Reef: Fast Succinct Non-Interactive Zero-Knowledge Regex Proofs

Sebastian Angel⋆ Eleftherios Ioannidis⋆ Elizabeth Margolin⋆ Srinath Setty† Jess Woods⋆
⋆*University of Pennsylvania* †*Microsoft Research*

Abstract

This paper presents Reef, a system for generating publicly verifiable succinct non-interactive zero-knowledge proofs that a committed document matches or does not match a regular expression. We describe applications such as proving the strength of passwords, the provenance of email despite redactions, the validity of oblivious DNS queries, and the existence of mutations in DNA. Reef supports the Perl Compatible Regular Expression syntax, including wildcards, alternation, ranges, capture groups, Kleene star, negations, and lookarounds. Reef introduces a new type of automata, *Skipping Alternating Finite Automata* (SAFA), that skips irrelevant parts of a document when producing proofs without undermining soundness, and instantiates SAFA with a lookup argument. Our experimental evaluation confirms that Reef can generate proofs for documents with 32M characters; the proofs are small and cheap to verify (under a second).

1 Introduction

Regular expressions (regex) are used to represent and match patterns in text documents in a variety of applications: content moderation, input validation, firewalls, biology, and more. Existing use cases assume that the regex and the document are both readily available to the querier so they can match the regex on their own with standard algorithms. But what about situations where the document is actually held by someone else who does not wish to disclose to the querier anything about the document besides the fact that it matches or does not match a particular regex? While slightly unusual, the ability to prove such facts enables interesting new applications:

• *Proving strong passwords.* Asymmetric or Augmented Password Authenticated Key Exchange (aPAKE) [48, 75, 78, 83] allow clients to register and authenticate to a server without disclosing their password to the server. However, aPAKE protocols have no mechanism for the server to confirm that the client chose a "strong password". This feature is crucial in corporate settings where password policies help prevent account compromise. Clients could convince the server of this fact with a proof that their secret password satisfies a password strength regex chosen by the server (e.g., at least 10 alphanumeric and one special character).

• *Disclosing content with redactions.* DomainKeys Identified Email (DKIM) [46] is a protocol whereby a sending mail server signs the header and payload of an email so that recipients can verify its authenticity. Journalists use DKIM signatures to establish the veracity of leaked emails. It might often be desirable to release a redacted version of an email (e.g., an email without a name) while allowing the public to confirm, via DKIM, the authenticity of the redacted email. By creating a regex that expresses the public content of the email, with redactions being expressed as wildcards with Kleene star, it is possible to show that the redacted email is derived from an email whose DKIM signature verifies under the sending mail server's public key. A similar idea is that of selectively disclosing fields in JSON web tokens [50] or verifiable credentials [15, 62] by "redacting" all other entries.

• *ODoH blocklisting. Oblivious DNS over HTTPS* [54] allow clients to obtain a domain's IP address without revealing which domain they are accessing. This technology improves privacy for users, but network administrators within organizations lose the ability to block certain sites (e.g., known malware domains) as they can no longer see which domains users query. One can reintroduce this functionality by asking clients to generate ZKPs showing that their DNS queries do not match a set of forbidden regexes before those packets are allowed through to the ODoH proxy. The same idea applies more generally to TLS traffic through middleboxes [45].

• *Proofs about genes.* DNA is used to establish ancestry or the presence of particular mutations. If sequencing companies (e.g., 23andme) were to provide users a signed commitment to their sequenced genome, users would be able to prove properties of their DNA (expressed as a regex) without having to disclose it in full. For instance, users could prove the presence of a certain genetic mutation when they order personalized medicine online or sign up to clinical trials.

In theory, the above applications can be designed with some suitable combination of encryption, commitments, signatures, and zero-knowledge proofs. In practice, creating efficient proofs over arbitrary unstructured text is far from trivial.

This is precisely the problem we tackle with *Reef*, a compiler and runtime system that allows an entity to *commit to a secret document and then subsequently prove that the document matches or does not match one or more public regexes without revealing anything else about the document.* Building Reef requires answering the following research questions:

1. How should one commit to a text document D?
2. Given a commitment to a document D, how can one *arithmetize* (i.e., express as some type of circuit) the statements "D matches/does not match a regex $\mathcal{R}$"?

3. What regex features are needed to enable realistic applications (e.g., quantifiers, alternation, lookarounds, etc.) and what is the best way to arithmetize these features?
4. What kind of zero-knowledge proof systems work well with the chosen commitment and arithmetization schemes?

To answer these questions, Reef marries new theoretical ideas and low-level techniques into a compiler that automatically arithmetizes arbitrary regexes. In particular, Reef:

Exploits NP checkers. Reef uses the common observation that checking the answer of a computation is often cheaper than finding the answer in the first place (either asymptotically or concretely). As a result, Reef does not arithmetize algorithms for finding regex matches/non-matches (e.g., Thompson's NFA [79], recursive backtracking). Instead, the prover in Reef computes the answer (i.e., finds the match and the relevant locations within the document, or establishes that there is no match) with a fast regex engine we built, and then proves that this answer satisfies criteria that implies the document has a match (or no match). Only this *NP checker* needs to be arithmetized and proven with a ZK proof system.

Reef's NP checker supports a wider class of regexes than all prior works, while also producing smaller arithmetizations. In particular, some works [16, 37] transform the regex into a DFA or NFA, and then prove that if one feeds the *entire* document into the automaton the final state is accepting/non-accepting. This approach results in $\mathcal{O}(|\mathsf{D}| \cdot |Q_{DFA}| \cdot |\Sigma|)$ constraints (or gates in some arithmetic or boolean circuit) to prove that there is a match, where D is the secret document, Q_{DFA} is the set of states in the DFA, and Σ is the alphabet. Three recent proposals, ZK-Regex [61], Zombie [84], and `zkreg` [69] reduce these costs: ZK-Regex and Zombie leverage Thompson's NFA (TNFA) and produce $\mathcal{O}(|\mathsf{D}| \cdot |Q_{TNFA}|)$ constraints, while `zkreg`'s use of Aho-Corasick DFA (ADFA) leads to $\mathcal{O}(|\mathsf{D}| + |Q_{ADFA}|)$ constraints.

Reef's NP checker is fundamentally different from the above approaches: it does not require feeding the entire document into the automata, only the relevant characters. This allows the prover to skip vast amounts of unnecessary work.

Introduces skipping automata. Above we allude to the idea of "skipping" irrelevant characters whenever possible. But how do we rigorously define this notion and what does "whenever possible" mean? To answer these questions, we introduce a new type of finite automata for regexes that we call *Skipping Alternating Finite Automata* (SAFA). SAFA generalize NFA to include the ability to change the *cursor* (i.e., the index of the next character to read in the input) following certain rules. SAFA allow Reef's prover to skip processing entire chunks of a document when the regex contains wildcard ranges such as "`.*`" or "`.{4,100}`" and let Reef handle *lookarounds*, which are common in password strength regexes and which no prior work supports.

Compared to prior works, Reef's NP checker can be expressed in $\mathcal{O}(\alpha \log(|\mathsf{D}| + |Q_{SAFA}| \cdot |\Sigma|))$ constraints, where $|Q_{SAFA}| \leq |Q_{TNFA}| \leq |Q_{ADFA}|$ and $\alpha = \mathcal{O}(|\mathsf{D}| \cdot L)$, where L is the number of lookarounds in the regex. There are two points worth emphasizing about the complexity of Reef's checker. First, SAFA have exponentially fewer states than TNFA and ADFA for many common regexes (§3.2). Second, α is much smaller than the above worst-case upper bound whenever Reef can skip characters. For instance, if $\Sigma = \{a, b, c\}$, the regex $\mathcal{R}$ = "`a.*b`" (meaning D has "a" eventually followed by "b") results in $\alpha = 2$ regardless of the size of D because Reef can skip all the wildcard characters. In contrast, $\mathcal{R}$ = "`^[a-b]*$`" (meaning D can contain any number of "a" or "b" characters but no "c") results in $\alpha = |\mathsf{D}|$ because we fundamentally have to check every character in the document to make sure it is not "c".

Leverages recursion. We observe that Reef's NP checker essentially performs the same high-level operations (looking up a character in the document and then transitioning to a new state) over and over. Such repeated structure is suitable for *recursive* zkSNARKs such as Nova [57], where the prover establishes that it ran some *step function* F, each time on a different input, until some terminating condition holds. Reef's termination condition is designed to allow the prover to safely stop proving as soon as the SAFA reaches an accepting state and the cursor points to the last character. This frees the prover from having to process the entire document (since in many SAFA the prover can skip to the last character without changing states) while hiding how many times F executes.

Commits to the document. Before Reef can be used, the document D needs to be committed in a form that allows Reef's NP checker to cheaply read arbitrary entries in D. Who generates the commitment depends on the application. In the gene example, the commitment is generated and signed by a trusted party (23andme). In the other applications, the commitment is generated by the user who must also supply a proof that ties the underlying document to the data in the application (e.g., the DKIM signature).

Reef uses a *polynomial commitment* [18, 27, 42, 59, 81, 85] for multilinear polynomials to commit to D, and a *lookup argument* [56] compatible with recursive proof systems. A lookup argument is a cryptographic protocol for proving that some entry exists in a public or committed table (polynomial) without revealing the entry. When the lookup argument is integrated into the step function F, it allows F to access any entries in D without revealing them to the verifier.

Supports table projections. Reef modifies the `nlookup` argument [56] to support lookup operations on *table projections*. Given a commitment to a table such as the document D, a projection is a smaller table D_{proj} derived from one or more contiguous chunks of D (the choice of which chunks of D are projected is public information). Reef then runs `nlookup` on D_{proj}, which incurs costs that are proportional to $|\mathsf{D}_{proj}|$. The verifier can still check that all lookups to D_{proj} were done correctly by using the original commitment to D.

$r, s ::= \alpha$	$\alpha \in \Sigma$
| `^` / `$`	document start / end
| `.`	wildcard character
| rs	concatenation
| $r \mid s$	alternation
| $r?$ / $r*$ / $r+$	quantifiers
| $[\alpha_i - \alpha_j]$	character classes
| $[\hat{}\alpha_i \ldots \alpha_j]$	negation of characters $\alpha_i \ldots \alpha_j$
| `r{m}` / `r{m,}` / `r{m,n}`	repetition ranges
| `(?=r)` / `(?<=r)`	lookahead / lookbehind

FIGURE 1—Reef supports the entire PCRE syntax [7] except for backreferences and subroutine references.

Table projections are a powerful construct in Reef and might be of independent interest. For example, a DNA chromosome results in a document D with tens of millions of entries. However, regexes on DNA usually have the form: $\mathcal{R}$ = "`.{1000}TT(T|C).{5000}CT(T|C|A|G).*`", which says that the first thousand entries are irrelevant, but right after we should see *TTT* or *TTC*, and 5000 entries later we should see *CTT*, *CTC*, *CTA*, or *CTG*; beyond that is irrelevant. SAFA allows Reef to skip all the irrelevant entries. However, in each step of the recursive proof system, `nlookup` internally invokes the *sum-check protocol* [60] which incurs costs linear in |D| (millions of entries) in order to prove the table accesses. With projections, `nlookup` runs the sum-check protocol over D_{proj} (under 10 entries).

Combines private and public tables. To efficiently express the state transitions and complex skipping rules in SAFA, Reef again uses a lookup argument. In particular, Reef stores SAFA's states, transitions, and skipping rules in a public table that both the prover and the verifier can derive from the regex. Given this table, the prover can, with one lookup, prove that it transitioned to the next state in the SAFA and advanced the cursor following the prescribed rules.

Having both a private table and a separate public table is undesirable because lookup arguments amortize their costs over many lookups (i.e., the more lookups to a table, the cheaper the per-lookup cost). If one has two tables, then queries to one table do not apply towards the amortization of queries to the other table. To remedy this situation, Reef shows how to combine both private and public tables into a single *hybrid* table (without leaking the contents of the private table) so that all lookups can be done on this combined table, improving amortization and eliminating repeated fixed costs.

We evaluate Reef on the applications described earlier and find that it can generate small proofs (tens of KB) in a few seconds, even for large documents such as DNA chromosomes.

2 Background

This section reviews regex matching, rank-1 constraint satisfiability (R1CS), NP checkers, and zero knowledge succinct non-interactive arguments of knowledge (zkSNARKs).

2.1 Regular Expression Matching

Given an alphabet Σ, a regex $\mathcal{R}$ is a pattern matching a set of strings, called the *language* of $\mathcal{R}$ or $\mathcal{L}[\![R]\!] \subseteq \Sigma^*$. Figure 1 outlines the basic syntax for the creation and combination of regexes that Reef supports.

Regexes are converted to *deterministic finite automata* (DFA) with known techniques [20, 26, 43, 51, 65, 77]. One can determine if a document matches a regex $\mathcal{R}$ by starting with the initial state and transitioning states on each character of the document until reaching a final state. If the final state an accepting states in the DFA, the document matches $\mathcal{R}$.

A common extension to regexes that Reef supports is *lookarounds* (e.g., positive or negative lookaheads and lookbehinds), a way to only match a pattern if is lead (or followed) by another pattern. For example, a password strength regex with two lookaheads might look like `^(?=.*[A-Z])(?=.*[!@#$&^*]).{10,}`, meaning it contains an upper case letter (`[A-Z]`), a special character from {`!`,`@`,`#`,`$`,`&`,`^`,`*`}, and has length at least 10 characters. The way to think about a lookaround such as "`^(?=`$\mathcal{R}$`)`" for some regex $\mathcal{R}$ is that $\mathcal{R}$ should be matched against the input string in the usual way, but once the match has been found, the *cursor* (i.e., the next position to process in the input string) should be reset back to what it was before the lookaround was processed. DFA/NFA have no notion of "resetting the cursor" and hence must simulate it by increasing the number of states exponentially [36].

2.2 zkSNARKs

A *zero-knowledge succinct non-interactive argument of knowledge* (zkSNARK) is a cryptographic protocol where a prover $\mathcal{P}$, convinces a verifier $\mathcal{V}$, that it knows a satisfying witness to some NP statement without revealing the witness. zkSNARKs typically target some variant of the NP complete problem of *circuit satisfiability* (e.g., R1CS [41, 72], Plonkish [40], AIR [21], CCS [73]), as one can represent arbitrary computations in this form. Informally, zkSNARKs are:

1. **Zero-knowledge:** The proof reveals no information to $\mathcal{V}$ beyond the fact that $\mathcal{P}$ knows a satisfying witness.
2. **Succinct:** The size of the proof and its verification is sublinear in the size of the satisfiability instance.
3. **Non-interactive:** No interaction between $\mathcal{P}$ and $\mathcal{V}$ besides the transferring of the computation's output and proof.
4. **Argument of knowledge:** $\mathcal{P}$ must convince $\mathcal{V}$ that it knows a witness that satisfies the instance. This argument is complete and computationally sound.

- *Perfect completeness:* If $\mathcal{P}$ knows a satisfying witness, $\mathcal{P}$ can always generate a proof that convinces $\mathcal{V}$.
- *Knowledge Soundness:* If $\mathcal{P}$ does not know a satisfying witness, it cannot produce a proof that $\mathcal{V}$ will accept, except with negligible probability.

2.3 Rank-1 Constraint Satisfiability (R1CS)

We focus on *rank-1 constraint satisfiability* (R1CS) as this is the arithmetization supported by the particular implementation of the zkSNARK we use [57], but all of our ideas apply to more general arithmetizations (e.g., CCS [73]). R1CS generalizes arithmetic circuit satisfiability, and an R1CS instance is given by a tuple $(\mathbb{F}, A, B, C, io, rows, cols)$, where $\mathbb{F}$ is a finite field, io is the public input and output of the instance, $A, B, C \in \mathbb{F}^{rows \times cols}$ are matrices, and $cols \geq |io| + 1$. The instance is satisfiable if and only if there exists a witness $w \in \mathbb{F}^{cols-|io|-1}$ that makes up a solution vector $z = (io, 1, w)$ such that $(A \cdot z) \circ (B \cdot z) = (C \cdot z)$, where $\cdot$ is the matrix-vector product and $\circ$ is the Hadamard product. The entry of z fixed at 1 allows constants to be encoded.

R1CS Arithmetization. Here we briefly explain how to turn a simple program into R1CS. Other works [19, 72, 74] have more complex examples. Suppose that $\mathcal{P}$ holds two elements $x_0, x_1 \in \mathbb{F}$ and wishes to convince $\mathcal{V}$ that y is the output of the following computation without leaking anything about x_0 or x_1 beyond what is implied by the result.

```
field foo(field x0, field x1) {
    field y;
    if (x0 == 30) { y = x1; } else { y = x0/x1; }
    return y;
}
```

To do so, we first express this function as a set of *constraints* (or equations) over elements in $\mathbb{F}$ that contain additions, subtractions, multiplications by constants, and at most one multiplication between variables. The result is:

$$\begin{aligned}
guard \times (x_0 - 30) &= 0 \\
guard \times (y - x_1) &= 0 \\
(1 - guard) \times (y - prod) &= 0 \\
x_0 \times inv - prod &= 0 \\
x_1 \times inv - 1 &= 0
\end{aligned}$$

To see why this represents the original computation, observe that we introduced auxiliary variables called *guard*, *inv*, and *prod*. Here, $\mathcal{P}$ is allowed to assign any values it wishes to y and the auxiliary variables, but let us assume that $\mathcal{P}$ provides the right values for x_0 and x_1 (this is usually enforced through the use of commitments). The only way that all six constraints are simultaneously satisfied is when: (1) $x_0 = 30$, $y = x_1$, and $guard = 1$ (there are many suitable values for the remaining variables); or (2) $x_0 \neq 30$, $guard = 0$, $y = prod = x_0 \times inv$, and $inv = x_1^{-1}$. As a result, if $\mathcal{P}$ claims that the output is y, and $\mathcal{P}$ can convince $\mathcal{V}$ that it knows a satisfying assignment for variables in the constraints given y, then $\mathcal{V}$ is assured that y is correct.

Appendix F shows how to convert these constraints into matrices A, B, and C. The solution vector z is $(y, 1, w)$, where $w = (x_0, x_1, guard, prod, inv)$ is $\mathcal{P}$'s secret witness.

2.4 NP checkers

While the above example is relatively simple it employs some clever tricks. In particular, it leverages *non-determinism* to transform expensive computations (branches and inverses) into cheap checkers that merely confirm the answers. For instance, if $\mathbb{F} = \mathbb{Z}_p$, *computing* $1/x$ with only additions and multiplications requires $\log(p)$ constraints via Fermat's little theorem (basically computing x^{p-2}). But in R1CS, we can just ask $\mathcal{P}$ to supply the inverse of x, *inv*, and simply *check* that *inv* is indeed the multiplicative inverse of x with one constraint: "$inv \times x - 1 = 0$". This is an example of an *NP checker*. There are many others used in SNARKs [19, 25, 49, 74, 82, 86].

In this work, we construct a novel NP checker for regex matching/non-matching based on a new type of automata.

3 Goals and standard approach

In Reef there are three parties: a committer $\mathcal{G}$, a prover $\mathcal{P}$, and a verifier $\mathcal{V}$ (in many cases $\mathcal{G}$ and $\mathcal{P}$ are the same entity). $\mathcal{G}$ generates a commitment *comm* for document D using random blind r, and provides $(comm, \mathsf{D}, r)$ to $\mathcal{P}$, and *comm* to $\mathcal{V}$. Later, $\mathcal{P}$ wishes to prove that D either does or does not match a regex $\mathcal{R}$ that is public and known to both $\mathcal{P}$ and $\mathcal{V}$. Given this setting, Reef has the following goals:

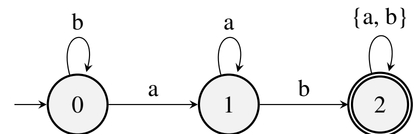
- **Completeness, Soundness, Succinctness, ZK**. These are analogs of the definitions given for zkSNARKs (§2.2) for the concrete R1CS instance that represents the statement "I know an opening of *comm*, and it matches $\mathcal{R}$" (or not).
- **Public verifiability**: The proof should be verifiable by anyone who has a commitment of the document and $\mathcal{R}$.
- **Expressiveness**: Reef should be able to support any regex written in PCRE syntax [7].

Additionally, our implementation of Reef achieves the following goal, though some settings might not need this and could use more efficient cryptographic primitives.

- **Transparency**: All cryptographic parameters for Reef should be generated without requiring a trusted setup.

3.1 A standard approach

As mentioned in Section 2.1, one can convert a regex into a DFA and then arithmetize its transition function δ. It boils down to a chain of if statements that takes as input the current state and current character in the document (both represented as field elements) and outputs the next state. For example, if the alphabet is $\Sigma = \{a, b\}$, and the regex is $\mathcal{R} =$ "`a+b.*`", the corresponding DFA would be:



Assuming that "a" maps to the field element 0, and "b" to 1, the corresponding δ transition is given by:

```
field delta(field state, field cur_char) {
    if (state == 0 && cur_char == 0) return 1;
    if (state == 0 && cur_char == 1) return 0;
    if (state == 1 && cur_char == 0) return 1;
    if (state == 1 && cur_char == 1) return 2;
    if (state == 2 && cur_char == 0) return 2;
    if (state == 2 && cur_char == 1) return 2;
    return -1; // invalid state or character
}
```

To express the computation of finding whether a committed document matches the regex, one would then: (1) open the commitment to obtain the document (an array of field elements); (2) call δ once for every character in the document in order; and (3) add a check at the end to see if the final state is one of the accepting states (another chain of if statements). The resulting `match` function is:

```
field match(field commit, field blind) {
  // commit is public input, blind is secret
  field[SIZE] document = open(commit, blind);
  field state = 0; // initial state

  for (i = 0; i < SIZE; i++) {
    state = delta(state, document[i]);
  }

  if (state == 2) { // accepting state in example
    return 1; // match
  } else {
    return 0; // no match
  }
}
```

One would then arithmetize this `match` function like in the example in Section 2.3. Indeed, this what some prior works do [16, 37]. Two recent works [61, 84] improve upon this design by converting the regex to a Thompson NFA (TNFA) [79] and performing additional optimizations.

3.2 Limitations of the standard approach

The previous standard approach has many drawbacks. We list the most salient ones here.

Insufficient Regex Expressiveness. Directly arithmetizing traditional finite state machines such as DFA, TNFA or Aho-Corasick DFA (AC-DFA) [17] fails to meet Reef's expressiveness goals. The most recent works in this area lack support for several common regex features.

For example, Zombie [84] lacks support for lookarounds. ZK-Regex [61] does not handle lookarounds, negations in character classes such as "`a[^[:space:]b`", or negations of entire matches (i.e., proving a non-match). Finally, `zkreg` [69], which is based on AC-DFA, only supports matching on a fixed set of strings. Unbounded repetition such as "`ab*c`" is unsupported, and negation of character classes, negation of entire matches, or wildcard ranges such as "`a.{100}b`" lead to an exponential number of states (2^{100}).

Poor scalability. The number of R1CS constraints produced by the standard approach for proving that a document D matches is $\mathcal{O}(|\mathsf{D}|\cdot|Q_{DFA}|\cdot|\Sigma|)$, where $|Q_{DFA}|$ is the number of states of the corresponding DFA. Zombie [84] improves this to $\mathcal{O}(|\mathsf{D}|\cdot|Q_{TNFA}|)$. But for applications where the document is millions of characters this still results in *billions of constraints*, even when the regex is small. In contrast, Reef's NP checker—based on SAFA (§5)—has $\mathcal{O}(\alpha\log(|\mathsf{D}|+|Q_{SAFA}|\cdot|\Sigma|))$ constraints, where $|Q_{SAFA}| \leq |Q_{TNFA}|$. As we discuss in Section 6.2, in the worst case $\alpha = \mathcal{O}(|\mathsf{D}|\cdot L)$, where L is the number of lookarounds in the regex; but in practice α is small (under 100 for even our largest document).

