta\) and \(\widehat{T}^1 : \theta\) such that (a) they coincide on public data and (b) \(Q(\widehat{T}^0) = Q(\widehat{T}^1)\).
3. The challenger picks a random bit \(b\), encodes the corresponding inputs \(\chi^b\), and generates \(R^b := \chi^b, v_k, T^b\).
4. Given \(R^b, \mathcal{A}\) returns his guess \(b'\), and wins iff \(b = b'\).

\((S, (K_i)) = (0..t-1), (D_i) = (0..t-1), P, V)\) is \((t, \varepsilon)\)-private when, for all \(\mathcal{A}\) running at most for time \(t\), we have \(Adv_{ps}^{\mathcal{A}} \leq \varepsilon\).

Note that we do not formally provide privacy protection against corrupted data sources. To strengthen our scheme against data source attacks, we would have to rerandomize all cryptographic material flowing from data sources to verifiers, which precludes our efficient use of homomorphic commitments.

We define integrity as a game in which an adversary has to produce an invalid but accepted response.

Definition 3 Given a source query \(Q\) and an adversary \(\mathcal{A}\), let \(Adv_{sm}^{\mathcal{A}} = Pr[\mathcal{A} \text{ wins}]\) where the event ‘\(\mathcal{A}\) wins’ is defined by the following game:

1. The challenger runs \(S\) and \(K_i\) to generate setup \(\chi\) and keys \(sk, v_k\); it provides \(\chi\) and \(v_k\) to \(\mathcal{A}\).
2. The adversary \(\mathcal{A}\) can adaptively corrupt data sources \(D_i\) to get their signing keys \(sk_i\) and, at the same time, it can obtain signed inputs \(T'_i := D_i(\chi, sk_i, T_i)\) for source inputs \(T_i\) of its choice.
3. Valid results are source values \(R = Q(\widehat{T})\) such that, for each \(i\), either \(i\) was corrupted or \(T_i\) was signed. The adversary wins if he outputs \(R^*\) such that \(V(\chi, v_k, R^*)\) returns any invalid result \(R^*\).

\((S, (K_i)) = (0..t-1), (D_i) = (0..t-1), P, V)\) is \((t, \varepsilon)\)-sound when, for all \(\mathcal{A}\) running at most for time \(t\), we have \(Adv_{sm}^{\mathcal{A}} \leq \varepsilon\).

Depending on the adversary, there can be zero, one, or numerous valid responses. In fact, depending on the query and the input tables, whether a response is valid may not even be efficiently checkable. The definition is, however, still meaningful.

4 Main Cryptographic Tools (Review)

Signatures A digital signature scheme allows everyone in possession of the verification key \(vk\) to verify the authenticity of data signed by the owner of the corresponding signing key \(sk\). We use signatures to let verifiers authenticate data sources. Instead of signing private data in the clear, data sources sign public commitments; thus, the resulting signature tags are also public.

Cryptographic groups Besides conventional digital signatures, for which we use standardized schemes, our remaining cryptographic tools can either be specified for composite order groups, obtained by computing modulo an RSA modulus, or for prime order groups with a bilinear pairing. We use the latter for our presentation and formal analysis as it offers both performance and conceptual advantages.

Let \(G, \hat{G},\) and \(G_T\) be groups of prime order \(q\). Let \(g \in G\) and \(\hat{g} \in \hat{G}\) be generators of \(G\) and \(\hat{G}\) respectively. A bilinear pairing is an efficiently computable function \(\hat{e} : G \times \hat{G} \rightarrow G_T\) that is bilinear, i.e. \(\forall a, b \in \mathbb{F}_q : \hat{e}(g^a, \hat{g}^b) = \hat{e}(g, \hat{g})^ab\) and non-degenerate, i.e. \(\hat{e}(g, \hat{g}) \neq 1\). Whenever possible we perform all operations in the base group \(G\) with the shortest representation.

Commitments A commitment scheme allows a user to commit to a hidden value such that he can reveal the committed value at a later stage. The properties of a commitment scheme are hiding: the committed value must remain hidden until the reveal stage, and binding: the only value which may be revealed is the one that was chosen in the commit stage. We use the perfectly hiding commitment scheme proposed by Pedersen [54]: given a group \(G\) of prime order \(q\) with generators \(g\) and \(h\), generate a commitment \(C_i\) to \(x \in \mathbb{F}_q\) by sampling a random opening \(\alpha \leftarrow \mathbb{F}_q^*\) and computing \(C_i = g^{\alpha h^i}\). The commitment is opened by revealing both \(\alpha\) and \(x\).

Two useful properties of Pedersen commitments are (i) their homomorphic property that allows to derive a commitment to the linear combination of input values; and (ii) their algebraic structure that allows for efficient zero-knowledge proofs. For RSA groups, we use commitments with similar properties [43, 34].

Zero-knowledge proofs [59, 39, 11] provide a verifying algorithm with an efficient means for checking the truth of a statement by guaranteeing that given access to a successful proof generation algorithm one can extract a secret witness for said truth. At the same time, zero-knowledge proofs [47, 46], and the related concepts of witness indistinguishable proofs [38, 32], allow the prover to keep this witness secret. We make use of a long line of work on efficient proofs of conjunctions of discrete logarithm (DL) representations [57, 28, 52, 32, 30, 18, 26, 33, 50]. For non-linear computations such as
multiplication, we use the approach of Brands [18], Camenisch [26], and Cramer and Damgård [31].

DL representation proofs are interactive protocols of three or more messages. To ease deployment and minimize communications, we use the Fiat-Shamir Heuristic [40] and replace random messages sent by the verifier with hash function computations. The resulting protocols can still be formally analyzed in the random oracle model [12, 62].

Proof compatible signatures The combination of zero-knowledge proofs and digital signatures allows us to prove authentication properties on private data, such as, for instance, the existence and properties of a matching row when performing a private lookup.

We use CL signatures [20], as they are compatible with DL representation proofs. The original scheme was proven secure under the Strong RSA assumption and requires groups with hidden order [6, 24]. Other CL signature proposals rely on a variety of assumptions based on bilinear pairings [21, 17, 3, 25] and require more standard prime order DL-representation proofs. We also use the scheme of [25], a good trade-off between security and performance. An additional benefit is that it is syntactically very close to RSA-based CL signatures.

To certify our lookup tables, data sources extend each row of the table with a CL signature. For instance, tables of triples of private integers \((x_0, x_1, x_2)\) are extended to tables with rows of the form \((x_0, x_1, x_2, e, v, A)\).