4 Improving the standard approach

One way to improve on the standard approach is to observe that the `match` function is well suited for a recursive proof system (this observation has been made many times in the context of other state machines such as blockchain rollups). In a *recursive zkSNARK* [22–24, 28–30, 55, 56], instead of arithmetizing the entire `match` function, we arithmetize one *step* of it. The result is:

```
field[3] match_step(field[] commit, field[] blind,
    field state, field cursor) {

  field cur_char = open_at(commit, blind, cursor);

  // accepting state and end of document (EOD)
  if (cur_char == EOD && state == 2) {
    return {0, 0, 1}; // match
  }

  state = delta(state, cur_char);
  return {state, cursor + 1, 0}; // not yet
}
```

The above `match_step` function takes as input a public *polynomial* [18, 27, 42, 52, 59, 81, 85] or *vector* [64] commitment (which could consist of multiple field elements) and the corresponding secret blind(s). These types of commitments have the nice property that they allow opening a particular entry within the commitment rather than having to open the entire document at once. `match_step` additionally takes the current state and the current cursor. If the current state is accepting and the cursor points to the end of D ("\$" in PCRE syntax, denoted by a special field element that the committer $\mathcal{G}$ appends to D to mark the end), D is a match and the return value is [0, 0, 1]. Else, `match_step` executes the DFA's δ function and returns the tuple [`state`, $cursor + 1$, 0].

A prover $\mathcal{P}$ in a recursive zkSNARK would then take the R1CS instance representing the `match_step` function, and produce a proof π_0 that establishes that running `match_step` correctly on a public commitment, private blinds, $state = 0$, and $cursor = 0$, produces the output *out*. Of course, proving a single step is not very useful (we could have done this without recursion); the key benefit is that a recursive proof system allows $\mathcal{P}$ to prove that it verified a prior proof (π_0 in this context) in addition to proving another `match_step` on the same public commitment, but the state and cursor returned

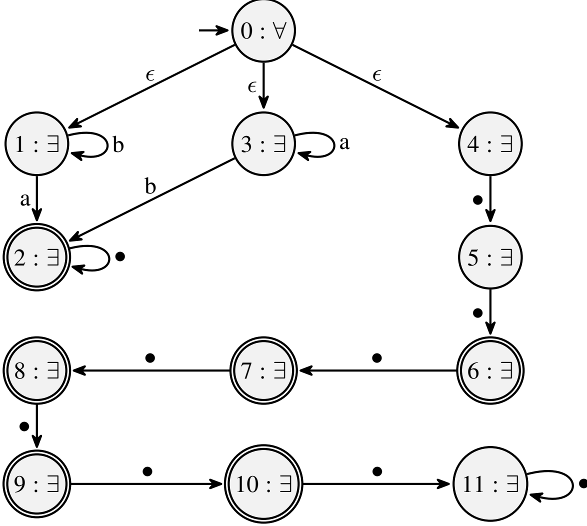


FIGURE 2—AFA for regex $\mathcal{R}$ = `^(?=.*a)(?=.*b).{2,6}$`.

by the prior step (*out*) which are bound by π_0. In this way, $\mathcal{P}$ can prove that, starting with $state = 0$ and $cursor = 0$, if $\mathcal{P}$ runs `match_step` *some* number of times, eventually $out = [0,0,1]$. The verifier $\mathcal{V}$ only learns this final value of *out* (and none of the intermediate values), in addition to a proof π_{final} that establishes that $\mathcal{P}$ checked all prior proofs and the last step was executed correctly.

This approach has four benefits. First, there is no need to unroll the loop and therefore the number of R1CS constraints is no longer fundamentally tied to the size of the document. This enables the second benefit: $\mathcal{P}$ can stop proving as soon as `match_step` outputs [0, 0, 1]. While in the construction presented so far $\mathcal{P}$ can only "stop" once it has gone through the entire document sequentially (so as to reach the EOD special character), Reef has the ability to skip many characters (possibly all the way to the end)—allowing the prover to stop without accessing the entire document. Third, breaking up the proof into small steps means that $\mathcal{P}$ can work on one step at a time, significantly reducing the amount of memory needed. Last, with recursive zkSNARKs like Nova [57], if $\mathcal{P}$ wants to prove the *same* step function many times (which is the case with `match_step`), there are significant performance gains.

5 Skipping Alternating Finite Automata

The use of recursion is a necessary first step in Reef, but it still falls short of our goals of expressiveness and efficiency.

In this section we introduce a new type of finite automaton called SAFA. The motivation for SAFA is twofold; avoid the state explosion problem for regex with lookarounds (§2.1) and capture the smallest set of characters within a document that must be checked in order to confirm that it matches a regex. We start by reviewing *Alternating Finite Automata* (AFA) which are a generalization of NFA. SAFA extend AFA to include the notion of *skips*.

5.1 Alternating Finite Automata (AFA)

AFA [32] are finite automata that generalize NFA by labeling states with an existential (∃) or a universal (∀) quantifier. An ∃ state is identical to a state in an NFA; the AFA merely reads the character at the current cursor, advances the cursor, and then transitions to any one of its possible next states. A ∀ state is very different. First, the AFA creates a *copy* of the remaining characters in the input string (starting at the current cursor until the end of the string) for *each* of its transitions (i.e., if there are 10 transitions it will create 10 copies of the input string). Then, in parallel, it transitions to every next state, and feeds each of those states their own independent copy of the input. For the AFA to accept an input string, all of the parallel branches need to end in accepting states. Intuitively, ∀ states capture the conjunction of multiple sub-automata, each of which operates independently on the provided input.

Formally, an AFA [32] is a 6-tuple $(Q, \Sigma, q_0, \lambda_q, \delta, F)$, where Q is the set of all states; Σ is the alphabet; $q_0 \in Q$ is the initial state; $\lambda_q : Q \to \{\forall, \exists\}$ is a labeling that assigns each state q either ∀ or ∃; $\delta \subseteq Q \times \Sigma \times Q$ is a transition relation that defines final states with respect to initial states and input characters; $F \subseteq Q$ is the set of accepting states.

Example. Suppose we want to match documents of length between 2–6 that contain "a" and "b" defined over $\Sigma = \{a,b,c\}$. This is given by the regex $\mathcal{R}$ = "`^(?=.*a)(?=.*b).{2,6}$`". Representing $\mathcal{R}$ as an NFA requires creating an automaton that accepts the alternation of all strings that contain both "a" and "b" and have length between 2 to 6 ("`ab`", "`.ab`", "`a..b`", "`.a.b.`", etc.). The minimal NFA for this has 17 states (the 16 shown here [10] plus a sink state for all invalid characters). In contrast, one can match $\mathcal{R}$ with the 11-state AFA given in Figure 2.

To understand this AFA, first recall *epsilon transitions*, which AFA inherit from NFA and which mean that the automaton can take any transition with an ϵ label without advancing the cursor or reading any character from the document. Second, notice the state at the top is labeled ∀, which means that after processing the document, all of its transitions (the 3 vertical branches) should end in an accepting state. The transitions of ∀ states are special in that each creates a private copy of the cursor initialized to the value of the cursor when the ∀ state is reached. As a result, states 1, 3, and 5 will all have their own cursors (i.e., advancing the cursor of the left branch does not affect the cursor of the right branch).

Consider for example the document $\mathsf{D} = acbcc$ which is accepted since the three branches out of state 0 run in parallel and each branch terminates in an accepting state. If instead $\mathsf{D} = bccbb$, the middle and right branches both terminate in accepting states, but the left branch does not.

The above example immediately shows that AFA could provide savings over the automata considered by prior works. Indeed, if a regex requires n states to be represented in an AFA, the same regex may require 2^{2^n} states in a DFA [36].

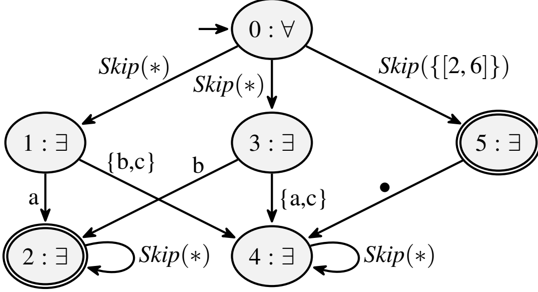


FIGURE 3—SAFA for regex $\mathcal{R}=$ `^(?=.*a)(?=.*b).{2,6}$` over alphabet $\Sigma = \{a, b, c\}$. *Skip*$(*)$ means the skip $\{[0,+\infty)\}$.

5.2 SAFA: Supporting Skips

AFA are a great way for Reef to increase the expressiveness of the supported regexes without incurring exponential costs, but AFA—just like DFAs and NFAs—are designed from the lens of "computation" rather than the lens of "verification". This fundamental distinction between compute and checking leaves a lot of opportunities unexplored.

As a concrete example, consider the regex $\mathcal{R} =$"`.*ab$`" and the document D ="`aaab`". AFA (much like NFA) represent "`.*`" by a single, non-accepting state, with the option to loop or progress forward with an ϵ transition. Finding the solution to the question "is $\mathsf{D} \in \mathcal{L}[\![\mathcal{R}]\!]$"? (meaning is D in the language defined by $\mathcal{R}$) requires computing both the case in which the first "a" in D matches the "`.*`" in $\mathcal{R}$ *and* the case in which it matches the "a" in $\mathcal{R}$. Confirming a match is simpler: given a path through the AFA for D, we just need to check that the path leads to an accepting state.

We can even take this concept further. When computing, bounded wildcard matching has to be explicitly unrolled. "`.{m,n}`", "`.{n}`", and "`.{n,}`" all require at least *n* transitions in an NFA or AFA. We see this in the right branch of the AFA in Figure 2 (states 4 through 10), where each state in "`.{2,6}`" has to be included explicitly.

But when checking, what if we could simply move the cursor forward by a number between 2–6 (inclusive), and carry on? Since "`.{2,6}`" is a wildcard, the content does not matter; what matters is that a *wildcard region* of the appropriate length exists. To express wildcard regions, we introduce *skips*. A skip is a finite set of non-overlaping intervals, $s = \{i_1, \ldots, i_n\}$, where each interval is of the form $i = [start, end]$ or $i = [start, \infty)$. Both *start* and *end* are non-negative integers and $start \leq end$; for $[start, \infty)$, the interval is unbounded on the right.

The idea is that when we reach a state that has a *skip transition* defined by some skip *s*, instead of reading a character from the input and transitioning to the next state based on the read value, the automaton advances the cursor by any amount within the intervals in *s*, and then moves to the next state.

Note that we need *s* to be a set rather than a single interval because of regexes such as "`(.{2,6}|.{8,10})a`" that have multiple acceptable disjoint wildcard regions. Also, observe that skips generalize epsilon transitions: we can simply define skip $\epsilon = \{[0, 0]\}$. Third, we can support Kleene-star wildcard regions with $Skip(*) = \{[0, \infty)\}$, meaning any cursor less-than the length of the document works. The use of the ∞ symbol allows the separation of SAFA from the document to which it is applied.

SAFA. With the above notion of skips we can then defineSkipping Alternating Finite Automaton (SAFA)* as the 8-tuple $(Q, E, \Sigma, q_0, \lambda_q, \lambda_e, \delta, \mathcal{F})$, where Q is the set of all states (nodes); E is the set of all transitions (edges); Σ is the alphabet; $q_0 \in Q$ is the initial state; $\lambda_q : Q \to \{\forall, \exists\}$ defines the label for each node q to be either $\forall$ or $\exists$; $\lambda_e : E \to Skip \uplus \Sigma$ sets the label for each edge e as either a skip s or $\alpha \in \Sigma$; $\delta \subseteq Q \times E \times Q$ is the transition relation; and $\mathcal{F} \subseteq Q$ is the set of accepting states. The symbol $\uplus$ is the set *disjoint union*.

Much of this definition should look similar to the AFA in Section 5.1. The only difference is the addition of two new fields: E and λ_e. E is simply the set of all transitions. λ_e can be thought of as an analog of λ_q, but over transitions instead of states. It labels each transition $e \in E$ as taking a single step via a character (as is the case in AFA and NFA), or as a skip, which does not consume any characters from the document but increases the cursor non-deterministically by some amount in *s*.

Example. We defer the formal definition of skips and the various transitions to Appendix C.4 . In Figure 3 we show the SAFA that corresponds to the AFA from Figure 2. The SAFA replaces the long chain of states (4–10) in the AFA with *Skip*$\{[2, 6]\}$. This compression is possible because ϵ (the identity element) followed by skip *s* is just *s*.

The examples in Figures 2 and 3 provide the intuition for why SAFA might be cheaper to represent in an NP checker than AFA, while also being computationally equivalent (though SAFA requires the automaton to "know" how much to skip ahead of time). We formalize the equivalence between SAFA and AFA by direct translation.

Theorem 5.1. *Let $\mathcal{S}$ denote a SAFA. There exists an AFA $\mathcal{A}$ such that the language $\mathcal{L}[\![\mathcal{S}]\!] = \mathcal{L}[\![\mathcal{A}]\!]$ is regular.*

The proof is in Appendix C.6.

5.3 Designing the SAFA `match_step` Function

Section 4 introduces a `match_step` function that is appropriate for recursive proof systems. Reef modifies this step function to support SAFA. Reef's `match_step` takes in two additional arguments: `cursor_move`, which is the quantity by which $\mathcal{P}$ plans on advancing the cursor in the next transition, and a *stack*. One can represent a stack very cheaply with a simple hash chain (a single field element). A new stack is simply the value $stack = 0$. To push a value *val* just append it to the hash chain $stack = H(stack||val)$. To pop a value from *stack*, $\mathcal{P}$ must supply a preimage of *stack*; the first part of the pre-image will be the new stack, the other part is the popped value. That said, in our specific setting we can implement an

```
field[4] match_step(field[] commit, field[] blind,
    field state, field cursor, field cursor_move,
    field stack) {

  field cur_char = open_at(commit, blind, cursor);

  if (cur_char == EOD) { // end of the document
    if !is_accept(state) {
      return {0, 0, 0, 0}; // no match
    }

    if (is_empty(stack)) {
      return {0, 0, 0, 1}; // match
    } else {
      // reached accepting state in one branch
      // but there are other branches.
      // process next branch
      stack, (state, cursor) = pop(stack);
      return {state, cursor, stack, 0};
    }
  }
  // special handling for forall state
  if (is_forall(state)) {
    for child in children(state) {
      stack = push(stack, (child, cursor));
    }
    stack, (state, cursor) = pop(stack);
    return {state, cursor, stack, 0};
  }
  // perform character or skip transition
  state, cursor = delta(state, cursor, cur_char,
    cursor_move);
  return {state, cursor, stack, 0};
}
```

FIGURE 4—Reef's step function using SAFA.

even more efficient version since we know ahead of time the maximum depth of the stack (which depends on the number of nested `forall` states and the number of transitions). The details are provided in Appendix H.

Reef's `match_step` function is given in Figure 4. A key attribute for SAFA is that for a document D to be considered a match for regex $\mathcal{R}$, all children of a `forall` state must reach accepting states. Additionally, all of these children must start from the same cursor position, which is private. In Reef's `match_step` function, when a `forall` node is reached a copy of the cursor and the state ID is pushed onto the stack for each of the node's children. When one of the child branches terminates in an accepting state, its sibling and the original cursor position are popped from the stack.

Reef's `delta` function is then:

```
field[2] delta(field state, field cursor,
    field cur_char, field cursor_move) {
  field state, min, max = lookup(state, cur_char);
  assert(min <= cursor_move <= max);
  assert(cursor <= cursor + cursor_move);
  cursor = cursor + cursor_move;
  return {state, cursor};
}
```

Reef relies on *lookup tables* for determining whether a transition is valid. This is discussed more in-depth in Section 6, but in the context of our `delta` function, they work as follows: given a current state, character, and proposed quantity by which to move the cursor, we use a lookup table to validate the next state, as well as the minimum and maximum quantity the cursor is allowed to move, based on the type of skip. For example, if the transition is a skip "`{[n,m]}`", then `min`$= n$ and `max`$= m$. If the transition is `Skip(*)`, then `min`$= 0$ and `max`$= |\mathbb{F}| - 1$. In addition, we check that the new cursor position is greater than or equal to the current cursor position (i.e., that the prover did not decrement the cursor through an arithmetic overflow). In all other cases `max`$=$`min`$= 1$.

6 SAFA and Document Lookup Tables

Reef uses two lookup tables. One lookup table is public and represents the SAFA character and skip transitions; $\mathcal{V}$ can derive this public table from the regex. The other lookup table represents the document and is *private* (i.e., its contents cannot be revealed to the verifier). In each invocation of Reef's `match_step` (§5.3), the document table is accessed to read the character at the current cursor, and then the transition table is accessed to determine the next state.

This section reviews lookup arguments (§6.1), how Reef organizes the SAFA transitions table (§6.2); how it commits to the private table representing the document (§6.3); how it supports *table projections* that help filter which entries in the private table are relevant to a particular regex (§6.4); and how it combines both the public and private tables into one *hybrid* table that reduces the fixed costs of the lookup argument and improves its amortization (§6.5).

6.1 Lookup arguments

There are cases where one would want to check that a value v in an R1CS instance is contained in some table T of size n. A way to do this when T is public is to "hardcode" T in the R1CS instance by expressing it as a cascade of `if` statements similar to how we arithmetized the DFA's δ function (§3.1). Then, we check that v matches one of these `if` statements and not the final `return`. This requires $\mathcal{O}(|T|)$ constraints per lookup. An asymptotically cheaper (but sometimes concretely more expensive) solution is to use a Merkle Tree where the leaves represent T. One passes the root of the tree as a public input and a secret Merkle proof; the R1CS instance computes $\log(n)$ hashes to confirm there is a path to the root given v.

Lookup arguments [35, 39, 56, 76] generalize this idea: given m values $\{v_0, \ldots, v_{m-1}\}$ each in $\mathbb{F}$, lookup arguments check that all m values are entries in a table $T \in \mathbb{F}^n$. Crucially, lookup arguments amortize the costs over the m checks such that as m increases, the per-lookup cost decreases.

`nlookup` [56]. We briefly describe `nlookup`, which is designed for recursive proof systems such as the one we use (§4). For now, assume that the table is public. Section 6.3 describes additional techniques to handle private tables.

Let T be a table with $n = 2^\ell$ elements and let $\widetilde{T}$ be a multilinear polynomial in ℓ variables such that for all $i \in \{0,1\}^\ell$,

$\widetilde{T}(i) = T[\text{to-int}(i)]$, where $\text{to-int} : \{0,1\}^\ell \rightarrow \{0,1,\ldots,n-1\}$ is a function that maps ℓ-sized bit strings to ℓ-bit integers in a natural manner. Given $\widetilde{T}$, one can then prove that a value $v \in T$ by producing a point $q \in \{0,1\}^\ell$ such that $\widetilde{T}(q) = v$. `nlookup`'s core idea is to reduce the task of checking m of these lookup proofs to evaluating $\widetilde{T}$ at a single point. To do this, the `nlookup` prover proves:

$$\sum_{i=1\ldots m} \rho^i \cdot v_i = \sum_{i=1\ldots m} \rho^i \cdot \sum_{j\in\{0,1\}^\ell} \widetilde{eq}(q_i,j) \cdot \widetilde{T}(j)$$

where $v_i \in \mathbb{F}$ is the i-th value claimed by the prover to be in T, $\rho \in \mathbb{F}$ is a random challenge chosen by the `nlookup` verifier, and $\widetilde{eq}$ is a designated multilinear polynomial for performing Boolean equality checks. This equality can be proven using the sum-check protocol [60].

On its own, this is sufficient for proving membership of a set of elements in T. However, `nlookup` is particularly beneficial in the case where we would want to look up m elements *multiple times* (e.g., during different iterations of the step function of a recursive proof system). Readers familiar with the sum-check protocol can recall that in the above description, the verifier has to evaluate $\widetilde{T}$ at a random point at the end of the sum-check protocol.

In the case where we want to lookup m elements, k separate times, `nlookup` leverages a folding scheme to fold all k evaluations of $\widetilde{T}$ into a single one. It does this by initializing a running claim $v_r = \widetilde{T}(q_r)$ where $q_r, v_r \in \mathbb{F}^\ell$, and q_r is chosen arbitrarily. To incorporate new lookup claims (i.e., polynomial evaluations) into this running claim, `nlookup` makes a slight modification to the polynomial above. In particular, the sum-check protocol is now run over the polynomial:

$$v_r + \sum_{i=1\ldots m} \rho^i \cdot v_i = \sum_{j\in\{0,1\}^\ell} \widetilde{eq}(q_r,j) \cdot \widetilde{T}(j) + \sum_{i=1\ldots m} \rho^i \cdot \sum_{j\in\{0,1\}^\ell} \widetilde{eq}(q_i,j) \cdot \widetilde{T}(j)$$

which incorporates the running claim over foldings.

Integrating `nlookup` into Nova. To use `nlookup` with Nova, we encode `nlookup`'s verifier as an R1CS NP checker. This involves implementing the sum-check verifier and the associated Fiat-Shamir transform involving hash computations as R1CS. We then invoke this NP checker inside Reef's step function whenever we want to enforce that a group of R1CS variables are set to values contained in a table.

The cost to represent the above NP checker is as follows. To look up m entries in a table of size n within a step function, the number of constraints depends on the above two components: (1) sumcheck verifier and variable assignment, which requires $\mathcal{O}(m \cdot \log n)$ constraints with small constant; and (2) Fiat-Shamir transform which requires representing $\mathcal{O}(\log n)$ hash function evaluations in constraints, and each hash function requires hundreds of constraints.

Since expressing the hash functions is the dominant cost, lookup arguments are designed to amortize this component over the batch of m lookups. This is in contrast to using Merkle Trees which requires $\mathcal{O}(m \log n)$ hash functions represented as constraints to handle m lookups.

Since the `nlookup` verifier is encoded as an NP checker in R1CS, the Nova prover actually needs to know the witness for this checker so that it can prove the satisfiability of the statement. To compute this witness, the Nova prover has to do $\mathcal{O}(n)$ finite field operations per series of m lookups. Also, outside of R1CS (after the Nova verifier has checked the proof), the verifier performs an additional $\mathcal{O}(n)$ finite field operations at the very end of the protocol. A more detailed explanation of the protocol can be found in [56, §7] and in Appendix in Figure 27.

6.2 SAFA Lookup table

The lookup table T that Reef uses to encode the SAFA has a row for each transition in the SAFA and 5 columns—current state, character, next state, minimum cursor move, and maximum cursor move. The function of each of these columns is covered in Section 5.3. To convert this into the multilinear polynomial $\widetilde{T}$ needed for `nlookup` we manifest T as a vector of elements; each element represents an entire row and is computed by hashing the corresponding 5 columns to produce a value in $\mathbb{F}$. After a lookup takes place in Reef's step function, the result is therefore a single hash digest. To obtain the columns, the step function has constraints that allow the prover to supply the five values of the column entries, followed by a check that confirms that the hash of these values matches the looked up digest.

Constraints for SAFA lookups. As we discuss in Section 6.1, the number of constraints required for m lookups in a table of size n using `nlookup` is $\mathcal{O}(m \cdot \log n)$ constraints plus $\mathcal{O}(\log n)$ hash functions expressed as constraints. Each of the m lookups represents one SAFA transition. The SAFA table is of size $\mathcal{O}(|Q_{SAFA}| \cdot |\Sigma|)$ in the worst case—a transition for every character from every state. If the step function processes one SAFA transition at a time then $m = 1$ and the number of constraints to represent the single lookup is $\mathcal{O}(\log(|Q_{SAFA}| \cdot |\Sigma|))$ plus $\mathcal{O}(\log(|Q_{SAFA}| \cdot |\Sigma|))$ hashes.

Constraints across all steps. While it may seem that the total number of transitions (and therefore steps) should be at most $\mathcal{O}(|\mathsf{D}|)$, that is not always the case. With no lookarounds, the total number of transitions is $\leq |\mathsf{D}|$. However, because SAFA may have multiple branches for lookarounds, certain parts of D may be looked up more than once. We thus upper bound the number of transitions by $\alpha = \mathcal{O}(|\mathsf{D}| \cdot L)$, where L is the number of lookarounds in the regex. The number of constraints needed to check all of the transitions is thus $\mathcal{O}(\alpha \log(|Q_{SAFA}| \cdot |\Sigma|))$ plus $\mathcal{O}(\alpha \log(|Q_{SAFA} \cdot |\Sigma|))$ hashes.

Of course, the whole point of using a lookup argument is to benefit from its amortization, which is why Reef places multiple SAFA transitions within a single step function based

on the results of our optimizing compiler (§7). As a result, $m \geq 1$, so each step function has m transitions but Reef needs to run m times fewer steps. In this case, the total number of constraints across all steps is $\mathcal{O}(\alpha \log(|Q_{SAFA}| \cdot |\Sigma|))$ plus $\mathcal{O}(\frac{\alpha}{m} \log(|Q_{SAFA} \cdot |\Sigma|))$ hash functions. One might think that the optimal case is to have all lookups in a single step (which maximizes the amortization), but this is not so because there are other considerations as we explain in Appendix H .

6.3 Committing to a document

To commit to a document D over an alphabet Σ, the committer $\mathcal{G}$ first maps each character in Σ to an element in $\mathbb{F}$. Then, $\mathcal{G}$ simply treats D as a vector in $\mathbb{F}^n$. At this point, $\mathcal{G}$ can commit to D using any vector or polynomial commitment [27, 59, 67, 81]. That said, we choose a polynomial commitment since Reef uses a lookup argument to access SAFA transitions anyway, so using a lookup argument to access D allows us to combine both lookup tables to get lower costs (§6.5).

Note that if the optional transparency goal is desired, then the commitment scheme must be transparent (§3).

Polynomial commitment. $\mathcal{G}$ treats the vector D as a multilinear polynomial $\widetilde{T}$ in evaluation form and commits to $\widetilde{T}$ with a polynomial commitment. A polynomial commitment is a tuple of algorithms ($\mathit{Setup}, \mathit{Commit}, \mathit{ProveEval}, \mathit{VerifyEval}$). Informally, *Setup* outputs public parameters pp; *Commit* takes pp, a polynomial $\widetilde{T}$, and outputs a hiding and binding commitment to $\widetilde{T}$, $C_{\widetilde{T}}$; *ProveEval* takes pp, $\widetilde{T}$, a point q, value v, and outputs a proof π_{poly} that $\widetilde{T}(q) = v$; *VerifyEval* takes pp, $C_{\widetilde{T}}$, q, π_{poly}, and v and outputs whether $\widetilde{T}(q) = v$.

In our implementation we use the Hyrax polynomial commitment (Hyrax-PC) [81, §6.1], but one could make other choices to get different tradeoffs (e.g., Dory [59] has smaller commitments but its *ProveEval* algorithm results in larger proofs and is more expensive).

Let the Pedersen commitment for a vector $x \in \mathbb{F}^n$ be:

$$\mathit{Pedersen}(x, b) = h^b \cdot \prod_{i=1}^{n} g_i^{x_i}$$

where $g_1, \ldots, g_n$ and h are public random generators of the group over which the zkSNARK is defined (Pallas elliptic curve [47] in our case) and $b \in \mathbb{F}$ is a secret random blind. Hyrax-PC treats T as the column-major order of a $\sqrt{n}$-by-$\sqrt{n}$ matrix M, and commits to each row of M using a Perdersen vector commitment. This means that the commitment in Reef is $\sqrt{n}$ group elements, and there are $\sqrt{n}$ random blinds.

Making `nlookup` zero-knowledge. `nlookup` [56] does not explicitly discuss a way to guarantee zero-knowledge during lookups. Here we give a concrete proposal, based on standard techniques [34, 71, 81]. As we describe in Section 6.1, the output of the recursive proof system will include an `nlookup` running value v_r purported to be the evaluation of the multilinear polynomial $\widetilde{T}$ at a public random point $q_r \in \mathbb{F}$ specified by the Fiat-Shamir transform. When T is public, $\mathcal{V}$ can simply compute $\widetilde{T}(q_r)$ and check if it equals v_r. This is what we do with the SAFA table (§6.2). However, when T is private, there are two issues: (1) $\mathcal{P}$ cannot give $\mathcal{V}$ the claim v_r in the clear, as v_r is a weighted sum of the contents of T and would leak information; and (2) $\mathcal{V}$ does not have access to T and hence cannot compute $\widetilde{T}(q_r)$ on its own.

We address these issues as follows. First, instead of outputting v_r in the clear, we have the `match_step` function output d, where $d = H(v_r||s_1)$ and s_1 is a random secret value that $\mathcal{P}$ chooses. $\mathcal{P}$ can make d available to $\mathcal{V}$ without revealing anything about v assuming H heuristically instantiates a random oracle. $\mathcal{P}$ then computes another proof, $\pi_{consistency}$, with a separate non-recursive zkSNARK (we use Spartan [71]) for the statement: "given commitment c and public input d, I know a v_r such that $d = H(v_r||s_1)$ and $c = g^{v_r}h^{s_2}$ for some s_1 and s_2", where g and h are appropriate generators of the polynomial commitment. In effect, $\pi_{consistency}$ establishes that $\mathcal{P}$ correctly transformed one type of commitment (d) that is cheap to compute in R1CS but is not useful to verify polynomial evaluations, into another type of commitment (c) for the same value v_r that can be used to verify polynomial evaluations. Furthermore, $\pi_{consistency}$ is very cheap to compute (≈ 300 constraints) as we make c an *outer commitment* [33] (i.e., a commitment that is native to the underlying proof system) and does not need to be expressed in R1CS at all.

Second, recall that $\mathcal{V}$ has access to a polynomial commitment of $\widetilde{T}$, $C_{\widetilde{T}}$. $\mathcal{P}$ can then give $\mathcal{V}$ a proof $\pi_{poly} = \mathit{ProveEval}(\widetilde{T}, q_r, v_r)$, which $\mathcal{V}$ can use alongside q_r, c, and $C_{\widetilde{T}}$ to confirm that $\widetilde{T}(q_r) = v_r$. The key idea is to realize that, in Hyrax [81] and similar polynomial commitments [27, 59], the first step of $\mathit{VerifyEval}(C_{\widetilde{T}}, q_r, \pi_{poly}, v_r)$ is for $\mathcal{V}$ to turn the claim v_r into the Pedersen commitment $g^{v_r}h^{s_3}$ for some s_3. However, $\mathcal{V}$ already has $c = g^{v_r}h^{s_2}$ and a proof $\pi_{consistency}$ that establishes that c is a valid Pedersen commitment for v_r. Hence, $\mathcal{V}$ can simply use c instead.

Security. Observe that the verifier sees d, c, $C_{\widetilde{T}}$, q_r, $\pi_{consistency}$ and π_{poly}. From this information, the verifier learns nothing about v_r beyond the fact that d and c commit to the same value, and that c is a commitment to a correct evaluation of a polynomial underneath the commitment $C_{\widetilde{T}}$ at point q_r. This is because $\pi_{consistency}$ and π_{poly} are both zero-knowledge arguments, and the three commitments d, c, and $C_{\widetilde{T}}$ are hiding.

6.4 Table projections

For proving m lookups over a committed document of size n, `nlookup`'s prover incurs $\mathcal{O}(n)$ operations over $\mathbb{F}$. Although these are not expensive group operations, when n is large (e.g., billions), this can be expensive. On the other hand, in some applications, it is public information that lookups will be made to particular portion of the document (though the actual content within that portion of the document is private). For example, a study may just care about DNA regions that start at publicly known offsets.

To address this, we describe an approach to run nlookup on a *projected* table (one that contains one or more "chunks" of an original table) such that the prover incurs costs proportional to the size of the projected table. Furthermore, the verifier still only needs a commitment to the original table. The core idea is to leverage certain basic facts about multilinear polynomials to reduce claims about a projected table to claims about the original table.

We begin with an overview, which we then generalize. Let T be the original table with $n = 2^\ell$ elements, and $\widetilde{T}$ be its multilinear extension as described in Section 6.1. Suppose we project T into a smaller table T'; $\widetilde{T'}$ is then a multilinear polynomial in $\ell' < \ell$ variables. It turns out that $\widetilde{T'}$ and $\widetilde{T}$ are related in a fundamental way. This is what enables us to run nlookup on T'. At the end of nlookup, the verifier is left with a claim about T', of the form $\widetilde{T'}(q_r) = v_r$. However, the verifier only has a commitment to the original table T. To address this, we transform this claim to an equivalent claim about an evaluation of $\widetilde{T}$, allowing the verifier to check the claim about $\widetilde{T}$ using a commitment to T. We now elaborate.

We use a concrete example, to provide intuition. Suppose that $T = [a,b,c,d,e,f,g,h]$, so $\widetilde{T}$ is a multilinear polynomial in $\ell = 3$ variables. Suppose the projected table is $T' = [c,d]$, so $\ell' = 1$. For this example, it follows that for all $q_r \in \mathbb{F}^{\ell'}$ $\widetilde{T'}(q_r) = \widetilde{T}(s, q_r)$, where $s = 01 \in \{0,1\}^2 = \{0,1\}^{\ell-\ell'}$. In the context of nlookup, to check that $\widetilde{T'}(q_r) = v_r$, the verifier can instead check $\widetilde{T}(s,q_r) = v_r$, where $s = 01$. A key take-away here is that for $0 \leq \ell' \leq \ell$, observe that a specified prefix $s \in \{0,1\}^{\ell-\ell'}$ "selects" a unique chunk of T and specifies a particular projection of size $2^{\ell'}$.

Note that this approach generalizes to project non-contiguous chunks of T. For simplicity, suppose that we want to project two chunks of T, specified with two selectors $s_1 \in \{0,1\}^{\ell'}$ and $s_2 \in \{0,1\}^{\ell'}$, where $0 \leq \ell' \leq \ell$. The projected table $T' = (L,R)$ is a vector of size $2^{\ell-\ell'+1}$ and L and R are vectors of size $2^{\ell-\ell'}$, so $\widetilde{T'}$ is a multilinear polynomial in $\ell - \ell' + 1$ variables. When we run nlookup with the projected table T', the verifier ends up with a claim about the projected table of the form $\widetilde{T'}(q_r) = v_r$, where $q_r \in \mathbb{F}^{\ell-\ell'+1}$. Again, derived from the properties of multilinear polynomials,

$$\begin{aligned}\widetilde{T'}(q_r) &= (1-q_r[0]) \cdot \widetilde{L}(q_r[1..]) + q_r[0] \cdot \widetilde{R}(q_r[1..]) \\ &= (1-q_r[0]) \cdot \widetilde{T}(s_1, q_r[1..]) + q_r[0] \cdot \widetilde{T}(s_2, q_r[1..])\end{aligned}$$

Thus to check if $\widetilde{T'}(q_r) = v_r$, the verifier can instead check if $(1-q_r[0]) \cdot \widetilde{T}(s_1, q_r[1..]) + q_r[0] \cdot \widetilde{T}(s_2, q_r[1..]) = v_r$, which makes two evaluation queries to $\widetilde{T}$. Note that this idea generalizes to projecting $k > 2$ non-contiguous chunks of T.

Low-cost padding to hide document size. In many settings, one would like to hide not just the content of D, but also its size. For example, if D is a password, revealing its size reveals the password's length. Projections allow the commitment generator $\mathcal{G}$ to pad the document to some upper bound (essentially for free) while allowing $\mathcal{P}$ to perform operations proportional to the unpadded document and without having to reveal the selector s to $\mathcal{V}$. Appendix G has the details.

6.5 Hybrid private/public lookup argument

Reef's step function (§5.3) looks up values from two tables: the public SAFA table (S) and the private document table (D). We can do this with two separate instances of nlookup, one for each table. However, this requires $m\log(|\mathsf{D}| \cdot |S|) + \mathcal{O}_H(\log(|\mathsf{D}| \cdot |S|))$ constraints where m is the number of lookups to each table per step.

Instead, we combine both tables into a single *hybrid* table, all while preserving the privacy requirements of the document table. Accessing this hybrid table requires only $2m\log(|D| + |S|) + \mathcal{O}_H(\log(|\mathsf{D}| + |S|))$ constraints. This optimization does not pay off only when one of the tables is multiple orders of magnitude larger than the other. But we never encountered an imbalance between $|\mathsf{D}|$ and $|S|$ large enough to nullify the benefits in any of our experiments.

$\mathcal{P}$ has access to S and D and can merge the tables by pretending they are two halves of a large table T and running the nlookup prover. At the end, $\mathcal{V}$ will end up with a single claim about the multilinear extension of T: $\widetilde{T}(q_r) = v_r$, where $q_r \in \mathbb{F}^\ell$ and $\ell = \log(2 \cdot \max(|D|, |S|))$. Since T in this case includes private data, $\mathcal{V}$ should not see v_r in the clear, and instead receives: $d = H(v_r || s_1)$, C_{v_r} (a Pedersen commitment to v_r), and a proof $\pi_{consistency}$ as we discuss in Section 6.3.

To verify $\widetilde{T}(q_r) = v_r$, $\mathcal{V}$ must treat the public and private parts of the large table as separate "indexable" chunks, similar to the way projections work. We define $\widetilde{T}(q_r)$ as:

$$\widetilde{T}(q_r) = (1-q_r[0]) \cdot \widetilde{S}(q_r[1..]) + q_r[0] \cdot \widetilde{\mathsf{D}}(q_r[1..]) = v_r$$

Notice that this means we need to arrange T such that it can be divided equally into a public half (indexed by $q_r[0] = 0$) and a private half ($q_r[0] = 1$). The smaller of the two tables will be padded to the size of the other, which is why $\ell = \log(2 \cdot \max(|\mathsf{D}|, |S|))$ above, and why the hybrid table becomes inefficient if one table is extremely larger than the other. Lookups to the public half of the table use exactly the same indices as before. Lookups to the private half will use the same indices as before added to $2 \cdot \max(|D|, |S|)$.

Given this structure, $\mathcal{P}$ evaluates $\widetilde{D}$ at the point $q_r[1..]$ and obtains a value $v_d \in \mathbb{F}$. $\mathcal{P}$ then generates a commitment C_{v_d} to v_d, and a proof $\pi_{poly} = \mathit{ProveEval}(\widetilde{D}, q_r[1..], v_d)$ that establishes that $\widetilde{D}(q_r[1..]) = v_d$. For its part, $\mathcal{V}$ computes $\widetilde{S}(q_r[1..]) = v_s$ on its own, and runs *VerifyEval* on π_{poly} using the document commitment, $C_{\widetilde{D}}$, and C_{v_d}.

So far, we have proceeded very similarly to the verification of the running claim in the non-hybrid model. But notice that $\mathcal{V}$ must still relate v_s and C_{v_d} to C_{v_r} in the following way:

$$(1-q_r[0]) \cdot v_s + q_r[0] \cdot v_d = v_r$$

This is done as follows. $\mathcal{V}$ computes C_L, which is a Pedersen commitment to the value on the left-hand-side of the

above equation using v_s and C_{v_d} (this requires only linear operations on Pedersen commitments, which are linearly homomorphic). $\mathcal{P}$ then proves that C_L and C_{v_r} commit to the same value using a Schnorr [70] zero-knowledge proof of equality π_{eq}.

Security. When the verifier computes the commitment C_L, it does not learn any additional information about v_d as the operations are done using C_{v_d} (C_{v_d} is a commitment that hides the underlying value v_d). Furthermore, π_{eq} proves that the values under the commitments C_L and C_{v_r} are the same without revealing any additional information.

7 Implementation

Reef is implemented in 14K lines of Rust and is open source [8]. We discuss the main components here and optimizations in Appendix H .

7.1 Compilation: from regex to R1CS

Reef has two levels of compilation. First, Reef compiles regexes written in standard PCRE syntax [7] (Figure 1) and produces a SAFA. From this SAFA, Reef generates the SAFA's transition lookup table and the `match_step` function discussed in Section 5.3. Since the `match_step` function uses lookups it also contains the checks that the `nlookup` verifier [56] must perform in each step. In particular, it contains a series of Fiat-Shamir challenges that we generate with the Poseidon hash function [44] using the Neptune library [3]. Finally, Reef uses the CirC [66] compiler to output R1CS instances that we convert to Bellman [1] instances.

7.2 Solving: finding the satisfying witness

Reef, given a document D, finds the witness to the R1CS instance representing `match_step` in two parts. First, Reef derives which paths in the SAFA to take, the skip values, the entries in D to read, and the rows in the transition table to look up. Reef's solver might be of independent interest and we discuss it in Appendix E.4. This solver only needs to run once and tells $\mathcal{P}$ how many steps to prove.

Second, for each step, Reef runs the `nlookup` prover, which we implement as there was no prior implementation, to generate the values that will satisfy the `nlookup` checks that were inserted in the corresponding `match_step`. The result of this and the SAFA solver are sufficient to construct the entire solution vector $z_i = (y_i, 1, w_i)$ where w_i is the witness and y_i is the output of step i.

7.3 Proving knowledge of the witness

For the proving and verifying, we use Nova [4], which we modify to make it zero-knowledge (the existing implementation was only succinct). This required changing 1.6K lines of Rust to hide the number of steps executed, and making the commitments hiding, and the folding scheme, sumcheck protocol, inner product argument, and SNARK zero-knowledge. Our modified version of Nova is open source [5].

8 Costs and Complexity analysis

In this section we discuss the asymptotic costs of all of the components of Reef. The analysis below considers the case where Nova [57] uses Pedersen commitments to commit to vectors, and Spartan [71] uses an IPA-based polynomial commitment scheme to compress incrementally generated proofs. Furthermore, for `nlookup` [56], the analysis considers the case where documents are committed with Hyrax's polynomial commitment scheme [81]. Finally, one of the basic operations of the above proof systems are *multiexponentiations*: given generators $g_1, \ldots, g_n$, and exponents $e_1, \ldots, e_n$, compute $g_1^{e_1} \cdot g_2^{e_2} \cdots g_n^{e_n}$. These are also called *multi-scalar multiplications* (MSM). These proof systems typically use Pippenger's algorithm [68] which can compute a size-n MSM in $\mathcal{O}(n\lambda/\log(n\lambda))$ group operations. We will ignore the security parameter λ and just treat a size-n MSM as $\mathcal{O}(n/\log n)$ group operations.

For simplicity, let $T = |\mathsf{D}| + |Q_{SAFA}|$ be the sum of the size of both the document and the SAFA lookup tables.

Committer's costs. Committing to a document D with Hyrax's polynomial commitment [81] requires the committer $\mathcal{G}$ to perform $\mathcal{O}(|\mathsf{D}|/\log\sqrt{|\mathsf{D}|})$ group operations.

Prover's costs. Ignoring the distinction between the arithmetization of hash functions and other operations, the contribution of the lookup argument towards Reef's step function is $\mathcal{O}(m \log T)$ R1CS constraints; Reef requires a total of $\mathcal{O}(\alpha/m)$ steps to finish processing a document. Nova performs $\mathcal{O}(m\log(T)/\log(m\log T))$ group operations per step. This results in $\mathcal{P}$ performing a total of $\mathcal{O}(\alpha\log(T)/\log(m\log T))$ group operations. The resulting proof π is of size $\mathcal{O}(\log(m\log T))$.

In addition, during each step, Reef needs to run the `nlookup` prover in order to generate the relevant portion of the satisfying witness for the R1CS instance. This requires computing the sumcheck protocol over the hybrid table, which necessitates $\mathcal{O}(T)$ field operations. If projections are used, then D is substituted with D_{proj} in the definition of T.

At the end of the protocol $\mathcal{P}$ needs to compute *ProveEval* in order to generate π_{poly} so that $\mathcal{V}$ can verify the private component of the hybrid table. This requires $\mathcal{P}$ to perform $\mathcal{O}(\sqrt{|\mathsf{D}|}/\log\sqrt{|\mathsf{D}|})$ group operations. The proof, π_{poly}, is of size $\mathcal{O}(\log|\mathsf{D}|)$.

Finally, our zero-knowledge extension to the lookup argument for D requires generating the proof $\pi_{consistency}$, which is done with a constant-size R1CS instance, and therefore $\mathcal{O}(1)$ group operations in Spartan [71].

$\mathcal{P}$ performs $\mathcal{O}(\alpha\log(T)/\log(m\log T) + \sqrt{|\mathsf{D}|}/\log\sqrt{|\mathsf{D}|})$ group and $\mathcal{O}(T)$ finite field operations in total.

Verifier's costs. The cost to the verifier $\mathcal{V}$ is $\mathcal{O}(m\log(T)/\log(m\log T))$ group operations in Nova to verify π. Further, $\mathcal{V}$ must invoke Hyrax's *VerifyEval* to check π_{poly}, which requires $\mathcal{O}(\sqrt{|\mathsf{D}|}/\log\sqrt{|\mathsf{D}|})$ group

Application	Document Size	SAFA States	SAFA Transitions
Redactions			
Small Email	415	331	42,318
Large Email	1,000	908	116,751
ODoH	128	36	4,012
Passwords			
Match	12	21	1,188
Non-Match	9	21	1,188
DNA			
Match	32.3×10^6	976	4,861
Non-Match	32.3×10^6	976	4,861

FIGURE 5—Document and SAFA size for evaluated applications

operations. Lastly, the verifier needs to evaluate the public component of the hybrid table which requires $\mathcal{O}(|Q_{SAFA}|)$ finite field operations.

$\mathcal{V}$ performs $\mathcal{O}(m\log(T)/\log(m\log T)+\sqrt{|\mathsf{D}|}/\log\sqrt{|\mathsf{D}|})$ group and $\mathcal{O}(|Q_{SAFA}|)$ finite field operations in total.

Alternate approach. In Appendix I we discuss how if we instantiate the SAFA table and the document commitment using a Merkle tree, the asymptotic costs of Reef are much lower (logarithmic number of group and finite field operations). However, the lower asymptotics do not translate to lower concrete costs due to much higher constants.

9 Evaluation

This section answers Reef's motivating questions: is proving general regular expression matching in zero knowledge practical for various applications and do Reef's optimizations meaningfully reduce the costs? Our results indicate that this is indeed the case.

9.1 Experimental Setup

Reef runs fine on a laptop (Intel Core i7 1.9 GHz, 16GB RAM) since its use of recursion means that $\mathcal{P}$ proves one step at a time and therefore uses little memory; at most 5.1 GB in our largest experiment. However, in order to run the baselines which require more memory, we run all of our experiments (including Reef) on a 16-core Intel Xeon Platinum 8253 CPU (2.20GHz) with 764 GB of RAM. We evaluate Reef over the applications discussed in Section 1: proving password strength, disclosing redacted emails, ODoH blocklisting, and genetic proving. For each of our use cases we evaluate documents and regexes of varying sizes. Figures 5 and 6 report the document sizes, SAFA sizes, and results for the largest instances based on SAFA size. However, full results, all document sizes, and a list of all regexes can be found in Appendix J.

9.2 Overall Performance

We start by showing the end-to-end results of Reef on our applications, averaged over 10 runs, and then later break down some of these costs to show the benefits of each of Reef's optimizations.