The verification equations for RSA and bilinear pairing based CL signatures are of the form \(Z = A^*R_0^{x_0}R_1^{x_1}R_2^{x_2}S^y\) and \(\tilde{e}(Z, g) = \tilde{e}(A, pk^*g^e)\tilde{e}(R_0^{x_0}R_1^{x_1}R_2^{x_2}S^y, g)\) respectively, where \((Z, R_0, R_1, R_2, S, pk)\) are group elements that form the components of the verification key \(vk\). Both verification equations can be proven using efficient DL representations. The security of these two schemes is based on the strong RSA assumption and the strong Diffie-Hellman (SDH) assumption respectively.

5 Compiler Architecture

Protocol Overview The ZQL compiler takes a source query, which contains no cryptographic computations, and automatically produces programs for each data source, for the prover, and for the verifier.

First, the compiler augments the source query with various cryptographic commitments to secrets and representation equations to generate a shared translation that will lead to both prover and verifier code. Some commitments are computed and signed by the data sources that certify the computation inputs, and simply passed to the prover and verifier programs. Others, representing intermediate secrets in the query, are interleaved with the source computation: for any such secret \(x\), the prover may sample a secret opening \(\alpha_x\), compute a Pedersen commitment \(C_x = g^{\alpha_x}h^x\), and send it to the verifier; and the verifier may check it using a zero-knowledge proof.

Linear relations between secrets do not require complex zero knowledge proofs, as they can be checked by the verifier simply by using the homomorphisms of Pedersen commitments. For example, a private sum \(z = x + y\) will have commitment \(C_z = g^{\alpha_x}C_x \ast C_y\). Such commitments need not be transmitted, as they can be recomputed by the verifier. On the other hand, non-linear relations between secrets, including multiplication and table lookup, require \(\Sigma\)-protocol proofs to be synthesized. For instance, to prove that \(z\) is the product of a secret \(x\) committed in \(C_x\) and a secret \(y\), one proves the conjunction of the representation equations \(C_x = g^zh^x\) and \(1 = g^x(C_x)^{-1}g^yh^z\). Note that the second equation uses a variable commitment \(C_x\) as a base.

All \(\Sigma\)-protocols used in the compiler come down to proving knowledge of the secret values underlying the discrete logarithm representations of public group elements, and equality relations between the secret values. Assume the ZQL query reduces to proving in zero-knowledge the representations \(\tilde{C} = g^x\tilde{e}(\tilde{x})\) of a number of commitments \(\tilde{C}\), represented by public group elements, using a number of secrets \(\tilde{x}\) (including secret openings). For the multiplication example above, we have two equations on five secrets: \(\tilde{C} \equiv (C_x, 1), \tilde{x} \equiv (x, \alpha_x, y, z, \alpha_z)\) and \(\tilde{e}(\alpha, \beta, \gamma, \delta, \epsilon) \equiv (g^{\alpha}, h^\beta, C_x^{-1}g^\delta h^\epsilon)\). The zero-knowledge protocol synthesized works as follows. The prover

1. samples a vector of random values \(\tilde{r}\), one for each secret in \(\tilde{x}\); we call \(\tilde{r}\) values the proof randomness;
2. computes the challenge \(c = H(\tilde{e}(\tilde{r}))\);
3. computes the responses \(\tilde{r} = \tilde{r} - \tilde{c} \ast \tilde{x}\), for all secrets.

The proof sent to the verifier consists of the public parameters and values, the commitments \(\tilde{C}\), the global challenge \(c\), and the responses \(\tilde{r}\). The verifier checks that \(H(\tilde{C} \ast g^x\tilde{e}(\tilde{r})) = c\), which ensures that the prover knows the secret values in the commitments [40, 12]. As detailed below, our compiled prover and verifier programs introduce secrets and process equations on the fly, depending on the query and its inputs.

Once the shared translation is decided, its specialization into prover and verifier code is relatively straightforward. It involves mainly ensuring the right data flows within the query processing to compute all commitments and responses, and to correctly verify them in the same order. The inputs of the shared translation also determine the data source programs that generate keys, compute commitments, and sign extended data.

Embedding cryptography within ZQL Our compiler mostly operates within ZQL, with F# and C++ back-ends to turn the compiled queries into executable code. This
enables us to reason about code in a simple, domain-specific language. To this end, Figure 3 supplements the source language of Figure 2 with the types and operators for expressing cryptographic operations. Expressions are extended with assert, used in the shared translation to embed proof obligations. As an invariant, all asserted equations \( \varphi \) must hold at runtime. We have types and operations for integers modulo \( q \) (\( \mathbb{F}_q \), written num), for group elements (elt\( G \)), and for bitstrings, and more specific sub-types to keep track of their usage. For instance, hash is the sub-type of bitstrings representing cryptographic hashes, and xopening is a sub-type of num tracking openings generated for the secret value \( x \). In our presentation, we use standard abbreviated forms for their operations; for instance we often omit group parameters, writing \( g^i \) for exp\( G \, g \, x \).

**Setup and Key Generation** The abstract setup 5 produces global parameters \( \chi \) supplied by our cryptographic runtime, including \( q \), the prime order of \( G \), \( \hat{G} \), and \( G_T \); and independent, random generators \( g \), \( h \), \( (R_i)_{i=0..\ell} \), \( S \), \( Z \) in \( G \); and \( \hat{G} \) in \( G \). Its fixed code is provided by our cryptographic libraries.

We use \( D_{LT} \subseteq 0..\ell - 1 \) to denote the subset of data sources that sign lookup table. The key generation \( K_i \) is defined as keygen\( \chi \) when \( \tau \) is a scalar or a table (\( i \notin D_{LT} \)), and as the CL-key generation let \( sk = sample() \, in \, sk, (\hat{g})^k \) when \( \tau \) is a lookup table (\( i \in D_{LT} \)). The data source code \( D_i \) is explained below, as we discuss these two representations.

**Shared Translation** We extend the source query with openings and commitments, but not yet with the corresponding proof randomness and responses.

The main difficulty of the translation is to select cryptographic mechanisms, and notably intermediate commitments, to run the private computation: for every private sub-expression, our compiled protocol may rely on zero, one, or more Pedersen openings and commitments, and it may allocate some proof randomness or not.

In this presentation, for simplicity, we give a formal translation that assumes that all source private integer variables are handled uniformly, with a commitment in the same group, sharing the same bases, and (later) with a proof randomness for the secret and for its opening. Figure 4 and 5 show how we translate types and expressions, respectively, in this special case. We discuss our general, more efficient compilation scheme below.

A source expression is public in a typing environment when all its free variables have public types. The translation leaves public types (1) and expressions (3) unchanged. The translation of a private integer expression is a triple of an integer for the source value, its opening, and its commitment, with the types given on line (2).

**Fresh commitments** Our compilation rules may require openings and commitments on their arguments, and may not produce openings and commitments on their results. Our compiler attempts to minimize those cases. Nonetheless, assuming for instance that we need a commitment for \( z \), we produce it on demand, using the expression abbreviation Commit \( z \) below.