Compilation. Compiling a regex to R1CS is the most time consuming part since it requires parsing the regex and generating the SAFA, lookup tables, and R1CS matrices. This includes the generation of the document commitment. However, this is typically a one-time cost and can be done in advance since the regex is public.

Solving (witness generation). Reef's witness generation includes the time to find the regex match, the right values for all the skips in SAFA, running the `nlookup` prover (whose output becomes a witness value to the step function), and finding the satisfying assignment to all R1CS variables. In most cases, all of this can be done in a few milliseconds; the exceptions are large documents (e.g., DNA or large emails) which require considerable time.

Proving. Proving time depends on document length, the regex complexity, how many steps the prover needs to run, and the size of each step. It includes the time to generate all the proofs, including the consistency and equality proofs of the hybrid table (§6.5). In Appendix H we discuss how Reef often batches many character and skip transitions into one step (leading to a larger step function but fewer total steps). Reef generally performs worse on regexes where the regex is similar to the document, as it gives Reef's prover fewer opportunities to skip and stop early. For example, the email redaction regexes are very similar to the original document, and hence result in more proving steps than some of the other regexes, and consequently larger proving time.

Reef's benefits are best exemplified with the DNA matching application, in which the document has over 32 million characters. Reef is able to generate succinct proofs for DNA in under 30 seconds (including both solving and proving) because it can avoid processing most of the document, thanks to its use of skips and projections.

Verification. The verifier's costs depend on the number of R1CS constraints for a single step (since Nova folds all steps into one), as well as the cost to evaluate the SAFA polynomial at a random point, and check the consistency polynomial evaluation, and the equality proof. Nova's current implementation uses Bulletproofs's [27] linear-time inner product argument on the folded instance (which we made zero-knowledge in our evaluation); so while it has logarithmic proofs it still has verification linear in the size of one step. This could be expensive when the step function is large, but our step functions are relatively small (under 100K constraints). As a result, verification in Reef takes less than 1 second in all of our applications and workloads.

Proof size. The proof column includes all materials needed for the verifier to check the prover's claim. This includes all commitments and auxiliary proofs (e.g., $\pi_{consistency}, \pi_{poly}, \pi_{eq}$). Reef is succinct so all proof sizes are sublinear (logarithmic) in the size of the statement being proven. However, Reef's

Application	Constraints per step	# of Steps	Compiler Time (s)	Solver Time (s)	Prover Time (s)	Verifier Time (s)	Proof Size (KB)	Commitment Size (KB)
Redactions								
Small Email	46,655	4	36.947	0.760	3.169	0.553	32.609	0.512
Large Email	65,727	7	217.628	3.221	5.923	0.701	33.361	1.024
ODoH	22,692	2	19.650	0.213	1.709	0.435	31.889	0.512
Passwords								
Match	19,982	5	17.960	0.067	2.573	0.418	31.665	0.128
Non-Match	20,728	6	18.636	0.357	2.963	0.416	31.761	0.128
DNA								
Match	81,722	8	62.351	12.830	17.708	0.908	34.417	131.072
Non-Match	81,722	1	62.357	3.006	10.838	0.915	34.417	131.072

FIGURE 6—Summary of all costs for the largest instance of each application evaluated in Reef. R1CS Constraints are for one step in Nova. Proof sizes include all the Nova zkSNARK proof as well as all auxiliary proofs (e.g., $\pi_{consistency}$). Commitment size measures the size of the document commitment. Reported times are the mean across 10 runs, and the standard deviation was less than 5% of the mean for all components and applications.

Application	DFA	DFA + Recursion	SAFA + nlookup	Reef
Redactions				
Small Email	76.300	1.721	0.760	0.733
Large Email	—	5.848	1.067	1.051
ODoH	2.064	0.640	0.409	0.362
Passwords				
Match	—	—	0.351	0.347
Non-Match	—	—	0.343	0.330
DNA				
Match	—	—	9.392	5.091
Non-Match	—	—	8.389	5.032

FIGURE 7—Maximum memory used (GB) by the Prover in Reef and baselines for our applications. Verifier's memory use is lower.

use of Hyrax means that document commitments consist of $\sqrt{|D|}$ group elements. When the document is very large, such as in DNA, this can be sizable.

9.3 Comparative Performance

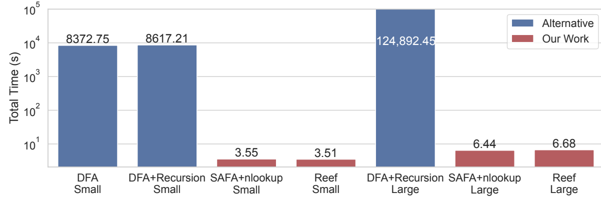
To contextualize the benefits of Reef, we compare it against several alternatives:

- **DFA**. This is the standard approach articulated in Section 3.1. To our knowledge, this is also the approach taken by the ZK-Email project [16]. We use Circom [2] to compile the match function and solve the corresponding R1CS instance since CirC [66] is not presently capable of compiling such large statements due to memory issues.
- **DFA + recursion**. This is the approach described in Section 4, which adds recursion and processes one character at a time. It uses a hash-chain as a vector commitment, which we believe is optimal (exactly one hash invocation) when accessing entries in the committed document sequentially. We use Circom and NovaScotia [6] to compile the step function and connect it with our zero-knowledge version of Nova (§7.3). Again, we are unable to compile these R1CS instances with CirC since they require expressing the (large) DFA delta function in constraints.
- **SAFA + lookup.** This is our implementation of Reef (§7) with SAFA and `nlookup`, but without projections (§6.4) or the hybrid table optimization (§6.5).

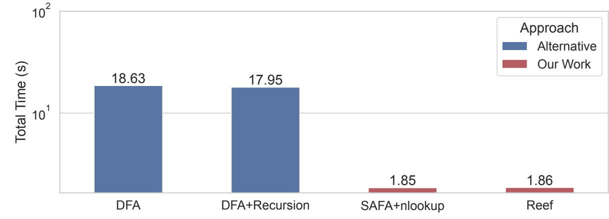
The metrics that we will consider in this section are memory usage and *end-to-end* completion time for the Prover, which includes both the time to solve and generate all witness values, and prove the satisfiability of the R1CS instance. Appendix J has additional graphs for these same experiments but separates the time for solving and proving for readers interested in understanding the contribution of each component towards the end-to-end time. One thing to consider is that Reef pipelines the generation of a proof for step i with the generation of the witness for step $i+1$ in parallel, as we discuss in Appendix H. As a result, the end-to-end time can sometimes be lower than the sum of the corresponding proving and solving times.

Results. Figure 7 shows the maximum memory use of Reef and the baselines for the same documents and regexes found in Figure 6. We are unable to run the password matching application with either of the DFA baselines due to its use of lookaheads, and the DNA application due to the massive R1CS instances (or number of steps) that are required. There are two observations: (1) using a recursive proof system has significant benefit in keeping the amount of memory required by the prover small since the prover only needs to prove one step at a time; and (2) the use of table projections in the DNA application means that the prover does not need to compute an expensive and memory-intensive sumcheck over the entire document, but rather works only over the projected table. This is why Reef uses less memory than SAFA + `nlookup`.

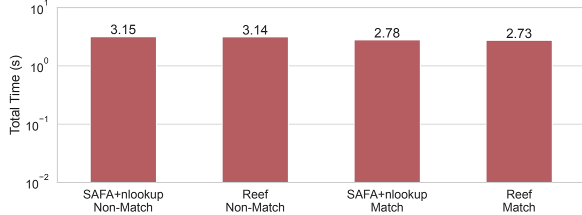
Figure 8 shows the end-to-end performance results. Across the board, SAFA +`nlookup` and Reef both dramatically outperform the DFA and DFA+Recursion approaches. Take for example *Redactions Small*. SAFA +`nlookup` and Reef took 3.55 and 3.51 seconds generate witnesses and prove, while the



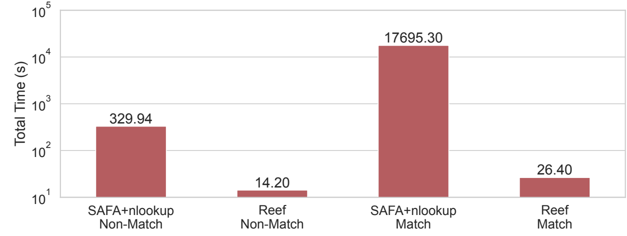
(a) Proof that a (small / large) committed email matches a redaction regex. DFA was unable to handle the large email.



(b) Proof that a committed document matches an ODoH regex.



(c) Proof that a committed password matches/does not match a password strength regex.



(d) Proof that a committed DNA document matches/does not match a DNA regex. Neither DFA nor DFA+recursion can handle this application.

FIGURE 8—Mean end-to-end completion time (which includes witness and proof generation) across 10 runs for proving that some committed document matches/does not match a regex with Reef and various alternatives. Standard deviations were less than 5% of the mean. Each subfigure describes a different application (regex) and type of document. The corresponding document sizes are found in Figure 6.

DFA baselines took over an hour. This suggests that Reef's ability to skip irrelevant parts of the document and the use of our zero-knowledge version of `nlookup` provides benefits.

One might notice that DFA + recursion actually performs *worse* than just DFA in the case of small email redactions. There are a few reasons for this. First, while each step can process multiple characters, because the circuit still relies on `for` loops, it can only process a few characters per step before the circuit becomes too large. Second, in each step, there is some non-trivial work that is performed to check if the document is a match (§3.1). Third, each step of Nova adds a check to ensure foldings are correct (≈20,000 constraints).

Note that Reef also suffers from the latter two overheads (though the specific invariants for checking a match in a SAFA are different than in a DFA). However, Reef's use of lookup tables allow it to efficiently process large numbers of characters per step, which amortizes the latter two overheads over a batch of characters. Indeed, one of our optimizations (Appendix H) is to process the optimal number of characters per step for a given regex in the SAFA's `match_step`, which amortizes these costs over the batch. We find this optimal value with a cost model that we have implemented in Reef's compiler.

The final impact to consider is that of Reef's additional optimizations. As discussed in Section 6.5, using hybrid tables reduces the number of constraints needed. This reduction is usually 1K–3K fewer constraints in the step function; full results are in Appendix J. This reduction in the size of the step function results in some small performance gains.

More significant is the impact of document projections in our DNA applications. Because common variants in the genome occur at known, fixed locations, by using projections (§6.4) Reef can skip over large parts of the genome directly to the start of the variant of interest. In the case of DNA matching, this results in a 50% reduction in proving time, and an over 99% reduction in solving time. While SAFA +`nlookup` can avoid the costs of a large document when it comes to proving, it still has to evaluate the sum-check protocol on the entire document for each step. When working with a document as large as DNA, this rapidly becomes prohibitive.

Takeaway. Reef handles a wide class of regexes at reasonable cost while producing succinct proofs. Each of Reef's optimizations provide benefits: SAFA allows expressing complicated regexes and skipping irrelevant parts of the document; recursion unleashes the power of SAFA by allowing the prover to prove only for as long as needed, and requiring much fewer memory; Reef's compiler picks the optimal number of characters to process per step for a regex to reduce the penalty of non-uniformity during recursion; hybrid lookup tables reduce the size of the step function; and projections make it possible for the prover to solve more efficiently when the location of relevance within the document is public.

10 Related Works

Reef relates to a series of very recent works on building proof systems for regexes [16, 61, 69, 84]. Reef aims to be as general as possible—targeting complex PCRE expressions and arbitrarily long documents. Reef achieves this by introducing SAFA, a brand new automata. In contrast, these other works target particular applications (middlebox packet inspection, malware hash membership tests) and use existing automata (DFAs or NFAs) enhanced with various encoding optimizations for their application domains. Reef can also handle these applications (and many others). It is unclear whether Reef

would achieve better performance on these applications over these tailored proposals as we have not yet done an empirical comparison (they were all developed concurrently with Reef). One exception is ZK Regex from the ZK Email Verify project [16], which is in effect the "standard" approach in our evaluation, and which Reef outperforms in all applications.

Another related area is that of *secure regex evaluation* [38, 53, 58, 61, 63, 80]. Here the goal is for one party to supply the regex $\mathcal{R}$ and another party the document D, and to determine whether $\mathsf{D} \in \mathcal{L}[\![\mathcal{R}]\!]$ without revealing their inputs. This is a multi-party computation, and the techniques used in this domain aim to express *computation* rather than *verification*, which is the main theme in our work (via NP checkers).

11 Discussion and Future Work

Reef is the most expressive zero-knowledge proof system for regexes to date. It excels in situations where the document is large and the match is small, or when the regex gives Reef many opportunities to skip unnecessary work. In contrast, works like Zombie [84] excel in the opposite regime (small documents or when the document closely matches the regex). We think there are opportunities to combine the techniques in these two approaches to obtain the best of both worlds.

Reef has the ability to prove regex matches (and non-matches), but an interesting extension is to support "search and replace". In such a setting, the prover would prove not whether there is a match for some regex but rather that some committed document is the result of performing a regex search and replace transformation on some other committed document. Another extension to Reef is to support context free grammars. We think a similar approach of developing a custom automata would work there, and Reef already uses a stack for SAFA, which we show is quite efficient.

Source Code

Our code is available at:
https://github.com/eniac/Reef.

Acknowledgements

We thank Justin Thaler for insightful discussions on the problem of matching with wildcards and Riad Wahby for clarifying some questions on Hyrax and its polynomial commitment. We also thank Zachary DeStefano, Paul Grubbs, and Michael Walfish for helping us better understand Zombie, and Xiang Fu for pointing out a typo in an earlier version of this paper and for helping us better understand `zkreg`. Weidong Cui encouraged us to formalize our automata, which ultimately led to our formal development of SAFA. This work was funded in part by NSF grants CNS-2045861, CNS-2107147, CNS-2124184, a gift from the Arcological Swiss Association, and DARPA contract HR0011-17-C0047. Any views expressed herein are solely those of the authors listed.

References

[1] bellman. https://github.com/zkcrypto/bellman.

[2] Circom. https://github.com/iden3/circom.

[3] Neptune. https://github.com/lurk-lab/neptune.

[4] Nova: Recursive SNARKs without trusted setup. https://github.com/microsoft/Nova.

[5] Nova: Recursive SNARKs without trusted setup. https://github.com/sga001/Nova.

[6] Nova-scotia. https://github.com/nalinbhardwaj/Nova-Scotia.

[7] Perl-compatible regular expressions (PCRE). https://www.pcre.org/original/doc/html/pcresyntax.html.

[8] Reef: A zkSNARK system for proving that a committed document matches a regex. https://github.com/eniac/Reef.

[9] Regex filters for pi-hole. https://github.com/mmotti/pihole-regex/blob/master/regex.list.

[10] https://bit.ly/reef-min-dfa, 2023.

[11] https://nordpass.com/most-common-passwords-list/, 2023.

[12] https://www.cs.cmu.edu/~enron/, 2023.

[13] https://www.ncbi.nlm.nih.gov/gene/672, 2023.

[14] https://www.ncbi.nlm.nih.gov/gene/675, 2023.

[15] Introduction to Microsoft Entra Verified ID. https://learn.microsoft.com/en-us/azure/active-directory/verifiable-credentials/decentralized-identifier-overview, 2023.

[16] Zk email verify. https://github.com/zkemail/zk-email-verify, 2023.

[17] A. V. Aho and M. J. Corasick. Efficient string hatching: an aid to bibliographic search. *Communications of the ACM*, 18, 1975.

[18] S. Ames, C. Hazay, Y. Ishai, and M. Venkitasubramaniam. Ligero: Lightweight sublinear arguments without a trusted setup. In *Proceedings of the ACM Conference on Computer and Communications Security (CCS)*, 2017.

[19] S. Angel, A. J. Blumberg, E. Ioannidis, and J. Woods. Efficient representation of numerical optimization problems for SNARKs. In *Proceedings of the USENIX Security Symposium*, 2022.

[20] V. Antimirov. Partial derivatives of regular expressions and finite automaton constructions. *Theoretical Computer Science*, 155(2):291–319, 1996.

[21] E. Ben-Sasson, I. Bentov, Y. Horesh, and M. Riabzev. Scalable, transparent, and post-quantum secure computational integrity. Cryptology ePrint Archive, Paper 2018/046, 2018. https://eprint.iacr.org/2018/046.

[22] E. Ben-Sasson, A. Chiesa, E. Tromer, and M. Virza. Scalable zero knowledge via cycles of elliptic curves. In *Proceedings of the International Cryptology Conference (CRYPTO)*, 2014.

[23] D. Boneh, J. Drake, B. Fisch, and A. Gabizon. Halo Infinite: Recursive zk-SNARKs from any Additive Polynomial Commitment Scheme. In *Proceedings of the International Cryptology Conference (CRYPTO)*, 2021.

[24] S. Bowe, J. Grigg, and D. Hopwood. Recursive proof composition without a trusted setup. Cryptology ePrint Archive, Paper 2019/1021, 2019. https://eprint.iacr.org/2019/1021.

[25] B. Braun, A. J. Feldman, Z. Ren, S. Setty, A. J. Blumberg, and M. Walfish. Verifying computations with state. In *Proceedings of the ACM Symposium on Operating Systems Principles (SOSP)*, 2013.

[26] J. A. Brzozowski. Derivatives of regular expressions. *Journal of the ACM (JACM)*, 11(4):481–494, 1964.

[27] B. Bünz, J. Bootle, D. Boneh, A. Poelstra, P. Wuille, and G. Maxwell. Bulletproofs: Short proofs for confidential transactions and more. In *Proceedings of the IEEE Symposium on Security and Privacy (S&P)*, 2018.

[28] B. Bünz and B. Chen. ProtoStar: generic efficient accumulation/folding for special sound protocols. Cryptology ePrint Archive, Paper 2023/620, 2023. https://eprint.iacr.org/2023/620.

[29] B. Bünz, A. Chiesa, W. Lin, P. Mishra, and N. Spooner. Proof-carrying data without succinct arguments. In *Proceedings of the International Cryptology Conference (CRYPTO)*, 2021.

[30] B. Bünz, A. Chiesa, P. Mishra, and N. Spooner. Proof-carrying data from accumulation schemes. In *Proceedings of the Theory of Cryptography Conference (TCC)*, 2020.

[31] P. Caron, J.-M. Champarnaud, and L. Mignot. A general framework for the derivation of regular expressions. *RAIRO-Theoretical Informatics and Applications*, 48(3):281–305, 2014.

[32] A. K. Chandra, D. C. Kozen, and L. J. Stockmeyer. Alternation. *Journal of the ACM (JACM)*, 28(1), 1981.

[33] C. Costello, C. Fournet, J. Howell, M. Kohlweiss, B. Kreuter, M. Naehrig, B. Parno, and S. Zahur. Geppetto: Versatile verifiable computation. In *Proceedings of the IEEE Symposium on Security and Privacy (S&P)*, May 2015.

[34] R. Cramer and I. Damgård. Zero-knowledge proofs for finite field arithmetic, or: Can zero-knowledge be for free? In *Proceedings of the International Cryptology Conference (CRYPTO)*, 1998.

[35] L. Eagen, D. Fiore, and A. Gabizon. cq: Cached quotients for fast lookups. Cryptology ePrint Archive, Paper 2022/1763, 2022. https://eprint.iacr.org/2022/1763.

[36] A. Fellah, H. Jürgensen, and S. Yu. Constructions for alternating finite automata. *International journal of computer mathematics*, 35(1-4):117–132, 1990.

[37] N. Franzese, J. Katz, S. Lu, R. Ostrovsky, X. Wang, and C. Weng. Constant-overhead zero-knowledge for ram programs. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, pages 178–191, 2021.

[38] K. B. Frikken. Practical private dna string searching and matching through efficient oblivious automata evaluation. In *Data and Applications Security XXIII: 23rd Annual IFIP WG 11.3 Working Conference, Montreal, Canada, July 12-15, 2009. Proceedings 23*, pages 81–94. Springer, 2009.

[39] A. Gabizon and Z. J. Williamson. plookup: A simplified polynomial protocol for lookup tables. Cryptology ePrint Archive, Paper 2020/315, 2020. https://eprint.iacr.org/2020/315.

[40] A. Gabizon and Z. J. Williamson. Proposal: The turbo-PLONK program syntax for specifying SNARK programs. https://docs.zkproof.org/pages/standards/accepted-workshop3/proposal-turbo_plonk.pdf, 2020.

[41] R. Gennaro, C. Gentry, B. Parno, and M. Raykova. Quadratic span programs and succinct NIZKs without PCPs. In *Proceedings of the International Conference on the Theory and Applications of Cryptographic Techniques (EUROCRYPT)*, 2013.

[42] A. Golovnev, J. Lee, S. Setty, J. Thaler, and R. S. Wahby. Brakedown: Linear-time and post-quantum snarks for r1cs. *Cryptology ePrint Archive*, 2021.

[43] G. Gramlich and G. Schnitger. Minimizing nfas and regular expressions. In *STACS 2005: 22nd Annual Symposium on Theoretical Aspects of Computer Science, Stuttgart, Germany, February 24-26, 2005. Proceedings 22*, pages 399–411. Springer, 2005.

[44] L. Grassi, D. Khovratovich, C. Rechberger, A. Roy, and M. Schofnegger. Poseidon: A new hash function for zero-knowledge proof systems. In *USENIX Security Symposium*, volume 2021, 2021.

[45] P. Grubbs, A. Arun, Y. Zhang, J. Bonneau, and M. Walfish. Zero-knowledge middleboxes. In *Proceedings of the USENIX Security Symposium*, 2022.

[46] T. Hansen, D. Crocker, and P. Hallam-Baker. Domainkeys identified mail (DKIM) service overview. https://www.rfc-editor.org/rfc/rfc5585.html, 2009. RFC 5585.

[47] D. Hopwood. The Pasta curves for Halo 2 and beyond. https://electriccoin.co/blog/the-pasta-curves-for-halo-2-and-beyond/, 2020.

[48] S. Jarecki, H. Krawczyk, and J. Xu. Opaque: An asymmetric PAKE protocol secure against pre-computation attacks. In *Proceedings of the International Conference on the Theory and Applications of Cryptographic Techniques (EUROCRYPT)*, 2018.

[49] K. Jiang, D. Chait-Roth, Z. DeStefano, M. Walfish, and T. Wies. Less is more: refinement proofs for probabilistic proofs. Cryptology ePrint Archive, Paper 2022/1557, 2022. https://eprint.iacr.org/2022/1557.

[50] M. Jones, J. Bradley, and N. Sakimura. JSON web token (JWT). https://datatracker.ietf.org/doc/html/rfc7519, 2015. RFC 7519.

[51] A. R. Karlin, H. W. Trickey, and J. D. Ullman. Experience with a regular expression compiler. Technical report, STANFORD UNIV CA DEPT OF COMPUTER SCIENCE, 1983.

[52] A. Kate, G. M. Zaverucha, and I. Goldberg. Constant-size commitments to polynomials and their applications. In *International Conference on the Theory and Application of Cryptology and Information Security (ASIACRYPT)*, 2010.

[53] F. Kerschbaum. Practical private regular expression matching. In *IFIP International Information Security Conference*, pages 461–470. Springer, 2006.

[54] E. Kinnear, P. McManus, T. Pauly, T. Verma, and C. A. Wood. Oblivious DNS over HTTPS. https://www.rfc-editor.org/rfc/rfc9230, 2022. RFC 9230.

[55] A. Kothapalli and S. Setty. SuperNova: proving universal machine executions without universal circuits. Cryptology ePrint Archive, Paper 2022/1758, 2022. https://eprint.iacr.org/2022/1758.

[56] A. Kothapalli and S. Setty. HyperNova: recursive arguments for customizable constraint systems. *Cryptology ePrint Archive*, 2023.