**Commit**

\[
\text{Commit } z \triangleq \\
\begin{align*}
\text{let } oz & : z \, \text{opening} = \text{sample}() \, \text{in} \\
\text{let } C_z & : z \, \text{commitment} = g^z \, \text{exp} G \, h^{oz} \, \text{in} \\
\text{assert } C_z & = g^z \, \text{exp} G \, h^{oz} \\
\end{align*}
\]

The translation is compositional, as can be seen on lines (4,5,6) in the figure. For instance, we translate let expressions by translating their two sub-expressions, and we translate source maps to maps that operate on their translated arguments.

The translation assumes prior rewriting of the source query into simpler sub-expressions. For instance, to compile the discriminant query of §2, we first introduce intermediate variables for the private product and the declassification, rewriting expression \( \downarrow (z \ast z - 4 \ast x \ast y) \) into

\[
ed \triangleq \text{let } p = z \ast z \, \text{in} \, \text{let } d = p - 4 \ast x \ast y \, \text{in} \, \downarrow d.
\]

As a sanity check, our translation preserves typing, in an environment extended with the constants used in our cryptographic libraries; variants of this lemma with more precise refinement types for the prover and verifier translation can be used to verify their privacy and integrity.

**Lemma 1 (Typing the shared translation)** Let \( \Gamma_0 \triangleq g, h, Z, R_0, \ldots, R_{\ell}, S : e l t G, \hat{g}, (p k_i)_{i \in D_{LT}} : e l t \hat{G} \). If \( \Gamma \vdash e : \rho \), then \( \Gamma_0, [\Gamma] \vdash [\Gamma \vdash e] : [\rho] \).

Next, we explain and illustrate the base cases of the shared translation on private expressions.
\[
\begin{align*}
\llbracket x : \tau(\varphi) \rrbracket &= x : \tau(\varphi) \quad \text{when } \tau \text{ is public;} \\
&\quad \text{otherwise:} \\
\llbracket x : \text{int}(\varphi) \rrbracket &= x : \text{int}(\varphi), \quad \text{ox} : x \text{ opening}, \\
C_z : x \text{ox commitment} \\
\llbracket \rho \text{ table} \rrbracket &= \llbracket \rho \rrbracket \text{ table}, \quad s : \text{tag} \\
\llbracket \rho \text{ looketable} \rrbracket &= (\rho, \sigma) \text{ table} \\
\sigma &= e : \text{num}, v : \text{num}, A : \text{elit}_G \\
[\varepsilon] &= \varepsilon \\
\llbracket x : \tau(\varphi), \rho \rrbracket &= \llbracket x : \tau(\varphi), \| \rho \rrbracket
\end{align*}
\]

Figure 4: Shared translation of types and environments

*Expressions affine in private variables* are translated by supplementing the expression with a linear expression on openings and an homomorphic product of commitments (7); we easily check that the resulting triple \((z, \alpha, C_z)\) is such that \(C_z = g^z \ast_G h^\alpha\). Note that the public constant \(a_0\) is not included in the opening computation.

*Expressions polynomial in private variables* are translated using an auxiliary representation equation for every product of private expressions, depending on the availability of openings and commitments—see translation rule (8). To illustrate affine and quadratic expressions, let us translate the discriminant query \(\theta \rightarrow \downarrow(e_d)\) where the source environment \(\theta = x : \text{int pub}, y : \text{int}, z : \text{int}\) specifies that \(x\) is public, whereas \(y\) and \(z\) are private. By definition, the translated environment \(\llbracket \theta \rrbracket\) is

\[
\begin{align*}
x &: \text{int pub}, \\
y &: \text{int, oy} &: y \text{ opening}, C_y &: y \text{oy commitment}, \\
z &: \text{int, oz} &: z \text{ opening}, C_z &: z \text{oz commitment}
\end{align*}
\]

and, from the translation invariant, we already know that \(C_y = g^y h^\alpha\) and \(C_z = g^z h^\beta\). Applying rules (4), (8), (7), and (10) and inlining the definition of \text{Commit} we arrive at the shared translation

\[
\text{let } p, op, C_p = \\
\text{let } p = z * z \text{ in} \\
\text{let } d' = \alpha * z \text{ in} \\
\text{assert } 1 = (C_y)^z \ast_G g^{-p} \ast_G h^{-d'}; \quad (E_1) \\
\text{let } op = \text{sample()} \text{ in} \\
\text{let } C_p = g^p \ast_G h^{op} \text{ in} \\
\text{assert } C_p = g^p \ast_G h^{op}; \quad (E_2) \\
(p, op, C_p) \\
\text{let } d, \text{ od, } C_d = \downarrow (p - 4 \ast x \ast y), (ap - 4 \ast x \ast \alpha y), (C_p \ast C_y^{4 * x y}) \downarrow d
\]

and we easily check that \(C_d\) is a commitment to \(z^2 - 4xy\) with opening \(op - 4x \ast \alpha y\). The code of the shared translation makes explicit the two representation equations for the private multiplication, presented more abstractly at the beginning of §5. Anticipating on the next stages of