[57] A. Kothapalli, S. Setty, and I. Tzialla. Nova: Recursive zero-knowledge arguments from folding schemes. In *Advances in Cryptology–CRYPTO 2022: 42nd Annual International Cryptology Conference, CRYPTO 2022, Santa Barbara, CA, USA, August 15–18, 2022, Proceedings, Part IV*, pages 359–388. Springer, 2022.

[58] P. Laud and J. Willemson. Universally composable privacy preserving finite automata execution with low online and offline complexity. *Cryptology ePrint Archive*, 2013.

[59] J. Lee. Dory: Efficient, transparent arguments for generalised inner products and polynomial commitments. In *Proceedings of the Theory of Cryptography Conference (TCC)*, 2021.

[60] C. Lund, L. Fortnow, H. Karloff, and N. Nisan. Algebraic methods for interactive proof systems. In *Proceedings of the IEEE Symposium on Foundations of Computer Science (FOCS)*, Oct. 1990.

[61] N. Luo, C. Weng, J. Singh, G. Tan, R. Piskac, and M. Raykova. Privacy-preserving regular expression matching using nondeterministic finite automata. Cryptology ePrint Archive, Paper 2023/643, 2023. https://eprint.iacr.org/2023/643.

[62] J. Miller, D. Waite, and M. Jones. JSON web proof. https://www.ietf.org/archive/id/draft-ietf-jose-json-web-proof-00.html, 2023.

[63] P. Mohassel, S. Niksefat, S. Sadeghian, and B. Sadeghiyan. An efficient protocol for oblivious dfa evaluation and applications. In *Topics in Cryptology–CT-RSA 2012: The Cryptographers' Track at the RSA Conference 2012, San Francisco, CA, USA, February 27–March 2, 2012. Proceedings*, pages 398–415. Springer, 2012.

[64] A. Nitulescu. SoK: Vector commitments. https://www.di.ens.fr/~nitulesc/files/vc-sok.pdf, 2021.

[65] S. Owens, J. Reppy, and A. Turon. Regular-expression derivatives re-examined. *Journal of Functional Programming*, 19(2):173–190, 2009.

[66] A. Ozdemir, F. Brown, and R. S. Wahby. Circ: Compiler infrastructure for proof systems, software verification, and more. In *2022 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2022.

[67] T. P. Pedersen. Non-interactive and information-theoretic secure verifiable secret sharing. In *Proceedings of the International Cryptology Conference (CRYPTO)*, 2001.

[68] N. Pippenger. On the evaluation of powers and related problems. In *Proceedings of the Annual Symposium on Foundations of Computer Science (SFCS)*, 1976.

[69] M. Raymond, G. Evers, J. Ponti, D. Krishnan, and X. Fu. Efficient zero knowledge for regular language. Cryptology ePrint Archive, Paper 2023/907, 2023. https://eprint.iacr.org/2023/907.

[70] C.-P. Schnorr. Efficient signature generation by smart cards. *Journal of Cryptology*, 4:161–174, 1991.

[71] S. Setty. Spartan: Efficient and general-purpose zkSNARKs without trusted setup. In *Proceedings of the International Cryptology Conference (CRYPTO)*, 2020.

[72] S. Setty, B. Braun, V. Vu, A. J. Blumberg, B. Parno, and M. Walfish. Resolving the conflict between generality and plausibility in verified computation. In *Proceedings of the ACM European Conference on Computer Systems (EuroSys)*, 2013.

[73] S. Setty, J. Thaler, and R. Wahby. Customizable constraint systems for succinct arguments. Cryptology ePrint Archive, Paper 2023/552, 2023. https://eprint.iacr.org/2023/552.

[74] S. Setty, V. Vu, N. Panpalia, B. Braun, A. J. Blumberg, and M. Walfish. Taking proof-based verified computation a few steps closer to practicality. In *Proceedings of the USENIX Security Symposium*, 2012.

[75] S. Shin, K. Kobara, and H. Imai. Security proof of AugPAKE. Cryptology ePrint Archive, Paper 2010/334, 2010. https://eprint.iacr.org/2010/334.

[76] T. Solberg. A brief history of lookup arguments. https://github.com/ingonyama-zk/papers/blob/main/lookups.pdf, 2023.

[77] C. Stanford, M. Veanes, and N. Bjørner. Symbolic boolean derivatives for efficiently solving extended regular expression constraints. In *Proceedings of the 42nd ACM SIGPLAN International Conference on Programming Language Design and Implementation*, pages 620–635, 2021.

[78] T. Taubert and C. A. Wood. SPAKE2+, an augmented PAKE. https://www.rfc-editor.org/rfc/internet-drafts/draft-bar-cfrg-spake2plus-08.html, 2022.

[79] K. Thompson. Programming techniques: Regular expression search algorithm. *Communications of the ACM*, 11(6), 1968.

[80] J. R. Troncoso-Pastoriza, S. Katzenbeisser, and M. Celik. Privacy preserving error resilient dna searching through oblivious automata. In *Proceedings of the 14th ACM conference on Computer and communications security*, pages 519–528, 2007.

[81] R. S. Wahby, I. Tzialla, A. Shelat, J. Thaler, and M. Walfish. Doubly-efficient zkSNARKs without trusted setup. In *Proceedings of the IEEE Symposium on Security and Privacy (S&P)*, 2018.

[82] C. Weng, K. Yang, X. Xie, J. Katz, and X. Wang. Mystique: Efficient conversions for zero-knowledge proofs with applications to machine learning. In *Proceedings of the USENIX Security Symposium*, 2021.

[83] T. Wu. The secure remote password protocol. In *Proceedings of the Network and Distributed System Security Symposium (NDSS)*, 1998.

[84] C. Zhang, Z. DeStefano, A. Arun, J. Bonneau, P. Grubbs, and M. Walfish. Zombie: Middleboxes that don't snoop. In *Proceedings of the USENIX Symposium on Networked Systems Design and Implementation (NSDI)*, 2024.

[85] J. Zhang, T. Xie, Y. Zhang, and D. Song. Transparent polynomial delegation and its applications to zero knowledge proof. In *Proceedings of the IEEE Symposium on Security and Privacy (S&P)*, 2020.

[86] L. Zhao, Q. Wang, C. Wang, Q. Li, C. Shen, and B. Feng. Veriml: Enabling integrity assurances and fair payments for machine learning as a service. *IEEE Transactions on Parallel and Distributed Systems*, 2021.

$r,s ::=$	$\emptyset$	Empty set
$\mid$	ϵ	Empty string
$\mid$	$\mathcal{C}$	Non-empty character set ($\emptyset \subset \mathcal{C} \subseteq \Sigma$)
$\mid$	rs	concatenation
$\mid$	$r + s$	Logical or (alternation)
$\mid$	$r \,\&\, s$	Logical and (conjunction)
$\mid$	r^*	Kleene-closure
$\mid$	$r\{n, m\}$	Bounded repetition ($n, m \in \mathbb{N}$)

FIGURE 9—Low-level Regex syntax in Reef. Additionally define the wildcard notation . to be the full character set Σ and n-repetition as $r\{n\} = r\{n, n\}$.

A Preliminaries

A.1 Monoids

A monoid is a triple $(A, \cdot, \epsilon)$ where A is the carrier set, $\cdot$ is the *append* operation, and ϵ is the identity element of append, such that the monoid equations apply.

- Associativity $a \cdot (b \cdot c) = (a \cdot b) \cdot c$.
- Left-identity $\epsilon \cdot a = a$.
- Right-identity $a \cdot \epsilon = a$.

A.2 Boolean Algebras

A boolean algebra (or boolean lattice) is a 6-tuple $(A, \top, \bot, \wedge, \vee, \neg)$ where A is the carrier set, $\top \in A, \bot \in A$ represent the *true* and *false* booleans.

The binary combinators $\wedge, \vee$ correspond to conjunction and disjunction respectively, and the unary $\neg$ corresponds to negation. A boolean algebra is closed in A under $\wedge, \vee, \neg$ and has the following equations.

- Associativity of $\vee, \wedge$.
- Commutativity of $\vee, \wedge$.
- Distributivity of $\wedge$ over $\vee$ and $\vee$ over $\wedge$.
- $\bot$ the unit of $\vee$.
- $\top$ the unit of $\wedge$.
- Annihilation for $\vee, \top$ and $\wedge, \bot$ respectively.
- Idempotence of $\vee, \wedge$. item Complement rules for $\vee, \wedge$ and $\neg$.

A.3 Kleene Algebras

A *Kleene Algebra* over carrier set A is the 8-tuple $(A, \top, \bot, \wedge, \vee, \neg, \cdot, \epsilon)$, where A is a Monoid $(A, \epsilon, \cdot)$ and A is also a Boolean algebra $(A, \top, \bot, \wedge, \vee, \neg)$. Additionally, the distributivity laws describe the interactions of $\wedge, \vee, \cdot$.

$$
\begin{aligned}
(a \wedge b) \cdot x &= (a \cdot x) \wedge (b \cdot x) \\
(a \vee b) \cdot x &= (a \cdot x) \vee (b \cdot x) \\
x \cdot (a \wedge b) &= (x \cdot a) \wedge (x \cdot b) \\
x \cdot (a \vee b) &= (x \cdot a) \vee (x \cdot b)
\end{aligned}
$$

In the proofs that follow, we take advantage of the fact that regular expressions form a Kleene algebra, we also show that *skips* over regions of the string also form a Kleene Algebra, as do the automata we introduce—*Skipping Alternating Finite Automata* (SAFA). We use this equivalence to Kleene algebras to prove SAFA are regular.

B Regular Expressions

The regular expression syntax in Reef is almost the one from Owens et al. [65], with the addition of *bounded repetition*, to preserve bounded skips for the SAFA compiler. The bounded repetition does not change the regular nature of the source language, as $r\{a, b\} = ra|ra+1|\ldots|rb$ which is regular, but for large a, b can grow quickly.

Given an alphabet Σ, the language accepted by regex r is defined as $\mathcal{L}[\![r]\!]$ in Figure 10).

Definition B.1.

$$
\begin{aligned}
\mathcal{L}[\![\emptyset]\!] &= \emptyset \\
\mathcal{L}[\![\epsilon]\!] &= \{\epsilon\} \\
\mathcal{L}[\![C]\!] &= C \text{ (where } C \text{ is a character class)} \\
\mathcal{L}[\![r \cdot s]\!] &= \{uv \mid u \in \mathcal{L}[\![r]\!],\ v \in \mathcal{L}[\![s]\!]\} \\
\mathcal{L}[\![r + s]\!] &= \mathcal{L}[\![r]\!] \cup \mathcal{L}[\![s]\!] \\
\mathcal{L}[\![r \& s]\!] &= \mathcal{L}[\![r]\!] \cap \mathcal{L}[\![s]\!] \\
\mathcal{L}[\![r^*]\!] &= \{\epsilon\} \cup \mathcal{L}[\![r \cdot r^*]\!] \\
\mathcal{L}[\![r\{n, m\}]\!] &= \begin{cases} \{\epsilon\} & \text{if } n = m = 0 \\ \mathcal{L}[\![r \cdot r\{0, m-1\}]\!] & \text{if } n = 0 \\ \mathcal{L}[\![r \cdot r\{n-1, m-1\}]\!] & \text{otherwise} \end{cases}
\end{aligned}
$$

FIGURE 10—The set of strings accepted by a regex r is the *language* $\mathcal{L}[\![r]\!] \subseteq \Sigma^*$.

B.1 Derivatives of regular expressions

Brzozowzki [26] defined the derivative of a regex r given a character $\alpha \in \Sigma$ as $d_\alpha(r)$, as another regular expression such that its language $\mathcal{L}[\![d_\alpha(r)]\!]$ contains all the suffixes $w \subseteq \Sigma^*$ of $\mathcal{L}[\![r]\!]$ with prefix α.

$$\mathcal{L}[\![r]\!] = \{\alpha w \mid w \in \mathcal{L}[\![d_\alpha(r)]\!]\}$$

Regex derivatives are the workhorse of the SAFA compiler. Before we can define regex derivatives for SAFAs, we follow Brzozowki's presentation and introduce the *nullable* predicate $v(r)$ in Figure 11. The predicate $v(r)$ is true if and only if the regex r accepts the empty string. Nullable regex correspond to *accepting* states in finite automata. We use $v(r)$ in Definitions 12, 13 to check if for a regex $r \cdot s$ the derivative $d_\alpha(r \cdot s)$ should be applied to r or s.

Definition B.2.

$$
\begin{aligned}
v(\epsilon) &= \mathit{true} \\
v(r^*) &= \mathit{true} \\
v(\emptyset) &= \mathit{false} \\
v(C) &= \mathit{false} \\
v(rs) &= v(r) \wedge v(s) \\
v(r+s) &= v(r) \vee v(s) \\
v(r\&s) &= v(r) \wedge v(s) \\
v(r\{n,m\}) &= \begin{cases} \mathit{true} & \text{if } n = 0 \\ v(r) & \text{otherwise} \end{cases}
\end{aligned}
$$

FIGURE 11—The predicate $v(r)$ is true when the regex r accepts the empty string

$$
\begin{aligned}
d_\alpha(\emptyset) &= \emptyset \\
d_\alpha(\epsilon) &= \emptyset \\
d_\alpha(C) &= \begin{cases} \epsilon & \text{if } \alpha \in C \\ \emptyset & \text{otherwise} \end{cases} \\
d_\alpha(r \cdot s) &= \begin{cases} (d_\alpha(r)s) \mid d_\alpha(s) & \text{if } v(r) \\ d_\alpha(r)s & \text{otherwise} \end{cases} \\
d_\alpha(r+s) &= d_\alpha(r) + d_\alpha(s) \\
d_\alpha(r\&s) &= d_\alpha(r) \& d_\alpha(s) \\
d_\alpha(r^*) &= d_\alpha(r) r^* \\
d_\alpha(r\{n,m\}) &= \begin{cases} \emptyset & \text{if } n = m = 0 \\ d_\alpha(r \cdot r\{0, m-1\}) & \text{if } n = 0 \\ d_\alpha(r \cdot r\{n-1, m-1\}) & \text{otherwise} \end{cases}
\end{aligned}
$$

FIGURE 12—The Brzozowski derivative for a regex r given character $\alpha \in \Sigma$ is $d_\alpha(r)$.

Given the nullable predicate $v(r)$, the Brzozowski derivative [26] [65] $d_\alpha(r)$ for a character $a \in \Sigma$ is defined as the regex matches string D given r matches initial string $\alpha \cdot D$, see Figure 12.

Antimirov imporves on regex deriviatives with *Partial Regex derivatives* [20], by observing the upper semilattice $(R, +, \emptyset)$ where R are sets of regex, provides us with the Associativity, Commutativity, Idempotence, and Zero-element laws already, from their set structure.

Caron et al. [31] generalize Antimirov's partial derivatives from sets to arbitrary *support structures*, a significant generalization. We define *Generalized Antimirov partial derivatives* $\partial_\alpha : \mathit{regex} \to B^+(\mathit{regex})$ as a function, given a character $\alpha \in \Sigma$, returns a *positive* boolean algebra over regex $B^+(\mathit{regex})$, meaning any and/or expression (no negation $\neg$) of regex.

In practice *Alternating Finite Automata (AFA)* which we introduce later as the basis of SAFA, alternate between *and* ($\wedge$) expressions and *or* ($\vee$) expressions at different states, so each state is either an $\wedge$ state or an $\vee$ state, a subset of $B^+(\mathit{regex})$.

Here's our definition of generalized partial derivative $\partial_\alpha(r)$ given character $\alpha \in \Sigma$ in Figure 13.

Definition B.3.

$$
\begin{aligned}
\partial_\alpha(\emptyset) &= \bot \\
\partial_\alpha(\epsilon) &= \bot \\
\partial_\alpha(C) &= \begin{cases} \epsilon & \text{if } \alpha \in C \\ \bot & \text{otherwise} \end{cases} \\
\partial_\alpha(r \cdot s) &= \begin{cases} \partial_\alpha(s) \vee \partial_\alpha(r) \cdot s & \text{if } v(r) \\ \partial_\alpha(r) \cdot s & \text{otherwise} \end{cases} \\
\partial_\alpha(r+s) &= \partial_\alpha(r) \vee \partial_\alpha(s) \\
\partial_\alpha(r\&s) &= \partial_\alpha(r) \wedge \partial_\alpha(s) \\
\partial_\alpha(r^*) &= \partial_\alpha(r) \cdot r^* \\
\partial_\alpha(r\{n,m\}) &= \begin{cases} \bot & \text{if } n = m = 0 \\ \partial_\alpha(r \cdot r\{0, m-1\}) & \text{if } n = 0 \\ \partial_\alpha(r \cdot r\{n-1, m-1\}) & \text{otherwise} \end{cases}
\end{aligned}
$$

FIGURE 13—The generalized partial regex derivative $\partial_\alpha(r)$ is a boolean expression of other regex.

B.2 Existing Automata

Let us review some basic automata theory which will be used later in proofs and as a useful step to understand the SAFA construction.

Start with *Nondeterministic Finite Automata* (NFA), a 5-tuple $(Q, \Sigma, q_0, \delta, \mathcal{F})$ in Figure 14. NFA start at an initial state q_0 and transition *nondeterministically* to subsequent states in Q, using the transition relation δ to determine the next step. Note the special ϵ nondeterministic transition, which does not consume a character and allows the NFA to arbitrarily (nondeterministically) take those transitions. This use of nondeterminism is crucial for the performance of Reef, as those choices correspond to existential witnessses.

An AFA [32] is a 6-tuple $(Q, \Sigma, q_0, \lambda_q, \delta, \mathcal{F})$ in Figure 15 which generalizes nondeterminism (the $\exists$ quantifier states) with the addition of the $\forall$ quantifier states. When a state is an $\exists$ state, based on the labeling function λ_q, it works like an NFA, any one descendant of state accepts means the whole state accepts. Dually, a $\forall$ state accepts if and only if *all* descendants of the state accept the string.

Q	:	the set of all states
$\Sigma \cup \{\epsilon\}$	:	The alphabet with ϵ (empty string)
$q_0 \in Q$	:	Initial state
$\delta \subseteq Q \times (\Sigma \cup \{\epsilon\}) \times Q$	:	Transition relation
$\mathcal{F} \subseteq Q$	:	Set of accepting states

FIGURE 14—Noneterministic Finite Automata (NFA) can chose ϵ transitions.

Q	:	the set of all states
Σ	:	The alphabet
$q_0 \in Q$	:	Initial state
$\lambda_q : Q \to \{\forall,\ \exists\}$	:	Label states $\forall$ or $\exists$
$\delta \subseteq Q \times \Sigma \times Q$	:	Transition relation
$\mathcal{F} \subseteq Q$	:	Set of accepting states

FIGURE 15—An AFA alternates between existential and universal states.

C Skipping Alternating Finite Automata

AFA introduce existential/universal states that allow Reef to represent both alternation and look-aheads respectively. Next, to efficiently represent sparse matches we introduce an automaton that can ignore irrelevant parts of the string.

Consider the regex `^.{1,1000}ab$` given string D, it means there is a cursor $1 \leq i \leq 1000$ and $\mathsf{D}_i = a$ and $\mathsf{D}_{i+1} = b$. The verifier has a nondeterministic choice of i at this point, as long as the remaining $\mathsf{D}_{(i,i+1)}$ characters match. We skip verifying those 1000 characters and only verify the inequality $1 \leq i \leq 1000$ instead.

We must now formally define the notion of *skip*, which was described informally so far. The properties of skips we look for is, short representation of sets of possible natural numbers, fast membership checks, and compositionality. Specifically we want not only the usual set combinators $\cup, \cap, \neg$ but also concatenation of skips $s + t$, for example in the regex `^.{2,3}.a$` we can see two wildcards compose and we get the equivalent skip `.{3,4}`. The datastructure we use to represent skips are *Interval sets*, a set of disjoint, ordered intervals of natural numbers.

C.1 Intervals

A *bounded interval* $[a, b]$ where $a \leq b,\ a, b \in \mathbb{N}$ represents the subset of natural numbers $\{\ i \mid a \leq i \leq b\}$, inclusive in both ends.

An *unbounded interval* $[a, \infty)$ represents the subset of natural numbers $\{\ i \mid a \leq i\}$, inclusive on the left, unbounded on the right. We represent intervals with the letters I, i.

C.2 Skips/Interval sets

An *interval set S* or a *skip* is a set datastructure, a collection of intervals $\{i_1, \ldots, i_n\}$ ordered by increasing starting point, such that there is no *overlap* between consecutive intervals.

Interval sets allow us to represent continuous ranges of natural numbers, for example the single interval $\{[a, b]\}$ where $a \leq b,\ a, b \in \mathbb{N}$, as well as disjoint sets, for example the set of numbers less-than-equal to a and greater-than-equal b as $\{\ [0, a], [b, \infty)\}$, assuming $a \leq b$. A right-open interval simply means the end-of-file determines when to stop skipping.

We consider an overlap of two intervals to be a difference of at most 1 between the end-point of the first and the start point of the second. So for example $[1, 2], [4, 5]$ do not overlap but both $[1, 3], [2, 4]$ and $[1, 2], [3, 4]$ overlap and are both equivalent to $[1, 4]$. Two intervals $[a_1, b_1],\ [a_2, b_2]$ overlap if $\max(a_1, b_1) + 1 \geq \min(a_2, b_2)$.

Intervals and Interval sets admit the familiar set operations *union* ($\cup$), *intersection* ($\cap$), *complement* ($\neg$) and additionally the operation *append*($+$), which is element-wise addition of interval bounds. We use set notation $i \in s$ and $n \in s$, to mean i is an interval in s, or n is a number contained in one of the intervals of s.

C.3 Operations on Interval sets

We define the combinators $\cup, \cap, \neg$ on intervals on Figure 16,17,18,19 as a preparation for defining the boolean closures of Interval sets. Notice $\cup, \cap, \neg$ on Intervals are not closed, they return an Interval Set S and as such, intervals do *not* form a boolean algebra. But this is fine as we can recover boolean closure for interval sets with these definition.

$$[a_1, b_1]\ \cup [a_2, b_2] = \begin{cases} \{[min(a_1, a_2),\ max(b_1, b_2)]\} & max(a_1, a_2) \leq min(b_1, b_2) \\ \{[a_2, b_2], [a_1, b_1]\} & b_2 < a_1 \\ \{[a_1, b_1], [a_2, a_2]\} & \text{otherwise} \end{cases}$$

$$[a_1, \infty)\ \cup [a_2, b_2] = \begin{cases} \{[a_2, b_2], [a_1, \infty)\} & b_2 < a_1 \\ \{[a_1, \infty)\} & b_2 \geq a_1 \leq a_2 \\ \{[a_2, \infty)\} & \text{otherwise} \end{cases}$$

$$[a_1, b_1]\ \cup [a_2, \infty) = \begin{cases} \{[a_1, b_1], [a_2, \infty)\} & b_1 < a_2 \\ \{[a_2, \infty)\} & b_1 \geq a_2 \leq a_1 \\ \{[a_1, \infty)\} & \text{otherwise} \end{cases}$$

$$[a_1, \infty)\ \cup [a_2, \infty) = \begin{cases} \{[a_1, \infty)\} & a_1 \leq a_2 \\ \{[a_2, \infty)\} & \text{otherwise} \end{cases}$$

FIGURE 16—The union of intervals is an interval set.

$$[a_1, b_1] \cap [a_2, b_2] = \begin{cases} \{[max(a_1,a_2),\ min(b_1,b_2)]\} & max(a_1,a_2) \le min(b_1,b_2) \\ \{\} & \text{otherwise} \end{cases}$$

$$[a_1, \infty) \cap [a_2, b_2] = \begin{cases} \{[a_2,b_2]\} & b_2 \ge a_1 \le a_2 \\ \{[a_1,b_2]\} & b_2 \ge a_1 > a_2 \\ \{\} & \text{otherwise} \end{cases}$$

$$[a_1, b_1] \cap [a_2, \infty) = \begin{cases} \{[a_1,b_1]\} & b_1 \ge a_2 \le a_1 \\ \{[a_2,b_1]\} & b_1 \ge a_2 > a_1 \\ \{\} & \text{otherwise} \end{cases}$$

$$[a_1, \infty) \cap [a_2, \infty) = \begin{cases} \{[a_2,\infty)\} & a_1 \le a_2 \\ \{[a_1,\infty)\} & \text{otherwise} \end{cases}$$

FIGURE 17—The intersection of intervals is an interval set.

$$\begin{aligned} \neg\, [0,\ \infty) &= \{\} \\ \neg\, [a,\ \infty) &= \{[0, a-1]\} \\ \neg\, [0,\ a] &= \{[a+1, \infty]\} \\ \neg\, [a,\ b] &= \{[0, a-1], [b+1, \infty)\} \end{aligned}$$

FIGURE 18—The complement of intervals is an interval set.