\[
\begin{align*}
\llbracket \Gamma \vdash e \rrbracket &= e \quad \text{when } e \text{ is public} \\
\llbracket \Gamma \vdash x \rrbracket &= \llbracket \Gamma(x) \rrbracket \quad \text{otherwise} \\
\llbracket \Gamma \vdash \text{let } \rho = e \text{ in } e_0 \rrbracket &= \llbracket \Gamma \vdash e \text{ in } \llbracket \Gamma, \rho \vdash e_0 \rrbracket \\
\llbracket \Gamma \vdash \text{map } (\rho \rightarrow e) T \rrbracket &= \text{map } (\llbracket \rho \rrbracket \rightarrow \llbracket \Gamma, \rho \vdash e \rrbracket) \llbracket \Gamma \vdash T \rrbracket \\
&\quad \text{where } \Gamma(T) = \rho \text{ table and } \Gamma, \rho \vdash e : \rho' \\
\llbracket \Gamma \vdash \text{fold } (a : \tau, \rho \rightarrow e) a T \rrbracket &= \text{fold } (\llbracket a : \tau, \rho \rrbracket \rightarrow \llbracket \Gamma, a : \tau, \rho \vdash e \rrbracket) \llbracket \Gamma \vdash a \rrbracket \llbracket \Gamma \vdash T \rrbracket \\
&\quad \text{where } \Gamma(T) = \rho \text{ table and } \Gamma, a : \tau, \rho \vdash e : \tau \\
\llbracket \Gamma \vdash a_0 + \sum_{i=1}^n a_i * x_i \rrbracket &= \sum_{i=1}^n a_i * x_i \\
&\quad \text{let } p = \text{int } x \ast y \text{ in} \\
&\quad \text{let } d' = \text{num } x \ast y \text{ in} \\
&\quad \text{assert } 1 = (C_y)^y \ast_G g^{-p} \ast_G h^{-d'}; \\
\text{Commit } p \\
&\quad \text{when } x \text{ and } y \text{ private} \\
\llbracket \Gamma \vdash \text{lookup } x_0 T_0 \rrbracket &= \text{lookup } x_0 \ T_0 \text{ in} \\
\text{let } x_1, \ldots, x_n, e, v \in \text{A} = \text{lookup } x_0 \ T_0 \text{ in} \\
\text{let } d, \text{ od, } C_d = \text{Commit } \text{(random())} \text{ in} \\
\text{let } p = d \ast e \text{ in} \\
\text{let } d' = \text{ad } \ast e \text{ in} \\
&\quad \text{assert } 1 = g^C_d g^{-p} h^{-d'} \\
&\quad \text{let } A' = A \ast h^{-d} \text{ in} \\
&\quad \text{assert } e(Z, \tilde{g}) \tilde{e}(1/A', p_k) = G_t \\
&\quad (\prod_{i=0}^{n} e(R_i, \tilde{g})^\sigma) \tilde{e}(A', \tilde{g})^\tau \\
&\quad \tilde{e}(S, \tilde{g})^{\gamma} \tilde{e}(h, p_k)^{\delta} \\
\text{Commit } x_1, \ldots, \text{Commit } x_n \\
&\quad \text{where } \Gamma(T_i) = (x_i: \text{int})_{i=0,n} \text{looketable} \\
\llbracket \Gamma \vdash x \rrbracket &= \downarrow x \text{ when } x \text{ private} \quad (10)
\end{align*}
\]

Figure 5: Shared translation of typed source expressions

the translation, the prover will compute \(C_p\), pass it to the verifier, and extend its challenge computation with equation \(E_2\), whereas the verifier will receive some \(C_p\) and use it to check this equation. Note that the cryptographic overhead depends on the target level of privacy: given instead a source environment \(\theta\) declaring that \(x\) is also private, the same discriminant expression would involve representation proofs for two private products.

*Private lookups* are translated using proofs of knowledge of signatures. To enable this, data sources extend input tables \(T : \rho \text{ looketable}, \rho\) is of the form \(x_0 : \text{int}, \ldots, x_n : \text{int}\) into tables \(T' : (\rho, \sigma)\text{table}\) with a CL signature at the end of each row, as follows:
Although this pre-processing may be expensive for large tables, it can be amortized over many queries.

A lookup within a source query, such as the one from the `blur` query of §2, is translated to a proof of possession of a CL signature. For instance, let us translate the expression `lookup c F` in environment

\[ \rho = F : (\text{city} : \text{int}, \text{country} : \text{int}) \text{lookupable}, \ e : \text{int}. \]

The environment is first translated to

\[ \llbracket \rho \rrbracket = (\text{city} : \text{int}, \text{country} : \text{int}, \sigma) \text{table}, \ c : \text{int}, \ oc : c \text{ opening}, C_e : c \text{ oc commitment} \]

The lookup itself is translated (using rule 9) to

\[ \llbracket \Gamma \vdash \text{lookup } c \ F \rrbracket = \]

let country, e, v, A = lookup c F in
let d, od, C_d = Commit(random()) in
let p, o' = d + e, od + e in
assert \( \Gamma \vdash \text{lookup } c \ F \rrbracket \)
let A' = A * h^{-d} \in
assert \[ \delta(Z, \hat{e}) \cdot \hat{e}(1, A', p_k) = c \]
\[ \hat{e}(R_0, \hat{g}), \hat{e}(R_1, \hat{g})^{\text{country}}, \]
\[ \hat{e}(A', \hat{g})^{x} \cdot \hat{e}(S, \hat{g})^{y} \cdot \hat{e}(h, \hat{g})^{y} \cdot \hat{e}(h, p_k)^{y} ; \]
Commit country

This code first looks for a signed tuple \((\text{city}, \text{country}, e, v, A)\) in \(F\) such that \(c = \text{city}\) and retrieves the remaining elements; it then proves knowledge of this tuple, without revealing which tuple is used in the proof, by blinding the element \(A\) of the signature. (Note that this proof internally relies on a proof of multiplicity.)

**Iterators and Commited Tables** ZQL supports tables with mixed public and private columns, as well as iterators `map` and `fold`. To enable processing on their private contents, data sources extend tables with commitments and sign them. For instance, here is the code for the provider of the table of cities for the `blur` query.

\[ D_i \triangleq \chi, sk, T \rightarrow \text{map} (x_0 \ldots x_n \rightarrow \]

let \(e = \text{random}()\) in
let \(v = \text{random}()\) in
let \(A = (\prod_{i=0}^{n} R^{z} Z^{-1})^{\pi \pi} \) in
\[ T \]

This code first uses `map` to extend each source integer with a fresh opening and commitment, using the `Commit` abbreviation; this yield the extended table \(X'\) passed to the prover. It then uses `fold` to compute the joint hash of these commitments, and finally signs the result. (In the hash computation, \(H_0\) is some fixed tag, and we omit a conversion from `elhG` to `hash`). As outlined at the end of this section, both the prover and the verifier perform some initial processing for these extended tables: the prover must show his knowledge of the representation for these commitments, and the verifier must verify the signature and the representation proofs for these commitments.

We illustrate the translation of the `map` iterator (5) on the `blur` query from §2. The translation of `fold` (6) is similar. The map expression of `blur` is translated to another map expression, in a translated environment that provides the extended input \(X : [x : \text{int}] \text{table} : \)

\[ \llbracket \Gamma \vdash \text{map} (c \rightarrow \text{lookup } c \ F) \rrbracket = \]

\[ \text{map} (c, oc, C_e \rightarrow \llbracket \Gamma, c : \text{int } \vdash \text{lookup } c \ F \rrbracket) X \]

and the translation continues with the `lookup` expression, as explained above.

**Prover Translation** Continuing from the result of the shared translation, the prover translation uniformly turns its assertions into a custom non-interactive \(\Sigma\)-protocol, in two passes, written \([\ldots]_1\) and \([\ldots]_2\), that produce code first for the message randomness, then for the responses.

Figure 6 defines these two passes, as well as the top-level query translation \([\ldots]_1\) that combines \([\ldots]_1\) and \([\ldots]_2\) with additional glue. Overall, the prover for a source query \(\theta \rightarrow e\) is thus defined using this translation after the shared translation: \(P \triangleq \llbracket [\theta \vdash e] \rrbracket_\text{PROVER.} \)

**First-message translation** In the first pass, \(H\) is the public hash incrementally computing the global challenge; \(a\) is the accumulated cryptographic evidence that will be sent to the verifier; and every private variable \(x\) is replaced with a pair \(x_t, x_e\) where \(x_t\) is the proof randomness for \(x\). (Openings are treated as any other secrets.) In combination with the shared translation, every private source expression becomes a tuple of the form \([\ldots]_1 : \)

\[ \langle x, t_x, \alpha, t_{\alpha}, C_{\alpha} \rangle \]

where \(x\) is the value of \(e\), \(t_x\) is the randomness for \(x\), \(t_{\alpha}\) is an opening for \(x\), \(t_{\alpha}\) is the randomness for \(\alpha\), and \(C_{\alpha}\) is a commitment to \(x\). For efficiency, all these additional values are optional in our compiler.

Compositively, the type translation \([\rho]\) maps shared environments to environments extended with an entry for each proof randomness, and leaves the other entries unchanged; the expression translation \([\ldots]\_1\) takes \(H\) and \(a\) as free variables and returns their updated values of the form \(\text{extend} \ldots (\text{extend } E_{\alpha}) \ldots E_a\), with one exponential expression \(E_i\) for each assertion in \(e\), and \(a, a_1, \ldots, a_n\) for each additional evidence \(a_i\) produced by \(e\).

We explain the main cases of the first-pass translation. Public expressions are (still) unaffected. Note that they may includes public expressions generated by the shared translation, such as products of commitments. Affine private expressions are translated homomorphically, adding a corresponding linear expression on the proof randomness. Private exponential computations yields evidence that must be communicated to the verifier; we add their results to \(a\). More complex private expression are supplemented with the sampling of a fresh message randomness.
for their result—we rely on the assertions introduced by
the shared translation to prove those assertions.

Assertions of equations of the form \( \varphi = \epsilon \) are
transformed into extensions of the global-challenge computation.
The left-hand-side must be a public expression, and is discarded.
The right-hand-side must be an expression.

\( x \) obtained by replacing each of these variables
with \( t_x \). The translation computes it, and extends \( H \) with the
result. Desclassifications are similarly translated: the
declassified value \( x \) is added to \( a \), and the hash is extended
with \( g^x \) to link it to its proof randomness (as if we were
translating \textbf{assert} \( g^x \)). Continuing with our example,
we give below the expression \( e_1 \), obtained by translating
the shared-translation of the discriminant query, after
removing the unnecessary commitment \( C_d \). (This code has
been rearranged for simplicity; the full code produced by
the translation rules appears in the full paper.)

\begin{verbatim}
let \( p = z + z \) in
let \( t_p = \text{random()} \) in
let \( a' = \varphi' \) in
let \( t_{\varphi'} = \text{random()} \) in
let \( H = \text{extend } H (\{C_y,g \} \cdot g^{t_p} \cdot g^h \cdot \varphi') \) in
let \( \alpha \) = \( \text{sample()} \) in
let \( t_\alpha = \text{random()} \) in
let \( C_\alpha = g^\alpha \cdot \text{random}() \) in
let \( H = \text{extend } H (\alpha C_p) (g^\varphi \cdot g^h \cdot \alpha) \) in
let \( d \) = \( p - 4 \cdot x \cdot y \) in
let \( t_w = t_p - 4 \cdot \alpha \cdot t_x \) in
let \( H = \text{extend } H (g^\varphi) g^d \) in
let \( a = a, (p,t_p), (\alpha',t_\alpha), (\alpha, t_\alpha), (\alpha, t_\alpha, \alpha, \alpha) \) in
\end{verbatim}

\( (H, a, d) \)

**Response Translation** In the second pass, after completing
the computation of the global challenge \( c \), we revisit
the collected evidence \( a \), and we replace every pair of a
private value \( x \) and associated proof randomness \( t_x \)
with the response \( r_x = t_x - c \cdot x \). This pass is defined by
induction on the type of \( a \), produced by the first-message translation,
which indicates where those pairs are. (Technically, this pass also needs to re-balance nested tuples,
as the prover produces \((\ldots (a_0, a_1), a_2), \ldots, a_n)\) whereas
the verifier consumes \((a_0, (a_1, \ldots, a_n))\); we omit those details.)
Continuing with the discriminant prover, the result-
ing value \( a : \delta \) binds the series of variables

\( (z, t_z), (\varphi, t_\varphi), (p,t_p), (\alpha',t_\alpha), (\alpha, t_\alpha), \alpha, \alpha) \)

and thus \( [\delta]_2 \) simply computes the responses for the five
pairs of secret and associated proof randomness:

\begin{verbatim}
[\delta]_2 = \let (z,t_z), (\varphi, t_\varphi), (p,t_p), (\alpha',t_\alpha), (\alpha, t_\alpha), \alpha, \alpha) in
let \( r_z = t_z - c \cdot z \)
let \( r_\varphi = t_{\varphi'} - c \cdot \varphi \)
let \( r_p = t_p - c \cdot p \)
let \( r_{\alpha'} = t_{\alpha'} - c \cdot \alpha' \)
let \( r_{\alpha} = t_{\alpha} - c \cdot \alpha \)
(r_z, r_\varphi, r_p, r_{\alpha'}, r_{\alpha}, C_p, d)
\end{verbatim}

**Top-Level Prover Translation (P)** We arrive at the following
code for the prover, given here for the discriminant query. (See Figure 6 for the general case.)
This prover relies on data sources extending both private
source inputs \( y \) and \( z \) with an opening, a commitment, and a signature on that commitment

\( x, (y, \sigma_y, C_y, \sigma_y), (z, \sigma_z, C_z, \sigma_z) \to \)

\begin{verbatim}
let \( H = \text{extend } (\text{extend } H_0 C_z C_z) \) in
let \( t_z = \text{random()} \) in
let \( t_{\sigma_z} = \text{random()} \) in
let \( a = (z, t_z), (\varphi, t_{\varphi} ) \) in
let \( H = \text{extend } H (g^\varphi \cdot g^h \cdot \alpha) \) in
let \( \delta, d = [\theta]_2 \) / \( \text{phase 1 detailed above} \)
let \( c = \text{finalize } H \) in
let \( a = [\delta]_2 \) / \( \text{phase 2 detailed above} \)
\end{verbatim}

\( (x, (C_y, \sigma_y), (C_z, \sigma_z), c) \)

In this code, \( H_0 \) is the hash of all public values used as
bases in the \( \Sigma \)-protocol, \( [\theta]_2 \) is the tuple type of the
(extended) provided data, and \([\theta]_\text{pub} \) is an expression
that extracts their public parts (including the plain signatures,
excluding lookup tables). The type \( \delta \) of the additional
evidence depends on the first-pass of the translation, and is used to drive the second part. In-between, the
final value \( H : \text{hash} \) is finalized into the global challenge
\( c : \text{num} \). The last line assembles the message passed from
the prover to the verifier, which consists of (1) the public parts
of the input data and of the result; (2) the additional evidence
for proving this result; and (3) the global challenge for verifying this proof.