Intervals can also be combined with the *append*$(+)$ operator and $\epsilon = [0, 0]$ as the identity element, forming a monoid in Figure 19.

$$\begin{aligned} [a_1,\ \infty) + [a_2, \infty) &= [a_1 + a_2, \infty) \\ [a_1,\ b_1] + [a_2, \infty) &= [a_1 + a_2, \infty) \\ [a_1,\ \infty) + [a_2, b_2] &= [a_1 + a_2, \infty) \\ [a_1,\ b_1] + [a_2, b_2] &= [a_1 + a_2, b_1 + b_2] \end{aligned}$$

FIGURE 19—Intervals are closed under *append*.

Interval set combinators in Figure 20 are then build up from the interval combinators. They are closed under boolean and monoidal operations and form a Kleene algebra A.3, with $\top = \{[0, \infty)\}, \bot = \{\}$ and combinators $\vee = \cup, \wedge = \cap$ and $\neg$ and monoidal combinator $+$ and ϵ the identify element. The Kleene Algebra properties hold by induction on the length of interval sets and taking cases for any possible interval (bounded/unbounded). In the Reef implementation, we additionally check the Kleene Algebra properties hold for interval sets using property-based tests.

Interval sets are also sets of natural numbers so we use set notation $i \in s$ to indicate membership of natural number i checked in sublinear time, proportional to the number of disjoint intervals in the Interval set.

We can now formally define a finite automaton that takes advantage of the skip structure. SAFA are a generalization of AFA with *skips* on their transitions.

C.4 SAFA formal definition

A Skipping Alternating Finite Automaton (SAFA) is an 8-tuple $(Q, E, \Sigma, q_0, bq, \lambda_q, \lambda_e, \delta, \mathcal{F})$ in Figure 21.
The only difference with the AFA definition is the addition edge labeling function λ_e. Transitions in a SAFA can be labeled as either a character set $C \subseteq \Sigma$ which match a character $\mathsf{D}_i \in C$, for document D, or a skip s, which does not consume a character, but increases the cursor nondeterministically by any $n \in s$.

We overload the notation $(q, s, q') \in \delta$ to indicate a skip transition s, or more precisely there exists an edge $e \in E$ such that $(q, e, q') \in \delta$ and $\lambda_e(e) = s$. Also overload $(q, \alpha, q') \in \delta$ to indicate a character $\alpha \in \Sigma$ transition, or there exists an edge $e \in E$ and character set $C \subseteq \Sigma$ such that $(q, e, q') \in \delta$ and $\lambda_e(e) = C$ and $\alpha \in C$. As the labeling function λ_e maps to a disjoint union there is no confusion with this notation.

C.5 SAFA Semantics

Now we can give the semantics of the language accepted by a SAFA $\mathcal{S}$ in Figure 22. If $\mathsf{D} \in \Sigma^*$ is a string with random-access, and $i \le |\mathsf{D}|$ a cursor in the string, define the mutually-recursive decidable procedure $\text{match}_{\mathcal{S}}$ which returns *true* if at state q, a string D at index i is accepted by SAFA $\mathcal{S}$.

With the auxilary definitions of Figure 22 in place, finally define the language recognized by a SAFA $\mathcal{S}$ as

$$\mathcal{L}[\![\mathcal{S}]\!] = \{\mathsf{D} \mid \text{match}_{\mathcal{S}}(q_0, \mathsf{D}, 0)\}$$

$$\begin{aligned} \{i,\ \overline{i_n}\} \cup s &= \{i \cup i_s \mid i_s \in s\} \cup \{\overline{i_n}\} \\ \{\} \cup s &= s \\ \\ \{i,\ \overline{i_n}\} \cap s &= \{i \cap i_s \mid i_s \in s\} \cap \{\overline{i_n}\} \\ \{\} \cap s &= \{\} \\ \neg\{i_1, \overline{i_n}\} &= \neg i_i \cap \neg\{\overline{i_n}\} \\ \neg\{\} &= \{[0, \infty)\} \\ \\ s_1 + s_2 &= \bigcup \{i_1 + i_2 \mid i_1 \in s_1,\ i_2 \in s_2\} \end{aligned}$$

FIGURE 20—Boolean and monoidal closure operations on interval sets.

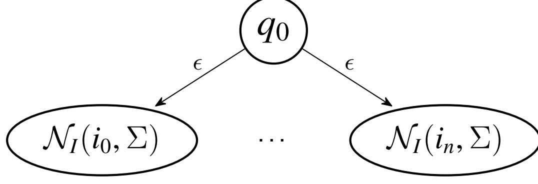


FIGURE 23—NFA $\mathcal{N}(s, \Sigma)$ for interval set $s = \{\overline{i_n}\}$.

Q	:	the set of all states (nodes)
E	:	the set of all transitions (edges)
Σ	:	The alphabet
$q_0 \in Q$	:	Initial state
$\lambda_q : Q \to \{\forall,\ \exists\}$	:	Label nodes, $\forall$ or $\exists$
$\lambda_e : E \to S \uplus C$	:	Label edges, skip or character set
$\delta \subseteq Q \times E \times Q$	:	Transition relation
$\mathcal{F} \subseteq Q$	:	Set of accepting states

FIGURE 21—SAFA formal definition over alphabet Σ generalizes AFA with skip edges.

Definition C.1.

$$\text{match}_{\mathcal{S}}(q, \mathsf{D}, i) \triangleq \begin{cases} q \in \mathcal{F} \land i = |\mathsf{D}| & \text{(accept condition)} \\ \text{match}_{\forall}(q, \mathsf{D}, i) & \text{if } \lambda_q(q) = \forall \\ \text{match}_{\exists}(q, \mathsf{D}, i) & \text{if } \lambda_q(q) = \exists \end{cases} \tag{1}$$

$$\text{match}_{\forall}(q, \mathsf{D}, i) \triangleq \forall\, e, q',\ (q, e, q') \in \delta \to \text{match}_E(q', e, \mathsf{D}, i) \tag{2}$$

$$\text{match}_{\exists}(q, \mathsf{D}, i) \triangleq \exists\, e, q',\ (q, e, q') \in \delta \land \text{match}_E(q', e, \mathsf{D}, i) \tag{3}$$

$$\text{match}_E(q', e, \mathsf{D}, i) \triangleq \begin{cases} \text{match}_{\mathcal{S}}(q', \mathsf{D}, i+1) & \text{if } \lambda_e(e) = \mathsf{D}[i] \\ \exists n \in s,\ \text{match}_{\mathcal{S}}(q', \mathsf{D}, i+n) & \text{if } \lambda_e(e) = s \\ \textit{false} & \text{otherwise} \end{cases} \tag{4}$$

FIGURE 22—SAFA semantics, mutually recursive predicate $\text{match}_{\mathcal{S}}$ is *true* iff string D at position $i \leq |\mathsf{D}|$ is accepted by SAFA $\mathcal{S}$.

C.6 SAFA are regular

To prove SAFA are regular and have the same computational power as the source regex language, we must provide an isomorphism from SAFA to another finite automaton which is known to be regular, like DFA, NFA or AFA, as well as an isomorphism to our source language to show that no expressive power is lost during compilation.

Source language $\iff$ DFA. The source language in Figure 9 is regular by giving a translation to the regular expression language with character sets [65], which is shown to be regular by equivalence to a DFA. Our *bounded range* expressions $r\{a, b\}$ in Figure 9 translate to an alternation of finite repetition $r\{a, b\} = r\{a\} + \ldots + r\{b\}$ where

$$r\{0\} = \epsilon$$
$$r\{i\} = r \cdot r\{i-1\}$$

DFA $\iff$ AFA. A construction is given by Fellah et al. [36]. The proof proceeds by constructing an intermediate NFA where states are $Q \times Q$ sized boolean matrices which correspond to truth tables of $B^+(Q)$ of AFA states. By the above, a n-state AFA is equivalent to an at most 2^n-state NFA and a 2^{2^n} DFA by the product construction [36]. The opposite direction is trivial, as all DFA are AFA with only existential nodes.

AFA $\iff$ SAFA. Finally, we show SAFA is regular by translation to AFA. We give a translation $\mathcal{N}(s, \Sigma)$ of skip s to NFA, given alphabet Σ. Then show substituting $\mathcal{N}(s, \Sigma)$ in place of s produces an AFA which recognizes the same language as the SAFA. The opposite direction, embedding an AFA to a SAFA is trivial; all AFA are SAFA without skips.

A skip is a finite set of intervals (Section C.2), whose union is a possibly infinite subset of the natural numbers $s = \{\overline{i_n}\} \subseteq \mathbb{N}$. Before we give an NFA construction for s we construct a NFA $\mathcal{N}_I(i, \Sigma)$ for an interval i and alphabet Σ.

Intervals to DFA. An NFA for a closed interval $[a, b]$ (where $a \leq b$) is given by the following construction

Definition C.2.

$$\begin{aligned} \mathcal{N}_c(a, b, \Sigma) = (& \\ & Q := \{q_1, q_2, \ldots, q_b\} \\ & \Sigma := \Sigma \\ & q_0 := q_1 \\ & \mathcal{F} := \{q_a, \ldots, q_b\} \\ & \delta := \{q_1 \xrightarrow{\Sigma} q_2, q_2 \xrightarrow{\Sigma} q_3, \ldots, q_{b-1} \xrightarrow{\Sigma} q_b\} \\ &) \end{aligned}$$

The NFA $N_{[a,b]} = \mathcal{N}_c(a, b, \Sigma)$ is a finite chain with accepting states $q_i,\ a \leq i \leq b$. Notice the final state q_b is an accepting state but does not transition. We will later give a substitution function of an NFA in a SAFA $\mathcal{S}$ edge $q_A \xrightarrow{s} q_B$ which adds an ϵ-transition from q_b to SAFA state q_B.

The $\mathcal{N}_c$ construction will not work for an open interval $i = [a, \infty)$ as it will result in an infinite chain of states. Instead, construct an NFA with an accepting self-loop as the last state.

Definition C.3.

$$
\begin{aligned}
&\mathcal{N}_o(a,\Sigma) = (\\
&\quad Q := \{q_1, q_2, \ldots, q_a\} \\
&\quad \Sigma := \Sigma \\
&\quad q_0 := q_1 \\
&\quad \mathcal{F} := \{q_a\} \\
&\quad \delta := \{q_1 \xrightarrow{\Sigma} q_2, q_2 \xrightarrow{\Sigma} q_3, \ldots, q_{b-1} \xrightarrow{\Sigma} q_b, q_a \xrightarrow{\Sigma} q_a\} \\
&\quad)
\end{aligned}
$$

Note the reflexive transition (q_a, Σ, q_a) on the only accepting state q_a means all a states must be traversed to accept. Now combine $\mathcal{N}_c(a,b,\Sigma)$ and $\mathcal{N}_o(a,\Sigma)$ to construct an NFA equivalent to an arbitrary interval, by conditionally branching on if the interval is closed or open.

Definition C.4.

$$
\mathcal{N}_I(i,\Sigma) = \begin{cases} \mathcal{N}_c(a,b,\Sigma) & \text{if } i = [a,b] \\ \mathcal{N}_o(a,\Sigma) & \text{if } i = [a,\infty) \end{cases}
$$

Interval sets to NFA. Given skip $s = \{\overline{i_n}\}$ and alphabet Σ construct an $n+1$ state NFA $\mathcal{N}(s,\Sigma)$ as follows. Assume no capturing of state identifiers—state $q \in Q_{\mathcal{N}_I(i,\Sigma)}$ does not appear in another $\mathcal{N}_I(i',\Sigma)$ where $i \neq i'$.

Definition C.5.

$$
\begin{aligned}
&\mathcal{N}(s,\Sigma) = (\\
&\quad Q := \{q_0\} \cup \bigcup_{i \in s} Q_{\mathcal{N}_I(i,\Sigma)} \\
&\quad \Sigma := \Sigma \\
&\quad q_0 := q_0 \\
&\quad \mathcal{F} := \bigcup_{i \in s} \mathcal{F}_{\mathcal{N}_I(i,\Sigma)} \\
&\quad \delta := \bigcup_{i \in s} \{(q_0, \epsilon, q_{0,\mathcal{N}(i,\Sigma)})\} \cup \bigcup_{i \in s} \delta_{\mathcal{N}(i,\Sigma)} \\
&\quad)
\end{aligned}
$$

The initial state q_0 is non-accepting. Then add a nondeterministic choice from q_0 to each one of $\mathcal{N}_I(i_n,\Sigma)$. As long as any $\mathcal{N}_I(i_n,\Sigma)$ reaches an accepting state, then $\mathcal{N}(s,\Sigma)$ accepts. This concludes the construction or $\mathcal{N}(s,\Sigma)$.

Substitution in SAFA. Now define formally a substitution procedure $\langle \mathcal{N}/e \rangle_{\mathcal{S}}$, for a SAFA $\mathcal{S}$, NFA $\mathcal{N}$, and SAFA edge $e \in E_{\mathcal{S}}$.

Remember that SAFA edges are a set $E_{\mathcal{S}}$ and map to either skips or character-sets via the labeling function $\lambda_e : E \to S \uplus C$. We build an edge set $E_{\mathcal{N}}$ and labeling function $\lambda_{\mathcal{N}} : E_{\mathcal{N}} \to C$ for NFA $\mathcal{N}$, such that

- No capturing: $E_{\mathcal{S}}$: $E_{\mathcal{N}} \cap E_{\mathcal{S}} = \emptyset$.
- $E_{\mathcal{N}}$ sound: $\forall e \in E_{\mathcal{E}},\ \exists q\ q'\ (q, \lambda_{\mathcal{N}}(e), q') \in \delta_{\mathcal{N}}$.
- $E_{\mathcal{N}}$ complete: $\forall q\ q'\ C,\ (q, C, q') \in \delta_{\mathcal{N}} \to \exists e \in E_{\mathcal{N}},\ \lambda_{\mathcal{N}}(e) = C$.

Also assume no capturing of states, $Q_{\mathcal{S}} \cap Q_{\mathcal{N}} = \emptyset$ are disjoint and their alphabets are equal. Let unique q_{src}, q_{dst}, such that $(q_{src}, e, q_{src}) \in \delta_{\mathcal{S}}$

Definition C.6.

$$
\begin{aligned}
&\langle \mathcal{N}/e_s \rangle_{\mathcal{S}} = (\\
&\quad Q := Q_{\mathcal{N}} \cup Q_{\mathcal{S}} \setminus \{q_{src},\ q_{dst}\} \\
&\quad E := E_{\mathcal{N}} \cup E_{\mathcal{S}} \setminus \{e_s\} \\
&\quad \Sigma := \Sigma_{\mathcal{S}} \\
&\quad \lambda_q := \lambda_{q,\mathcal{S}} \cup \{q \mapsto \exists \mid q \in Q_{\mathcal{N}}\} \\
&\quad \lambda_e := \lambda_{e,\mathcal{S}} \cup \lambda_{\mathcal{N}} \\
&\quad \delta := \delta_{\mathcal{S}} \\
&\qquad \cup \{(q, e, q') \mid (q, C, q') \in \delta_{\mathcal{N}},\ C \subseteq \Sigma,\ \lambda_{\mathcal{N}}(e) = C\} \\
&\qquad \cup \{(q_{src}, e_\epsilon, q_{0,\mathcal{N}})\} \cup \{(q_F, e_\epsilon, q_{dst}) \mid q_F \in \mathcal{F}_{\mathcal{N}}\} \\
&\quad \mathcal{F} := \mathcal{F}_{\mathcal{S}} \\
&\quad)
\end{aligned}
$$

The substitution construction may look complex but is intuitive to understand. For SAFA $\mathcal{S}$, NFA $\mathcal{N}$ and SAFA edge e_s with transition $(q_{src}, e, q_{src}) \in \delta_{\mathcal{S}}$, remove the transition (q_{src}, e, q_{src}) from $\delta_{\mathcal{S}}$ and replace it with $\mathcal{N}$ by connecting q_{src} and q_{dst} to the initial and accepting states of $\mathcal{N}$ respectively.

Notice the accepting states $\mathcal{F}_{\mathcal{S}}$ of $\mathcal{S}$ do not change, so if we can prove language equivalence of NFA $\mathcal{N}$ with the replaced edge e_s, we should be able to prove a key lemma for the SAFA $\iff$ AFA proof. The definition of a language for an edge $e \in E_{\mathcal{S}}$ comes directly from the match$_E$ rule in Figure 22.

Definition C.7.

$$
\mathcal{L}[\![e_{\mathcal{S}}]\!] = \begin{cases} C & \text{if } \lambda_e(e_{\mathcal{S}}) = C \subseteq \Sigma \\ \{\Sigma^n \mid n \in s\} & \text{if } \lambda_e(e_{\mathcal{S}}) = s \end{cases}
$$

Either the edge e maps to a character set $C \subseteq \Sigma$ and the language is all the single characters in the set C, or a skip s and the language is all the n-length strings, for every $n \in s$.

The language of an NFA is the textbook definition, where δ^* is the transitive-reflexive closure of the δ relation.

Definition C.8.

$$
\mathcal{L}[\![\mathcal{N}]\!] = \{w \mid (q_0, w, q_F) \in \delta^*_{\mathcal{N}}\}
$$

Define the language *suffix* at state $q \in Q_{\mathcal{S}}$ for a SAFA as a generalization of $\mathcal{L}[\![\mathcal{S}]\!]$ to a given start state $q \in Q_{\mathcal{S}}$ instead of initial state q_0. This definition gives us a strong induction hypothesis to use in the following lemma.

Definition C.9.

$$\mathcal{L}[\![\mathcal{S} : q]\!] = \{D \mid \text{match}_{\mathcal{S}}(q, D, 0)\}$$

Taking a transition $e \in E_{\mathcal{S}}$ prepends the language $\mathcal{L}[\![e_{\mathcal{S}}]\!]$ to all the suffixes of the destination of e. Note the concatenation operator $\cdot$ is overloaded, to mean the pairwise concatenation of the product of two sets.

Lemma C.1. $(q, e, q') \in \delta_{\mathcal{S}} \rightarrow \mathcal{L}[\![\mathcal{S} : q]\!] = \mathcal{L}[\![e_{\mathcal{S}}]\!] \cdot \mathcal{L}[\![\mathcal{S} : q']\!]$

Proof. The proof proceeds by induction on the derivation $\mathcal{L}[\![\mathcal{S} : q]\!]$. In the base and the inductive case, perform case analysis on $\lambda_e(e_{\mathcal{S}})$.

1. If $\lambda_e(e_{\mathcal{S}}) = C \subseteq \Sigma$ then the language $\mathcal{L}[\![e_{\mathcal{S}}]\!] = C$. Prepend each character in the character set C to $\mathcal{L}[\![\mathcal{S} : q']\!]$. The strings $\mathsf{D} \in C \cdot \mathcal{L}[\![\mathcal{S} : q']\!]$ are matched by $match_E(q', e_{\mathcal{S}}, \mathsf{D}, 0)$ for the base case and $match_E(q', e_{\mathcal{S}}, \mathsf{D}, i+1)$ in the induction step.

2. If $\lambda_e(e_{\mathcal{S}}) = s$ then the language $\mathcal{L}[\![e_{\mathcal{S}}]\!] = \{\Sigma^n \mid n \in s\}$. The strings $D \in \Sigma^n \cdot \mathcal{L}[\![\mathcal{S} : q']\!]$ are matched by $match_E(q', e_{\mathcal{S}}, \mathsf{D}, 0)$ for the base case and $match_E(q', e_{\mathcal{S}}, \mathsf{D}, i+1)$ in the induction step.

□

Now what is left is to show the substitution operation respects language equivalence between edge $e_{\mathcal{S}}$ and NFA $\mathcal{N}$.

Lemma C.2. $\mathcal{L}[\![\mathcal{N}]\!] = \mathcal{L}[\![e_{\mathcal{S}}]\!] \rightarrow \mathcal{L}[\![\langle \mathcal{N} / e_{\mathcal{S}} \rangle_{\mathcal{S}}]\!] = \mathcal{L}[\![\mathcal{S}]\!]$

Proof. This is a straightforward application of (Lemma C.1). □

We need two more auxilary lemmas. First, describe how interval set composition translates to langage union.

Lemma C.3. $\mathcal{L}[\![\{i_{n+1}, \overline{i_n}\}]\!] = \{\Sigma^m \mid m \in i_{n+1}\} \cup \mathcal{L}[\![\{\overline{i_n}\}]\!]$

The second is similar, an interval set composition in the NFA construction (Definition C.5) translates to language union.

Lemma C.4.

$$\mathcal{L}[\![\mathcal{N}(\{i_{n+1}, \overline{i_n}\}, \Sigma)]\!] = \{\Sigma^m \mid m \in i_{n+1}\} \cup \mathcal{L}[\![\mathcal{N}(\{\overline{i_n}\}, \Sigma)]\!]$$

Both lemmas are straightforward to prove from their definition. Finally, prove the NFA construction for interval sets $\mathcal{N}_I(s, \Sigma)$ has the same language as skip s for all possible skips.

Lemma C.5. $\lambda_e(e_s) = s \rightarrow \mathcal{L}[\![\mathcal{N}(s, \Sigma)]\!] = \mathcal{L}[\![e_{\mathcal{S}}]\!]$

Proof. For skip $s = \{\overline{i_n}\}$ prove this statement by induction on the number of intervals n.

1. For the base case, $n = 1$ as skips are non-empty sets of intervals, then $\mathcal{L}[\![\mathcal{S}(e_s)]\!] = \{\Sigma^n \mid n \in i_1\}$ and $\mathcal{N}(s, \Sigma) = \mathcal{N}_I(i_1, \Sigma)$ as only one epsilon transition is possible from $\mathcal{N}(s, \Sigma)$, the one to $\mathcal{N}_I(i_1, \Sigma)$. By inspecting the δ relations in $\mathcal{N}_c(i_1, \Sigma)$ and $\mathcal{N}_o(i_1, \Sigma)$ (Definition C.4), both recognize exactly $\{\Sigma^n \mid n \in i_1\}$.

2. For the inductive case, $n' = n + 1$, use the auxilary lemmas (Lemma C.3, Lemma C.4) to translate composition of interval sets to language union, as well as composition of interval NFA to language union. Both lemmas produce a language union with $\{\Sigma^m \mid m \in i_{n+1}\}$ which cancel out. The result is exactly satisfied by the induction hypothesis.