**Verifier Translation** Also following the shared translation,
the prover translation leaves the public parts of the
query unchanged, and it incrementally re-computes
the challenge using the responses and additional evidence
prepared by the prover for the private parts of the query.
Figure 7 gives the compositional translation applied
to the result of the shared translation, and the top-
level translation \([\cdot]\_\text{VERIFIER} \). In combination, the verifier
is defined as \( V = [\cdot]_2 \cdot [\cdot]_\text{VERIFIER} \).

**Compositional translation** \([\cdot]_v \) In the verification pass,
\( H \) is the public hash incrementally re-computing the
global challenge, \( a \) is the received evidence consumed
by the verifier, and every private variable \( x \) is replaced
with a (public) response variable \( r_x \)—the type translation
\([\rho]_v \) performs this replacement. In combination with
the shared translation, every private source expression now
yields a tuple of the form \( r_x, r_{\alpha}, C_x \) where
\( r_x \) and \( r_{\alpha} \) are (presumably) responses associated with the exponents
committed to \( C_x \). (Again, all these values are actually
optional in the compiler.)

The verifier expression \([\epsilon]_v \) takes free variables \( H \)
and \( a \), and additionally returns the updated \( H \) and the rest of \( a \). Public expressions are unchanged. Private
expressions are discarded, and replaced with response expressions,
either computed (for affine expressions) or read
off the evidence $a$ (for more complex expressions). Note that the translation of affine expressions includes a term $-c + a_0$ for the constant, to ensure that, given correct responses for its free variables, the translation of an expression also produces a correct response.

Assertions of equations of the form $e_p = e_s$ are translated to hash computations, by computing the expression $(e_p)^r = e_s$, where $e_r$ is obtained from $e_s$ by replacing every variable $x$ with $r_x$, and by extending $H$ with the result. Declassifications $\downarrow$ are similarly translated by reading $x$ off the evidence $a$ and extending the hash with $g^{x+c+e_x}$. For instance, continuing with the discriminant query, the (simplified) verifier translation $\llbracket [ \rho ] \mapsto [ e_a ] \rrbracket$, is

\[
\begin{align*}
\text{let } r_p, r_h, r_{ch}, C_p, d, a = a \text{ in} \\
\text{let } H = \text{extend } (H ((C_p)^{e_p} \ast G^{x+c} \ast G^{h+e_h})) \text{ in} \\
\text{let } H = \text{extend } \left( \text{extend } H C_p \right) ((C_p)^{e_p} \ast G^{x+c} \ast G^{h+e_h}) \text{ in} \\
\text{let } r_d = r_p - 4 + x + r_h \text{ in} \\
\text{let } H = \text{extend } \left( \text{extend } H g^{d^2} \right) ((g^{d^2})^x \ast G^{h+s}) \text{ in} \\
\left( H, a, d \right)
\end{align*}
\]

Top-Level Verifier We finally give below the top-level verifier translation, also for our sample discriminant query; see Figure 7 for additional details.

\[
\begin{align*}
x, C, a, c &\rightarrow \\
\text{verify } v_k C_a C_z; \\
\text{let } H = \text{extend } (H ((C_z)^{e_z} \ast G^{x+c} \ast G^{h+e_h})) \text{ in} \\
\text{let } H = \text{extend } H C_z \ast G^{x+c} \ast G^{h+e_h} \text{ in} \\
\text{let } H, a, d = \llbracket [ \theta ] \mapsto [ e_a ] \rrbracket_v \text{ in } H; \\
\text{check } c = \text{finalize } H.
\end{align*}
\]

The prover first verifies the signatures on the two received commitments for $y$ and $z$; it starts the challenge re-computation on the representation equation for input $z$ (since we need a response for $z$ an $\alpha z$ to check the proof of the square $z^2$), then proceeds with the verification for the query expression; it checks that the received and re-computed challenges match; it finally returns the public result $d$ (unless of course verify or check raised an error.)

6 Security Theorems

Consider a well-typed ZQL source query $Q \triangleq \theta \rightarrow_e e$, with $\ell$ input variables $\theta = (\chi_i : \tau_i)_{i=0,..,\ell-1}$, that declassifies only its result and its translation $(S, (K_i, D_i))_{i=0,..,\ell-1}, (P, V)$. We give our main results based on the definitions of $\ell$. We refer to the full paper for the proof outlines, and for a discussion of automated, type-based verification for the compiled protocols. For functional correctness and soundness, we also suppose that there is no source-program overflow—formally, integers and their operations are computed modulo $q$.

Theorem 1 (Functional Correctness)

$(S, (K_i, D_i))_{i=0,..,\ell-1}, (P, V)$ is correct.

Theorem 2 (Perfect Privacy)

$(S, (K_i, D_i))_{i=0,..,\ell-1}, (P, V)$ is $(t, 0)$-private.

Our soundness theorem below is in the random-oracle model, requiring that extend and finalize are independent random oracles. It assumes that the Discrete Logarithm (DL) and Strong Diffie Hellman (SDH) assumptions hold—to guarantee the security of commitments and CL signatures, respectively—and assuming that the $\ell_{CMA}$ conventional signatures primitives of data-sources are chosen message attack secure (CMA).

Theorem 3 (Computational Soundness)

$(S, (K_i, D_i))_{i=0,..,\ell-1}, (P, V)$ is $(t, \ell)$-sound, where the execution time $t$ and success probability $\epsilon$ are respectively lower- and upper-bounded by the corresponding parameters of the assumptions.

Concretely, let $t_{DL}$, $t_{SDH}$, $t_{CMA}$, and $\epsilon_{DL}$, $\epsilon_{SDH}$, $\epsilon_{CMA}$ be those parameters, for large enough bounds on the number of calls to their primitives. If $t < t_{CMA} - t_{red1}$, $t < (t_{DL} - t_{red2})/2$, and $t < (t_{SDH} - t_{red3})/2$, where the $t_{redi}$ are small constants, then $\epsilon < \epsilon_{CMA} \cdot Q \cdot \sqrt{t_{DL} + (\ell - t_{CMA}) \cdot \epsilon_{SDH} + Q^2/q}$, where $Q$ is the number of random oracle queries made by $\alpha$ and $q$ is the order of $G$ and thus also the size of the challenge.

In contrast with our privacy theorem, which is information-theoretic, our concrete-security soundness theorem is somewhat more cumbersome than the asymptotic security theorems often found in theoretical cryptography, but it remains closer to reality, in which cryptographic primitives come with concrete security bounds, and thus provides guidance for configuring these primitives to achieve adequate security.