□

SAFA to AFA recursive definition. The last construction that converts a SAFA to an AFA is now possible. We give a well-founded recursion procedure *unskip*, based on the number of skips in SAFA $n_{\mathcal{S}} = |\{e \mid e \in E_{\mathcal{S}}, \lambda_e(e) = s\}|$ which will substitute skip n on each iteration, for $0 \leq n \leq n_{\mathcal{S}}$.

Definition C.10.

$$\begin{aligned} \text{unskip}(0,\ \mathcal{S}) \quad &= \mathcal{S} \\ \text{unskip}(n+1,\ \mathcal{S}) \quad &= \langle \mathcal{N}(s_n, \Sigma) / s_n \rangle_{\text{unskip}(n, \mathcal{S})} \end{aligned}$$

This procedure runs once for all $e \in E_{\mathcal{S}}$ in $\mathcal{S}$, where $\lambda_e(e) = s$ is a skip and substitutes s for its equivalent NFA $\mathcal{N}(s, \Sigma)$ until there are no more skips. By this definition $\text{unskip}(n_{\mathcal{S}}, \mathcal{S})$ is an AFA.

Next to show $\mathcal{S}$ and $\text{unskip}(n_{\mathcal{S}}, \mathcal{S})$ are equivalent in terms of the regular languages they recognize, we prove

Lemma C.6. $\mathcal{L}[\![\mathcal{S}]\!] = \mathcal{L}[\![unskip(n_{\mathcal{S}}, \mathcal{S})]\!]$

Proof. We proceed by induction on the number of skips $n_{\mathcal{S}}$.

1. For $n_{\mathcal{S}} = 0$, there are no skips in $\mathcal{S}$, then $\text{unskip}(0, \mathcal{S}) = \mathcal{S}$ an AFA, the automata are equal and their languages are equal.

2. For $n_{\mathcal{S}} = n + 1$, assume $\mathcal{S}_n$ is an AFA with all skips already substituted and the induction hypothesis $\mathcal{L}[\![\mathcal{S}]\!] = \mathcal{L}[\![\mathcal{S}_n]\!]$. We must prove $\mathcal{L}[\![\mathcal{S}]\!] = \mathcal{L}[\![\mathcal{S}_{n+1}]\!]$.

   (a) Let s_n the current skip to substitute, then $\mathcal{L}[\![\mathcal{S}_{n+1}]\!] = \mathcal{L}[\![\text{unskip}(n+1, \mathcal{S}_n)]\!] = \mathcal{L}[\![\langle \mathcal{N}(s_n, \Sigma) / s_n \rangle_{\mathcal{S}_n}]\!]$ by unfolding the definition of *unskip*.

   (b) The key equality to conclude the proof is by (Lemma C.2) $\mathcal{L}[\![\langle \mathcal{N} / s_n \rangle_{\mathcal{S}_n}]\!] = \mathcal{L}[\![\mathcal{S}_n]\!]$, provided that $\mathcal{L}[\![\mathcal{N}(s_n, \Sigma)]\!] = \mathcal{L}[\![s_n]\!]$, which we proved (Lemma C.5).

   (c) All that is left is exactly the induction hypothesis $\mathcal{L}[\![\mathcal{S}]\!] = \mathcal{L}[\![\mathcal{S}_n]\!]$ which concludes the proof.

□

$$\frac{}{\emptyset \preceq r}\ \textsc{Bot} \qquad \frac{}{r \preceq r}\ \textsc{Refl} \qquad \frac{\alpha \in \Sigma}{\alpha \preceq .}\ \textsc{Wild}$$

$$\frac{v(r) = \mathit{true}}{\epsilon \preceq r}\ \textsc{Nil} \qquad \frac{r \preceq s}{r^* \preceq s^*}\ \textsc{Star} \qquad \frac{}{r \preceq .^*}\ \textsc{Top}$$

$$\frac{i \leq j}{r\{i,j\} \preceq r^*}\ \textsc{RepStar} \qquad \frac{r \preceq s \quad i_2 \leq i_1 \quad j_1 \leq j_2}{r\{i_1,j_1\} \preceq s\{i_2,j_2\}}\ \textsc{Rep}$$

$$\frac{r \preceq s \quad i \leq 1 \leq j}{r \preceq s\{i,j\}}\ \textsc{Rep1} \qquad \frac{r \preceq s \quad r' \preceq s'}{rr' \preceq ss'}\ \textsc{App}$$

$$\frac{r \preceq s \quad r \preceq u}{r \preceq s \mid u}\ \textsc{Alt} \qquad \frac{r \preceq u}{r \,\&\, s \preceq u}\ \textsc{AndL}$$

$$\frac{s \preceq u}{r \,\&\, s \preceq u}\ \textsc{AndR}$$

FIGURE 24—A partial ordering on regex $r \preceq s$ iff the language of r is subset (or equal) of the language of s.

D Compiling Regular Expressions to SAFA

We present here a recursive compilation procedure from regex to SAFA, based on generalized Antimirov derivatives (Section 13). Assume syntactic sugar expansion (Section E.1) and regex normalization by weak equivalence $\simeq$ is already done.

Start with a fully normalized regex r, alphabet Σ. Create an empty SAFA given alphabet Σ and states of type $B^+(r)$ then run this recursive procedure.

Given a regex r,

1. If state r exists in the SAFA, return. Otherwise add the new state r to Q.
2. Extract a skip $r \xrightarrow{s} r'$ (Section E.3) from r, if possible. Then s is the skip interval set and r' is the remaining regex when no more wildcards can be extracted. Label state r an $\exists$ state by $\lambda_q(r) = \exists$ and add to it a new outgoing edge e such that $(r, s, r') \in \delta$ and $\lambda_e(e) = s$. Recurse for r'.
3. Otherwise, for each character $\alpha \in \Sigma$ take the derivative of r with respect to α to be a boolean algebra expression $\partial_\alpha(r)$ (Section B.1) in disjunctive normal form (Section A.3) and add one transition for each character $(r, \alpha, \partial_a(r)) \in \delta$.
   (a) In DNF, the derivative $\partial_\alpha(r) = \bigvee_i (\bigwedge_j r_{i,j})$ proceed to add i existantial $\exists$ states $(\bigwedge_j r_{i,j})$ and for each j add a forall $\forall$ state $r_{i,j}$.
   (b) Then add ϵ-transitions $(\partial_\alpha(r), \epsilon, \bigwedge_j r_{i,j}) \in \delta$ for each i, as well as $(\bigwedge_j r_{i,j}, \epsilon, r_{i,j}) \in \delta$ for each j.
   (c) Recurse for each leaf state $r_{i,j}$.

Note, the number of new states added in step 3(a) is $\mathcal{O}(|\Sigma| \cdot i \cdot j)$. In practice, however, we noticed regex are not nested as much so $i \cdot j$ is small.

E Regular expression preprocessing

E.1 Syntactic sugar

All of the PCRE syntax in Figure 1 can be expressed in terms of the simpler syntax in Figure 9. The more interesting transformation is from a look-ahead to a boolean conjunction (&), indicating that *both* the look-ahead and the rest of the regex must match the string.

E.2 Regular expression normalization

The next step, after preprocessing syntactic sugar, is regex normalization, converting to a simpler, smaller regex by use of the weak syntactic equivalence [65] equations. Weak equivalence $r \equiv v$ is given in Figure 25 and is a nested recursive definition. Weak equivalence internally uses the *refinement* relation $r \preceq s$ in Figure 24, which corresponds to language inclusion.

$$\forall r,\ s,\ r \preceq s \quad \text{iff} \quad \mathcal{L}[\![r]\!] \subseteq \mathcal{L}[\![s]\!]$$
$$\forall r,\ s,\ r \equiv s \quad \text{iff} \quad \mathcal{L}[\![r]\!] = \mathcal{L}[\![s]\!]$$

The proof of the above is simple, albeit tedious. To show the weak equivalence is sound and complete with respect to language equality, proceed by induction on the derivation of the $r \equiv s$ relation in the left-to-right direction, and by induction on the definition of $\mathcal{L}[\![r]\!]$ in the right-to-left direction, similarly for $r \preceq s$.

The benefit of introducing this weaker notion of equivalence is a syntactic normalization procedure for regex. As Owens et al. [65] show, this normalization procedure is fast and successfully minimizes the number of states of the compiled automaton (SAFA) and thus the final step function size, which is proportional to the number of states.

E.3 Extract skips

We define the rules for extracting skips from the begining of regex, one at a time, with the partial function $r \xrightarrow{S} r'$ that extracts skip S (Figure 26).

E.4 SAFA solver

At a high-level, the SAFA solver algorithm is given by the SAFA semantics (Section C.5). A side-note, an additional advantage of non-determinism is the room for parallelization. We take advantage of non-determinism to parallelize the SAFA solver in at least three places using a threadpool.

1. On match$_\forall$ we parallelize solving across edges $e \in E$ then join and wait on the results.
2. On match$_\exists$ we parallelize solving across edges $e \in E$, but instead of join we race the threads. The first thread to find a solution returns and the rest are killed.

$$\frac{}{. \xrightarrow{\{[1,1]\}} \epsilon} \text{DOT} \qquad \frac{r \xrightarrow{\{\}} \epsilon}{r^* \xrightarrow{\epsilon} \epsilon} \text{EMPTY*} \qquad \frac{r \xrightarrow{\epsilon} \epsilon}{r^* \xrightarrow{\epsilon} \epsilon} \text{NIL*}$$

$$\frac{r \xrightarrow{s} \epsilon}{r^* \xrightarrow{\{[0,\infty)\}} \epsilon} \text{STAR*} \qquad \frac{r \xrightarrow{\{\}} \epsilon}{r\{0,j\} \xrightarrow{\epsilon} \epsilon} \text{RE_1}$$

$$\frac{r \xrightarrow{\{\}} \epsilon \quad i \neq 0}{r\{i,j\} \xrightarrow{\{\}} \epsilon} \text{RE_2} \qquad \frac{r \xrightarrow{\epsilon} \epsilon \quad i \leq j}{r\{i,j\} \xrightarrow{\epsilon} \epsilon} \text{RNIL}$$

$$\frac{r \xrightarrow{s} \epsilon \quad i \leq j}{r\{i,j\} \xrightarrow{s^i \cup \ldots \cup s^j} \epsilon} \text{RANGE} \qquad \frac{r_1 \xrightarrow{s_1} \epsilon \quad r_2 \xrightarrow{s_2} r'}{r_1 r_2 \xrightarrow{s_1 + s_2} r'} \text{APPR}$$

$$\frac{r_1 \xrightarrow{s} r_1'}{r_1 r_2 \xrightarrow{s} r_1' r_2} \text{APPL}$$

FIGURE 26—Inference rules for a partial, recursive function $r \xrightarrow{s} r'$ extracting skip s from the head position of a regex r and returning the tail r'

3. On match$_s$ we parallelize our search for different values of $n \in s$ and race the threads again. Even though skips s are unbounded, the string is bounded so we limit our solution search from $min(s)$ to $max(max(s), |D|)$.

The benefits of parallelization in the solver are concrete and is a contribution outside the cryptographic benefits of SAFA. Using SAFA we improve the performance of regex matching by taking advantage of non-determinism in branches and wildcards.

F Matrix representation of R1CS

We repeat our running example of constraints over $\mathbb{F}$:

$$\begin{aligned}
guard \times (x_0 - 30) &= 0 \\
guard \times (y - x_1) &= 0 \\
(1 - guard) \times (y - prod) &= 0 \\
x_0 \times inv - prod &= 0 \\
x_1 \times inv - 1 &= 0
\end{aligned}$$

We would like to convert these constraints to matrices A, B, and C such that $(A \cdot z) \circ (B \cdot z) = (C \cdot z)$, where $\cdot$ is the matrix-vector product and $\circ$ is the Hadamard product. There should only exist a solution vector $z = (io, 1, w)$, with witness $w \in \mathbb{F}^{cols - |io| - 1}$ when this set of constraints is satisfiable.

In the example from Section 2.3, y is the only public variable in io. The variables $x_0, x_1, guard, prod$, and inv are known only to $\mathcal{P}$, so they make up our witness w. So $z = (y, 1, x_0, x_1, guard, prod, inv)$.

First, we shuffle some of the constraints so that each is of the form *(addition term) * (addition term) = (addition term)*:

$$\frac{r \preceq s \quad s \preceq r}{r \simeq s} \text{ANTISYM} \qquad \frac{r \simeq s}{s \simeq r} \text{SYMM}$$

$$\frac{}{r \simeq \epsilon r} \text{APPNL} \qquad \frac{}{r \simeq r\epsilon} \text{APPNR} \qquad \frac{}{\emptyset \simeq \emptyset r} \text{APPZL}$$

$$\frac{}{\emptyset \simeq r\emptyset} \text{APPZR} \qquad \frac{}{r\{i,j\}r\{i',j'\} \simeq r\{i+i',j+j'\}} \text{APPREP}$$

$$\frac{s \simeq r}{s^* \simeq r^*} \text{STAR} \qquad \frac{r \preceq s}{r \mid s \simeq s} \text{ALTLEQ}$$

$$\frac{r \simeq s \quad i' \leq j}{r\{i,j\} \mid s\{i',j'\} \simeq r\{min(i,i'), max(j,j')\}} \text{ALTREP}$$

FIGURE 25—Weak regex equivalence $r \simeq s$ iff the language of r is equal to the language of s.

$$\begin{aligned}
guard \times (x_0 - 30) &= 0 \\
guard \times (y - x_1) &= 0 \\
(1 - guard) \times (y - prod) &= 0 \\
x_0 \times inv &= prod \\
x_1 \times inv &= 1
\end{aligned}$$

We create the corresponding R1CS matrices A, B, C:

$$A = \begin{bmatrix} 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \end{bmatrix}$$

$$B = \begin{bmatrix} 0 & -30 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & -1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

$$C = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Notice all of the matrices have 6 rows, since there are 6 multiplication constraints, and 8 columns, since the length of z is 8. If $\mathcal{P}$ has (for example) $x_0 = 10, x_1 = 5$, and wants to prove that $y = 2$, a vector $z = (2, 1, 10, 5, 0, 5, 2, 5^{-1})$ satisfies this R1CS instance. Note that we use 5^{-1} as the inverse of 5 in $\mathbb{F}$.

It is easy to see that this z only satisfies $(A \cdot z) \circ (B \cdot z) = (C \cdot z)$ when the assignments $y = 2$, $x_0 = 10$, $x_1 = 5$, $guard = 0$, $tmp = 5$, $prod = 2$, $inv = 5^{-1}$ satisfy our original constraints.

G Low cost padding

The obvious way to hide the size of a string D is by constructing a string D' which is equal to D but padded with dummy characters to some suitable upper bound. If the padding is chosen to be $0 \in \mathbb{F}$, then the committer $\mathcal{G}$ has to do no extra work, since $g^0 = 1$ for all generators g of the polynomial commitment scheme. However, the `nlookup` prover and the `nlookup` verifier (which is embedded within Reef's step function) still need to do work proportional to $|\mathsf{D}'|$ for each step: linear for the prover and a logarithmic number of constraints to express the verifier, plus $\mathcal{O}(|\mathsf{D}'|)$ operations at the end for *ProveEval* and $\mathcal{O}(\sqrt{|\mathsf{D}'|})$ operations for *VerifyEval* when we use the Hyrax (§6.3). If the upper bound is chosen to be large (e.g., $|\mathsf{D}'| = 2^{30}$), the cost to the prover would be prohibitive.

We observe that the same ideas in table projections that allow the prover to do less work can be used here: the prover's work during each step can be made linear in $|\mathsf{D}|$ (the unpadded document). The key observation is that given that D is a subset of D', and that padding is just 0s, it is possible for the prover to project the entries in the table corresponding to D, *without having to reveal to the verifier the selector s*. Consequently, the verifier learns nothing about D except for $|\mathsf{D}'|$, and the prover is able to save considerable costs.

The basic idea is to pad the multilinear extension of the document strategically with 0, and commit to this padded multilinear extension. This is, in spirit, committing to a larger document, $\widetilde{\mathsf{D}'}$. The prover can then run a slightly larger `nlookup` in the step function that looks to be operating over a larger document, so the size of the real document is hidden. But we leverage the structure of our multilinear polynomial so that work to generate the commitment and the work of $\mathcal{P}$ to generate `nlookup` witnesses is closer to the work done in the case of the original, smaller document.

Given a multilinear extension to the original document, $\widetilde{D}(x_0,...,x_{\ell-1})$, of length $|D| = 2^{\ell}$, we generate a multilinear extension to a larger document, of length $|D'| = 2^{\ell'}$:

$$\widetilde{D'}(p_0,...,p_{\ell'-\ell-1},x_0,...,x_{\ell-1}) = \widetilde{D}(x_0,...,x_{\ell-1})$$

$\widetilde{D'}$, will evaluate the same way $\widetilde{D}$ does on any point, "throwing away" its padding variables, $p_0,...,p_{\ell'-\ell}$. It is committed to by inserting zeros into the multilinear extension's coefficient commitment vector for every term that includes any padding variable p_i. (This will be a predictable pattern.) Although the literal document D' (i.e. the evaluations of $\widetilde{D'}$ over the boolean hypercube) never has to be materialized, it may be helpful to visualize it. For each padding variable added to the input's of $\widetilde{D'}$, the document size doubles, and the document repeats itself.

Notice that we are treating the document's multilinear extension as a vector of coefficients, rather than a vector of evaluations, which is a change from the original description. This does not change any of our previous cost evaluations, nor does it prevent the use of the projection or hybrid table optimizations. Implementation would require tweaking the Hyrax code (or writing/using code for any commitment scheme that supports inner product).

For example, if $\widetilde{D}(x_0,x_1) = 7+5x_0+3x_1+2x_0x_1$, the commitment to a extension with one padding variable, $\widetilde{D'}(p_0,x_0,x_1)$ will be to the vector $[7,0,5,3,0,0,2,0]$. The final check of that $\widetilde{D'}(q_0,q_1,q_2) = v$ will be done with an inner product argument that proves $\langle[7,0,5,3,0,0,2,0],[1,q_0,q_1,q_2,q_0q_1,q_0q_2,q_1q_2,q_0q_1q_2]\rangle = v$. The actual document D is $[7,12,10,17]$, and if materialized, D' would be $[7,12,10,17,7,12,10,17]$. The key here is that the commitment "zeros out" the padding variables, without revealing to the verifier which variables are padding.

This varies from typical projections in that our padding variables are not known to $\mathcal{V}$, since knowing the length would leak things about the length of the document.

The work to generate the commitment is the same as if we did not have padding—any generator exponentiated by 0 is 1. So $\mathcal{G}$ does not have to do extra exponentiations or multiplications for this larger commitment.

When producing sumcheck witnesses (as part of producing `nlookup` witnesses), $\mathcal{P}$ has to calculate evaluations over $\widetilde{D'}$ of the form:

$$\widetilde{D'}(r_0,...,r_{i-1},x,b_{i+1},...,x_{\ell'-1})$$
$$r_i \in \mathbb{F}$$
$$x \in \{0,1\}$$
$$b_i \in \{0,1\}$$

Instead, it can calculate evaluations over $\widetilde{D}$. Because of the structure of $\widetilde{D'}$, the evaluations over $\widetilde{D}(x_0,...,x_{\ell-1})$ can be calculated once, and reused to mimic evaluations over $\widetilde{D'}(p_0,...,p_{\ell'-\ell-1},x_0,...,x_{\ell-1})$, no matter what the values of $p_0,...,p_{\ell'-\ell-1}$. This ends up being $\mathcal{O}(|D| + \log(\frac{|D'|}{|D|}))$ work, instead of $\mathcal{O}(|D'|)$. The log factor covers any doubling of the precalculated $\widetilde{D}$ evaluations that have to be done to pad "extra" `nlookup` rounds (since there are now $\log(|D'|)$ rounds in the step function).

At the end of our protocol, $\mathcal{V}$ must verify a claim of the form $\widetilde{D'}(q_r) = v_r$, where $q_r \in \mathbb{F}^{\ell'}$. This is done in the usual way using an inner product argument and our commitment to $\widetilde{D'}$, and implies consistency of all of our lookups with the original $\widetilde{D}$.

H Implementation Optimizations

H.1 Batching

To leverage the amortization of `nlookup`, Reef reads a batch of $m \geq 1$ characters and transitions from the hybrid table within each step function. This results in having to perform $\frac{|\alpha|}{m}$ where $\alpha = \mathcal{O}(|\mathsf{D}| \cdot L)$. The benefit is that `nlookup` requires $\mathcal{O}(m \log n) + \mathcal{O}_H(\log n)$ constraints for each step when

looking up m entries from a table of size n, and the hash component is typically the dominant cost. As we discuss in Section 6.2, this results in $\mathcal{O}(\alpha \log(|Q_{SAFA}| \cdot |\Sigma|))$ constraints plus $\mathcal{O}(\frac{\alpha}{m} \log(|Q_{SAFA} \cdot |\Sigma|))$ hashes.

Since the hash component is the dominant cost, one might wonder whether setting $m = \alpha$ is optimal, as it minimizes the impact of the hash component. But this has a variety of issues.

First, we cannot actually set m to α since the actual value of α depends on the document and the R1CS instance is created independent of the document. This means we would need to set m to be the worst-case α which grossly overestimates its actual value (as we show in our evaluation).

Second, if there is a single step then there is no recursion. If there is no recursion, then this means that Reef cannot skip work because it cannot finish early—it has to perform all the operations in the single step. A corollary of this is that to benefit from the skipping powers of SAFA, Reef needs steps to be of a reasonable granularity (not too big).

Third, Nova actually benefits from having many steps because folding is cheaper than proving. If there is a single step, then there is no folding taking place and the entire cost is proving.

Fourth, a larger step function leads to larger proof sizes and a more expensive verifier since the size of proofs in our version of Nova are logarithmic in the size of the step function, and require work linear in the size of the step function to verify (owing to our use of Bulletproofs [27] inner product argument).

As a consequence of the above, the relationship between the size of the final proof, the number of constraints, the total computational cost, the ideal batch size, and the number of steps is not linear and requires careful tuning since it depends on many factors including the regex itself. Reef's compiler contains a cost model that takes into account all of the above (and a few other low-level concerns) and decides on the best batch size.

H.2 Optimized stack

Rather than using the hash chain stack construction, Reef represents a stack using a vector of field elements and a stack pointer field element:

```
field[stack_size+1] push(
  field[2*stack_size] stack,
  field stack_ptr, (field child, field cursor)) {
    for i in stack_size {
    if i == stack_ptr {
      stack[i] = (child, cursor);
      stack_ptr += 1;
    }
  }
  return {stack, stack_ptr};
}

field[stack_size+3] pop(field[2*stack_size] stack,
  field stack_ptr) {
  for i in stack_size {
    if i == stack_ptr {
      (popped_child, popped_cursor) = stack[i];
      stack_ptr -= 1;
    }
  }
  return {stack, stack_ptr, popped_child,
    popped_cursor};
}
```

The stack needs to be big enough to accommodate all of the children for all of the nested `forall` nodes on any particular path. This number, `stack_size` is calculated during the step function compilation. This is usually more efficient than a hash chain stack.

Additionally, since pushes to and pops from the stack only need to happen under certain conditions (encountering a `forall` state or finishing transversal of a branch), it is a waste of constraints to include pop constraints and multiple sets of push constraints for every lookup in the batch. Reef instead uses constraints that may perform a single pop or several pushes during only the first lookup. If during witness generation, $\mathcal{P}$ needs to perform a pop/push and does not currently have access to the correct (first) lookup, it is allowed to "loop" on the current state, consuming ϵ characters, until the lookup constraints are available. Obviously, if the batch size is set badly, this could become inefficient. We choose batch size carefully; during table generation, Reef walks over the SAFA in a depth-first search, and takes note of the length of paths between `forall` nodes and accepting states. The batch size is the average length of these paths. Users of Reef can also override this mechanism and set the batch size themselves.