7 ZQL applications

The expressivity of ZQL stems from the ease with which the primitive operators can be composed to build larger queries. We illustrate this by providing queries for applications in prior literature.

In the setting of smart metering, a meter issues signed private readings, and a household needs to compute their bill on the basis of a public tariff policy that maps each reading to a fee over time. A number of custom privacy protocols have been proposed to do this [55, 48]. One such billing policy takes a table of public times and private readings, and a household needs to compute their bill in the table $T$ using $\text{lookup}$ and sums the results.
Another popular application in the literature involves pay-as-you-drive insurance schemes. Such schemes require drivers to fit a black box in their car that records their driving habits, and allow the insurer to compute a premium based on the safety of the driving, as well as distance or time. The use of zero-knowledge protocols to support such automotive settings, including road usage billing and tolling has been well established in the literature [5, 61, 44].

An example policy used by a UK auto insurance pilot scheme involves recording the segment of road travelled, the distance and the speed and use those to subtract “points” from a virtual driving license. Points are linked to the magnitude of speed violations on the road segments travelled. The insurance rate per mile is then computed as a function of the points subtracted, up to a threshold where the insurance becomes invalid.

We can express such a policy in ZQL using a table for the recorded road segments used, and lookup tables to encode the speed limit of road segments, the penalty points per magnitude of violation, and finally the insurance premium for a certain number of points:

```zql
let pay_as_you_go
(Segments : (int * int * int * int) table)
(Limits : (int * int) lookuptable )
(Penalties : (int * int) lookuptable )
(Rates : (int * int) lookuptable )
let points =
  sum ((time, road, speed, miles) ->
    let limit = lookup road Limits
    lookup (speed − limit) Penalties) Segments
let rate = lookup points Rates
let miles =
  sum ((time, road, speed, miles) -> miles) Segments
  ↓ (miles * rate)
```

The **pay as you go** application makes extensive use of lookup tables to simulate traditional database half-joins between tables. The values of these tables are largely arbitrary and related to the insurance policy. We note that to fully secure this insurance mechanism, some information about the start and end times of the segments must also be signed by the black box and verified to avoid malicious replays or omissions. We also note that, depending on policies, the query leaks information from individual secret inputs to the computed premium. Securing against source query leakage is beyond the remit of ZQL, but could be achieved by adapting differentially private schemes [36].

The final example illustrates how ZQL lookups can be used to approximate real functions, including trigonometric functions and division which is not yet natively supported. The **hc** table has a large domain (~ 1 million items) but can be reused across multiple operations. Other tables have a relatively small domain related to the distances of the points compared.

8 Discussion

**Prototype implementation & limitations** Our compiler uses the language development and testing facilities of F#: we program source queries as (a small subset of) F#, then extract the ZQL abstract syntax tree (AST) through reflection. The compilation pipeline performs ZQL type-checking, applies the shared translation, and finally produces the data-source, prover and verifier code. Each of these steps operates on well-typed ZQL expressions. This enables us to share many optimizations as ZQL-to-ZQL transformations.

Besides standard optimizations, the compiler supports a more general variant of `lookup` primitive, named **find**, that returns any lookup-table row that meets a condition expressed as a boolean expression on the whole content...
of the row. This provides more flexibility on the use of lookup tables, but its compilation is more complex.

In addition to cryptographic code, ZQL also synthesizes a custom marshaller and un-marshaller for the cryptographic evidence and results of the query. Following the ZQL approach, this code is specialized and compiled for a specific proof. Hence, the size and location of all fields, parametrized on the input table lengths, in known at compile time and there is no need to rely on a general-purpose parser, a component that is traditionally a source of security flaws.

We support three distinct compiler back-ends:

Concrete F# The main branch of the compiler transforms and compiles the final ZQL data source, prover and verifier into F# code, linked to the standard .NET big integer libraries, or to proprietary managed libraries that support pairing based cryptography.

Symbolic F# The second branch of the compiler is linked against symbolic execution libraries for all the operators and primitives. Interestingly, since the F# branch makes extensive use of abstract types in the final prover and verifier, there is no need to write a separate symbolic execution environment: the mathematical functions can simply be replaced with equivalents computing on symbolic polynomials. The resulting code jointly computes the execution time and the proof size, as polynomial expressions of the input lengths and the unit costs of each cryptographic operation. We use symbolic execution to predict the performance of the compiler, and hope to use it in the future to choose between alternative optimization strategies at compile time.

Concrete C++ Finally, we support compilation of the verifier to native C++ code, linked with high performance native big integer libraries. This branch involves transforming the functional ZQL verifier and un-marshaller code into an imperative program and optimizing it using standard low-level techniques such as removing dead code, removing spurious copies, and minimizing memory re-allocations. The resulting native program takes a proof as an input, and outputs the verified result. The native branch does not support on-the-fly compilation and execution, and currently works for RSA groups only. Yet the resulting binary can be easily deployed where .NET runtimes are not available.

The process of compiling a query remains fast even on small devices. Thus, a service could simply send ZQL queries to the user, to be reviewed, compiled, then executed locally. To this end, our compiler also has an API that takes source ZQL ASTs, compiles them to F#, then also compiles and dynamically loads the resulting F# code. This is likely to be faster, cheaper, safer and more reliable than providing custom binaries every time the query is updated.

The prototype compiler is still subject to limitations. For instance, some optimizations, such as moving de-classifications up in the dataflow to minimize the size of the Σ-protocol, or batching some exponential computations, could be systematically applied.

Performance Evaluation Table 1 illustrates the performance of ZQL code for the three applications presented in Section 7. It provides the execution time for the F# provers and verifiers, as well as the size of the proof, for different security parameters of RSA (1024 bits, 2048 bits) and the pairing based cryptography over a 254 bits Barreto-Naehrig curve (BN254). The smart_meter_bill readings table is of size $l_{\text{read}} = 5$ and the pay_as_you_go query road segments table is of size $l_{\text{seg}} = 25$. This means that for the 1024 bit RSA branch, the prover can process a meter reading every $\sim 120mS$ or a segment of road every $\sim 360mS$. The proof size for the pairing based branch is $\sim 755$ bytes per reading and $\sim 1921$ bytes per segment. As expected, the pairing based proofs are more compact than their RSA counterparts for the same or even higher levels of security: a 254 bits curve provides about 128 bits of security which correspond to a 3072 bits RSA modulus. This is further aggravated by the lack of tightness in RSA-based security reductions [8]. Prover timings take into account the generation of random numbers. We note that these numbers, while slow by the standards of non-privacy friendly computation, are perfectly adequate for computing bills and insurance premiums in real time.

Besides the main F# backend we experimented with a C++ back-end that compiles to a native verifier. Although more performant in absolute terms, the native verifier is not significantly faster than its F# counterpart. The RSA 1024 bit computation of the pay_as_you_go verifier took 4.290mS as compared with the F# backend using native big integer binding that took 5.111mS. Profiling the C++ execution indicates that more than 90% of the time is spent inside the modular multiplication function performing exponentiations. Thus, improving the performance of ZQL comes down to either faster exponentiations (through batching, multi-exponentiation or hardware) or reducing the number of operations required through more aggressive simplification of the protocols.

Finally, Table 1 illustrates the output of the symbolic execution engine on these three applications, in a configuration that measures the number of exponentiations ($E$), pairings ($\hat{e}$), and signature verification operations ($\text{sig}$) in terms of the length of the input tables ($l_{\text{read}}$ and $l_{\text{seg}}$), and ignore all other costs.

Where next? The current ZQL language is subject to some intrinsic limitations, and we are actively exploring options to overcome them.

1http://www.cryptopp.com/wiki/Security_Level
Many of the limitations are cryptographic and could be overcome by applying more advanced protocols. For example, lookup and find are currently restricted to externally signed tables. Lookup tables based on accumulators [13] or vector commitments [27] would be more flexible and may reduce cost. At a lower level, table processing leads to many similar cryptographic operations in a data-parallel style. Batch proof and verification techniques and homomorphic signature schemes could speed them up [10]. Well known, zero-knowledge proofs for disjunctions, would allow ZQL branching statements. The shared translation could bundle multiple secrets per commitment. Alternatively one could also employ completely different low-level proof engines, e.g., [53]. We note that choosing automatically the best encoding and technique, as well as compiling them in a compositional manner are challenging open problems. For some preliminary work in this direction see [41].

On the language design side, we illustrated in §7 how functions can be approximated though lookups. ZQL could automate and optimize the process by compiling data sources that calculate and sign function-tables appropriately. Finally, by design, our source language shields programmers from cryptography, and this may hinder power-users that wish to customize our compilation scheme, or experiment with its variants. Similarly, some users may wish to rely on external, unverified procedures, and use ZQL only to validate their results. Advanced APIs exposing the internals of the ZQL compiler without breaking its invariants would help them.

Acknowledgments
The authors would like to thank Ian Goldberg for early discussions of languages for zero-knowledge proofs and the advantages of compilation versus interpretation, and Nikhil Swamy for his comments.

Table 1: Performance for our three applications: runtime, and communicated proof sizes. The smart_meter_bill readings table is of size \( \ell_{read} = 5 \), the pay_as_you_go query road segments table is of size \( \ell_{seg} = 25 \), the gps_distance is between two points.

<table>
<thead>
<tr>
<th>Examples (branch)</th>
<th>prover (mS)</th>
<th>verifier (mS)</th>
<th>proof size (Bytes)</th>
</tr>
</thead>
<tbody>
<tr>
<td>smart meter bill (1024)</td>
<td>586</td>
<td>599</td>
<td>6,106</td>
</tr>
<tr>
<td>smart meter bill (2048)</td>
<td>3,498</td>
<td>3,148</td>
<td>10,585</td>
</tr>
<tr>
<td>smart meter bill (BN254)</td>
<td>1,374</td>
<td>2,092</td>
<td>3,773</td>
</tr>
<tr>
<td>smart meter bill (symbolic)</td>
<td>( E + 16 \cdot E \cdot \ell_{read} + 6 \cdot \ell_{read} \cdot \hat{e} )</td>
<td>( 6 \cdot E + 14 \cdot E \cdot \ell_{read} + 8 \cdot \ell_{read} \cdot \hat{e} +</td>
<td>\hat{sgv}</td>
</tr>
<tr>
<td>pay as you go (1024)</td>
<td>5,314</td>
<td>5,111</td>
<td>57,368</td>
</tr>
<tr>
<td>pay as you go (2048)</td>
<td>32,442</td>
<td>30,859</td>
<td>100,099</td>
</tr>
<tr>
<td>pay as you go (BN254)</td>
<td>8,305</td>
<td>12,261</td>
<td>28,819</td>
</tr>
<tr>
<td>pay as you go (symbolic)</td>
<td>( 15 \cdot E + 40 \cdot E \cdot \ell_{seg} + 12 \cdot \ell_{seg} \cdot \hat{e} + 6 \cdot \hat{e} )</td>
<td>( 29 \cdot E + 36 \cdot E \cdot \ell_{seg} + 29 \cdot \ell_{seg} \cdot \hat{e} + 8 \cdot \hat{sgv} )</td>
<td>( 167 +</td>
</tr>
<tr>
<td>gps dist (1024)</td>
<td>501</td>
<td>529</td>
<td>5044</td>
</tr>
<tr>
<td>gps dist (2048)</td>
<td>3,017</td>
<td>2,889</td>
<td>8629</td>
</tr>
<tr>
<td>gps dist (BN254)</td>
<td>841</td>
<td>1,253</td>
<td>2751</td>
</tr>
<tr>
<td>gps dist (symbolic)</td>
<td>( 60 \cdot E + 18 \cdot \hat{e} )</td>
<td>( 71 \cdot E + 24 \cdot \hat{e} + 4 \cdot \hat{sgv} )</td>
<td>( 233 +</td>
</tr>
</tbody>
</table>

References

Towards ensuring client-side computation


First stage:

\[ \Gamma \vdash e \vdash H, a, e \] when \( e \) public expression, that is, whose variables are all public in \( \Gamma \).

\[ \Gamma \vdash e \vdash H, a, \underbrace{a_1 + \sum_{i=1}^n a_i \ast x_i}_v \] when the \( x_i \) are private and the \( a_i \) public:

\[ (\Gamma(a_i)) = \text{pub num}_{0 \ldots n}, (\Gamma(x_i)) = \text{num}_{0 \ldots n} \]

\[ \Gamma \vdash e \vdash H, a, \underbrace{a_1 \ast r_t}_v \] when \( \Gamma \vdash e ; \rho \) non-linear private expression (including assoc, random, opening...)

\[ \Gamma \vdash \text{assert } \underbrace{e_\rho = \text{extend } H, e_\rho, a, e}_v \] when \( e_\rho \) public and \( e_s \) algebraic on private exponents

\[ \Gamma \vdash \xi \vdash a, a, \underbrace{a \ast v}_v \] where \( x \) ranges over the private variables bound in \( \rho \)

Second stage:

\[ \xi \vdash e \vdash \text{Verifier} \] (see full paper for \text{map} and \text{fold})

Figure 6: Prover Translation (see full paper for \text{map} and \text{fold})

\[ \xi \vdash e \vdash \text{Verifier} \] (see full paper for \text{map} and \text{fold})