H.3 Pipelined solving and proving.

Reef also optimizes the solving/proving pipeline; $\mathcal{P}$'s solver runs in parallel with the thread that produces the folded cryptographic proof for $\mathcal{P}$. The solver thread calculates a witness for step i and hands it off to the prover, which is able to focus on proving step i while the solver moves on to generating witnesses for step $i+1$.

I Alternate instantiation of RAM with better asymptotics

The majority of the costs in Section 8 come from our use of lookup arguments and polynomial commitments. However, Reef can easily swap the lookup argument and polynomial commitment and use a Merkle Tree to represent the SAFA table and the document (assuming the hash function heuristically instantiates a random oracle). This gives us the efficient random access memory we need in Reef.

With Merkle trees using a SNARK-friendly hash function like Posseidon [44], we have the following cost profile. Here we redefine $T = |\mathsf{D}| \cdot |Q_{SAFA}|$, since Merkle Trees do not amortize requests and hence there is no benefit in combining the public SAFA table and the private document table. We thus assume we have two separate Merkle trees.

Commitment generation. $\mathcal{G}$ must perform $\mathcal{O}(|\mathsf{D}|)$ finite field computations.

Prover's cost. For processing a batch of m characters at a time, the step function has $\mathcal{O}(m \log T)$ constraints, and there are a total of $\mathcal{O}(\alpha/m)$ steps to finish processing a document. This results in $\mathcal{P}$ performing a total of $\mathcal{O}(\alpha \log(T)/\log(m \log T))$ group operations in Nova. The resulting proof π is of size $\mathcal{O}(\log(m \cdot \log T))$.

To generate the witness for each step, $\mathcal{P}$ also needs to perform $\mathcal{O}(m \log T)$ finite field operations in order to generate the appropriate Merkle proofs (though these could be pre-generated and stored for later use).

In total, there are $\mathcal{O}(\alpha \log(T)/\log(m \log T))$ group and $\mathcal{O}(\alpha \log T)$ finite field operations.

Verifier's cost. The total cost to the verifier $\mathcal{V}$ is simply $\mathcal{O}(m \log(T)/\log(m \log T))$ group operations in Nova to verify π. There is no need for any auxiliary proofs.

Discussion. While clearly this alternate approach is asymptotically better, our experiments reveal that arithmetizing so many hash functions leads to very large R1CS instances in practice, and hence why we choose to rely on more complicated lookup arguments that amortized these costs. Of course, there is likely to be some document and SAFA size for which this alternate approach is better. Fortunately, the main contributions of our work: Reef's `match_step` design and SAFA, are orthogonal to the proof system (as long as it is recursive) and the way that random access memory is instantiated.

J Applications

Here we recount the full results from our experimental evaluation of Reef for our motivating applications. We start by discussing the origin and rationale behind our test data.

Password Strength We randomly generated our good password set. Our bad password set was selected at random from the NordPass list of the top 200 most common passwords [11], which is a list of weak passwords. Our regex indicates strong passwords, of a certain length, with required characters from several different fields (uppercase and lowercase alphabet characters, numbers, and special characters).

Email Redactions For our redactions, we use the Enron email dataset [12]. Our small instance is their smallest instance, and our large instance was randomly selected. Our regexes indicate redacted versions of both.

ODoH Blocklisting We use a regex filter for Pi-hole [9], which is a DNS sinkhole, for our oblivious DNS over HTTPS regexes. While blocklisting would traditionally prove non-matching, to better compare to existing work we instead prove matching. Our queries are randomly generated.

Genetic Matching For our evaluation we consider three common mutations of the BRCAI and BRCAII genes. Mutations in these genes are commonly linked to most forms of breast cancer. The base pairs for these genes, as well as for common mutations are all publicly available from the US National Institutes of Health [13, 14].

Full results. The results are in Figures 29–34.

Modified nlookup Protocol

The typical nlookup protocol happens interactively between an nlookup prover and nlookup verifier. We describe a modified version of nlookup where the "bigger" Reef prover aims to proves a successful set of lookups to the Reef verifier by encoding the nlookup verifier as R1CS. More details can be found in Sections 6.1 and 6.3.

The Reef prover wants to prove b batches of m lookups (per batch) in a certain table T. We say $\widetilde{T}$ is the multilinear extension of that table, parameterized over $\ell = \log(|T|)$ variables. There is one batch per Nova folding.

For each batch $\beta \in [0, b]$:

1. Input: $m+1$ evaluation points $(q_1, v_1), ..., (q_m, v_m), (q_r, v_r)$ of a multilinear polynomial $\widetilde{T}$, such that $\widetilde{T}(q_i) = v_i$. For $i \leq m$, $q_i \in \{0,1\}^\ell$ and $q_r \in \mathbb{F}^\ell$. We refer to this last point as the "running claim". The first running claim ($\beta = 0$) can be an arbitrary point in $\widetilde{T}$.
2. The nlookup verifier, simulated by a hash computation in R1CS, chooses challenge ρ.
3. The Reef prover proves:

$$v_r + \sum_{i=1...m} \rho^i \cdot v_i = \sum_{j \in \{0,1\}^\ell} \widetilde{eq}(q_r, j) \cdot \widetilde{T}(j) + \sum_{i=1...m} \rho^i \cdot \sum_{j \in \{0,1\}^\ell} \widetilde{eq}(q_i, j) \cdot \widetilde{T}(j)$$

   - This is done using the sum-check protocol, encoded in R1CS over $\log(|T|)$ rounds. The left-hand side is the claim the sum-check prover makes. Each round, a degree 2 polynomial is sliced off of the right-hand-side of the equation in response to a random challenge from the sum-check verifier. This challenge is again simulated in R1CS by a hash computation, to produce a non-interactive protocol.
   - We maintain a Poseidon hash sponge in R1CS that absorbs a table commitment (if the table is not public), and all of the (q_i, v_i) pairs, including the running claim. The binary q_i elements can be packed into a smaller number of field elements for efficiency. This sponge is squeezed to produce ρ. Then, for each sum-check round, it absorbs the "messages" sent by the prover (describing the polynomial slice), and is squeezed to produce the sum-check verifier's random challenge.
   - Note the polynomial $\tilde{eq}(x, e) = \prod_{i=1}^{\ell}(e_i \cdot x_i + (1 - e_i) \cdot (1 - x_i))$ is a multilinear extension of a function that outputs 1/0 depending on whether $x == e$ or not.
   - During the last sum-check round, the sum-check verifier is required to check the evaluation of the right-hand side, over a random vector r of length ℓ in $\mathbb{F}$ in place of j. The nlookup verifier will indeed evaluate all of right-hand side except for $\widetilde{T}(r)$, which it will delay until the next $\beta + 1$ batch. It will set the next running claim $q_r = r, v_r = \widetilde{T}(r)$.

After b batches are completed, the nlookup verifier is required to confirm the final running claim, $\widetilde{T}(q_{r,\beta}) = v_{r,\beta}$.

1. Delaying this evaluation (the most expensive part of the right-hand side evaluation) until the end amortizes its cost over b batches.
2. The Reef verifier simulates the nlookup verifier by doing this computation separately from the main nlookup proof. If the table is public, this is straightforward. In the case of a private table, this check has to be done over a commitment to the table. (So there must be a commitment that supports polynomial evaluation.) In Reef, this check is proved by π_{poly} (notation from the body of the paper).
3. Additionally, in the case of a private table, the Reef verifier should not see $v_{r,\beta}$, since it leaks some information about the table. Instead, the verifier is handed $H_r = H(v_{r,\beta}||blind)$ and a commitment to $v_{r,\beta}$. This commitment is used to verify π_{poly}. The relationship between the commitment and the hash is proven by a separate proof, $\pi_{consistency}$.

FIGURE 27—The full nlookup protocol, with our zero-knowledge modifications. See [56] for the original protocol.

Full Reef Protocol

Here we describe the full Reef protocol. Black text indicates normal, non-optimized actions. Blue text indicates changes made when using hybrid tables. See Section 6.5. Red text indicates changes made when using projections. See Section 6.4. Green text indicates changes made when using low cost padding. See Appendix G.

1. A committer (who can also be the prover) uses a polynomial commitment scheme *(Setup, Commit, ProveEval, VerifyEval)*, in our case, *Hyrax-PC* to commit to a multilinear extension $\widetilde{D}$ of the private document D; Both the prover and verifier can access this commitment. This commitment can be reused across multiple proofs.
   (a) Committer runs *Setup* to produce pp, and $Commit(pp, \widetilde{D})$ to produce $C_{\widetilde{D}}$.
      - Projections and hybrid tables do not require any changes to the commitment.
      - Since low cost padding hides the length of the document in the commitment, this requires extending D to D_{ext}.

   (b) Committer sends $pp, C_{\widetilde{D}}$ to the prover and verifier, and blinding information about $C_{\widetilde{D}}$ to only the prover.

2. The Reef prover and verifier choose a public regex to create a proof about. Both compile the regex to a SAFA. They generate a set of R1CS constraints that verifies a single batch of a lookups using two `nlookup` protocols, one for the SAFA lookups (over multilinear extension $\widetilde{S}$ that describes the SAFA transition table) and one for the character lookups in the document (over multilinear extension $\widetilde{D}$). The rounds done by the $\widetilde{S}$ `nlookup`'s sum-check engine is $\ell = \log_2(|S|)$. The rounds in $\widetilde{D}$ `nlookup` is $\ell = \log_2(|D|)$. This R1CS compilation is a deterministic public process that they can do independently.
   - Using hybrid tables requires prior agreement and a change to the R1CS: there is only one `nlookup` protocol (over multilinear extension $\widetilde{T}$ of the combined table), that performs twice as many lookups. The rounds done by this `nlookup` protocol is $\ell_h = \log_2(|T|) = \log_2(2 \cdot max(|S|, |D|))$.
   - When using projections, the `nlookup` over $\widetilde{D}$ as $\ell_p = \log_2(|D_{proj}|)$ rounds, where D_{proj} is the section(s) of the document that actually need to be processed. The prover and verifier agree on where this section is, and the $\log_2(|D|) - \ell_p$ public boolean variables that index that section, referred to as q_{idx}.
   - Low cost padding: the number of rounds in the $\widetilde{D}/\widetilde{T}$ `nlookup` will technically depend on $|D_{ext}|$ rather than $|D|$, though the verifier will not be aware of this distinction.
   - Obviously, it is possible to use these three optimizations in any combination, though some combinations will be less useful than others, depending on the document and regex.

3. The prover:
   (a) iterates through (the appropriate subset of) the document. For each batch of document characters, it generates the appropriate `nlookup` witnesses (this requires running linear-time sum-check solver), and other bookkeeping witnesses the regex R1CS needs. The prover folds each new batch into it's proof of lookups, π. This `nlookup` witnesses include the final running claims for both `nlookup` protocols. The first we refer to as (q_s, v_s), where $\widetilde{S}(q_s) = v_s$. The second is (q_d, v_d), where $\widetilde{D}(q_d) = v_d$.
      - Hybrid: there is a single running claim, (q_r, v_r), over the single table, such that $\widetilde{T}(q_r) = v_r$.
      - Projections: The sum-check witnesses for the $\widetilde{D}/\widetilde{T}$ `nlookup` are calculated over $\widetilde{D}_{proj}$.
      - Low cost padding: The sum-check witnesses for the $\widetilde{D}/\widetilde{T}$ `nlookup` are calculated over $\widetilde{D}$ with some additional work to mimic $\widetilde{D}_{ext}$.

   (b) generates $\pi_{consistency}$, that proves the hash H_d of v_d is consistent with a commitment C_{v_d} to v_d.
      - Hybrid: We use the hash H_r of v_r and the commitment C_{v_r} to v_r instead.

   (c) runs $ProveEval(\widetilde{D}, q_d, v_d)$ to generate π_{poly}, that proves $\widetilde{D}(q_d) = v_d$
      - Hybrid: $ProveEval(\widetilde{D}, q_r[1..], v_d)$ generates π_{poly} instead. The prover also generates π_{eq}, a proof that $(1 - q_r[0]) \cdot v_s + q_r[0] \cdot v_d = v_r$.
      - Projections: $ProveEval(\widetilde{D}, q_{idx}||q_d, v_d)$ generates π_{poly} instead.

   (d) sends the verifier $\pi, \pi_{consistency}, \pi_{poly}, q_s, v_s, q_d, H_d$ and C_{v_d}.
      - Hybrid: Instead of q_s, v_s, q_d, H_d, the prover sends q_r, H_r, C_{v_r}, as well as π_{eq}, which proves $(1 - q_r[0]) \cdot v_s + q_r[0] \cdot v_d = v_r$.

4. The verifier:
   (a) verifies π and checks that the public part of π's witness makes sense; that is, that SAFA traversal starts at state 0 and ended at an accepting state (requiring an end of file character to be "seen" in the document), and the cursor stack starts and ends totally empty.

   (b) verifies $\pi_{consistency}$, using H_d and C_{v_d}. (Hybrid: H_r and C_{v_r} are used instead.)

   (c) runs $VerifyEval(C_{\widetilde{D}}, q_d, \pi_{poly}, v_d)$ and checks $\widetilde{S}(q_s) = v_s$ "in the clear".
      - Hybrid: $VerifyEval(C_{\widetilde{D}}, q_r[1..], \pi_{poly}, v_d)$ is called instead. verifier does not need to verify $\widetilde{S}(q_s) = v_s$ (these values don't exist). It must verify π_{eq}, which requires the verifier to compute it's own commitment to $(1 - q_r[0]) \cdot v_s + q_r[0] \cdot v_d$, which is possible using q_r, C_{v_d}, and a computation of $v_s = \widetilde{S}(q_r[1..])$ done in the clear.
      - Projections: $VerifyEval(C_{\widetilde{D}}, q_{idx}||q_d, \pi_{poly}, v_d)$ is called instead.

FIGURE 28—The full Reef protocol, with modifications required for each optimization.

Application	Document ID	Regex ID	R1CS Constraints	Doc. Length	# Steps	Compile Time (s)	Solving Time (s)	Proving Time (s)	Verifying Time (s)	Proof Size (KB)	Commit Size (KB)	Max Memory Usage (GB)
Redactions												
	Small Email	r1	46,655	415	4	36.947	0.760	3.169	0.553	32.609	0.512	0.733
	Large Email	r2	65,727	1,000	7	217.628	3.221	5.923	0.701	33.361	1.024	1.051
ODoH												
	5f558	r3	18,437	128	3	16.180	0.081	1.904	0.424	31.793	0.512	0.330
	25424	r4	22,692	128	2	19.650	0.213	1.709	0.435	31.889	0.512	0.362
	55824	r5	23,148	128	1	19.676	0.040	1.366	0.422	31.857	0.512	0.387
	21d97	r6	18,409	128	2	16.092	0.028	1.601	0.409	31.761	0.512	0.329
	49b9a	r7	18,433	128	2	16.095	0.028	1.579	0.407	31.761	0.512	0.323
	b8f74	r8	18,263	128	2	16.030	0.028	1.560	0.406	31.761	0.512	0.330
	3b4ed	r9	17,177	128	2	15.415	0.023	1.572	0.398	31.761	0.512	0.308
	24448	r10	18,865	128	2	16.241	0.029	1.572	0.414	31.761	0.512	0.335
	b329c	r11	18,241	128	2	16.075	0.025	1.575	0.405	31.761	0.512	0.326
	6f74a	r12	18,241	128	2	15.995	0.028	1.574	0.413	31.761	0.512	0.323
	83a9c	r13	17,785	128	2	15.795	0.025	1.569	0.411	31.761	0.512	0.313
	5410f	r14	17,617	128	1	15.797	0.014	1.344	0.400	31.761	0.512	0.318
	a0514	r15	17,365	128	1	15.578	0.013	1.292	0.399	31.697	0.256	0.314
	b4ebd	r16	17,617	128	3	15.789	0.034	1.869	0.412	31.761	0.512	0.314
Passwords												
Match	dcdc9	r17	19,982	12	5	17.960	0.067	2.573	0.418	31.665	0.128	0.347
	43db4	r17	19,982	12	5	17.975	0.067	2.571	0.415	31.665	0.128	0.341
	91edc	r17	19,982	12	5	17.936	0.066	2.581	0.409	31.665	0.128	0.337
	2bcf2	r17	19,982	12	5	17.897	0.068	2.597	0.411	31.665	0.128	0.343
	10bf0	r17	19,982	12	5	18.086	0.079	2.599	0.415	31.665	0.128	0.341
	aff42	r17	19,982	12	5	17.901	0.068	2.577	0.421	31.665	0.128	0.347
	edde7	r17	19,982	12	5	18.011	0.067	2.555	0.413	31.665	0.128	0.350
	1539c	r17	19,982	12	5	17.904	0.067	2.580	0.413	31.665	0.128	0.344
	7bfcc	r17	19,982	12	5	17.992	0.067	2.585	0.414	31.665	0.128	0.344
	dfa02	r17	19,982	12	5	17.899	0.071	2.600	0.410	31.665	0.128	0.344
Non-Match	e73ee	r17	20,728	8	7	18.638	0.410	3.284	0.425	31.761	0.128	0.328
	b5f3a	r17	20,728	8	6	18.610	0.345	2.979	0.415	31.761	0.128	0.331
	fd1e7	r17	20,725	6	6	18.550	0.344	2.944	0.418	31.729	0.064	0.327
	db267	r17	20,725	3	5	18.731	0.285	2.631	0.414	31.729	0.064	0.326
	40867	r17	20,728	8	6	18.566	0.360	3.004	0.411	31.761	0.128	0.326
	f4a98	r17	20,725	6	6	18.640	0.359	2.978	0.417	31.729	0.064	0.332
	7474f	r17	20,728	8	6	18.578	0.349	2.969	0.424	31.761	0.128	0.332
	b20ef	r17	20,725	6	6	18.634	0.354	2.961	0.415	31.729	0.064	0.334
	27ba9	r17	20,728	7	6	18.594	0.355	2.976	0.427	31.761	0.128	0.329
	304b5	r17	20,728	9	6	18.636	0.357	2.963	0.416	31.761	0.128	0.330
DNA												
Match	BRCA1 Var1	r18	35,306	43,054,295	2	47.766	2.157	18.425	0.837	33.761	262.144	7.937
	BRCA1 Var2	r19	50,783	43,054,295	5	52.834	5.494	20.254	0.856	33.761	262.144	8.078
	BRCA2 Var1	r20	81,722	32,325,508	8	62.351	12.830	17.708	0.908	34.417	131.072	5.091
Non-Match	BRCA1 Var1	r19	34,940	43,054,295	1	47.610	1.779	16.417	0.849	33.761	262.144	7.998
	BRCA1 Var2	r18	50,783	43,054,295	1	53.268	2.275	18.101	0.878	33.761	262.144	8.024
	BRCA2 Primary	r20	81,722	32,325,508	1	62.357	3.006	10.838	0.915	34.417	131.072	5.032

FIGURE 29—Summary of all costs for all applications evaluated in Reef. R1CS Constraints are for the step function in Nova. Times are averaged across 10 runs, standard deviation was less than 5% for all components and applications.

Application	Document ID	Regex ID	R1CS Constraints	Doc. Length	# Steps	Compile Time (s)	Solving Time (s)	Proving Time (s)	Verifying Time (s)	Proof Size (KB)	Commit Size (KB)	Max Memory Usage (GB)
Redactions												
	Small Email	r1	49,144	415	4	38.994	0.458	3.263	0.570	32.801	0.512	0.764
	Large Email	r2	75,812	1,000	6	223.189	1.726	5.735	0.685	33.585	1.024	1.073
ODoH												
	5f558	r3	22,573	128	3	20.769	0.131	2.022	0.429	31.953	0.512	0.354
	25424	r4	25,129	128	2	21.506	0.136	1.757	0.459	32.049	0.512	0.409
	55824	r5	25,576	128	1	21.813	0.054	1.396	0.456	32.017	0.512	0.415
	21d97	r6	22,193	128	2	19.683	0.066	1.672	0.428	31.857	0.512	0.352
	49b9a	r7	22,217	128	2	19.671	0.065	1.687	0.431	31.857	0.512	0.354
	b8f74	r8	22,094	128	2	19.662	0.067	1.694	0.429	31.889	0.512	0.357
	3b4ed	r9	21,009	128	2	21.501	0.062	1.677	0.426	31.857	0.512	0.344
	24448	r10	22,020	128	2	21.976	0.065	1.681	0.425	31.825	0.512	0.354
	b329c	r11	21,138	128	2	18.893	0.061	1.668	0.427	31.825	0.512	0.330
	6f74a	r12	21,749	128	2	19.319	0.064	1.674	0.427	31.857	0.512	0.350
	83a9c	r13	21,305	128	2	19.128	0.064	1.685	0.424	31.857	0.512	0.344
	5410f	r14	20,515	128	1	18.945	0.036	1.352	0.418	31.793	0.512	0.346
	a0514	r15	20,589	128	1	18.779	0.036	1.346	0.430	31.761	0.256	0.347
	b4ebd	r16	21,149	128	3	19.677	0.092	1.991	0.431	31.857	0.512	0.340
Passwords												
Match	dcdc9	r17	21,002	12	5	18.814	0.157	2.653	0.421	31.697	0.128	0.349
	43db4	r17	21,002	12	5	18.837	0.149	2.673	0.423	31.697	0.128	0.343
	91edc	r17	21,002	12	5	18.759	0.148	2.641	0.413	31.697	0.128	0.346
	2bcf2	r17	21,002	12	5	18.796	0.150	2.662	0.424	31.697	0.128	0.348
	10bf0	r17	21,002	12	5	18.644	0.147	2.652	0.422	31.697	0.128	0.347
	aff42	r17	21,002	12	5	18.687	0.151	2.651	0.422	31.697	0.128	0.346
	edde7	r17	21,002	12	5	18.687	0.149	2.661	0.417	31.697	0.128	0.344
	1539c	r17	21,002	12	5	29.531	0.153	2.679	0.418	31.697	0.128	0.345
	7bfcc	r17	21,002	12	5	18.598	0.151	2.652	0.411	31.697	0.128	0.354
	dfa02	r17	21,002	12	5	19.013	0.150	2.652	0.423	31.697	0.128	0.350
Non-Match	e73ee	r17	21,721	8	7	22.067	0.336	3.281	0.431	31.793	0.128	0.344
	b5f3a	r17	21,721	8	6	22.711	0.290	3.025	0.426	31.793	0.128	0.345
	fd1e7	r17	21,401	6	6	19.478	0.287	2.973	0.426	31.729	0.064	0.342
	db267	r17	21,401	3	5	19.234	0.242	2.688	0.435	31.729	0.064	0.334
	40867	r17	21,721	8	6	19.718	0.291	3.004	0.428	31.793	0.128	0.345
	f4a98	r17	21,401	6	6	19.282	0.295	3.017	0.423	31.729	0.064	0.343
	7474f	r17	21,721	8	6	19.600	0.285	3.005	0.431	31.793	0.128	0.342
	b20ef	r17	21,401	6	6	19.128	0.291	2.989	0.431	31.729	0.064	0.344
	27ba9	r17	21,721	7	6	19.336	0.283	3.009	0.431	31.793	0.128	0.345
	304b5	r17	21,721	9	6	22.028	0.292	2.997	0.428	31.793	0.128	0.339
DNA												
Match	BRCA1 Var1	r18	44,698	43,054,295	2	92.504	2,074.962	11.644	0.900	34.305	262.144	16.848
	BRCA1 Var2	r19	71,818	43,054,295	4	174.760	7,829.746	17.003	1.053	35.089	262.144	17.173
	BRCA2 Var1	r20	96,296	32,325,508	8	73.533	13,407.752	14.952	0.976	35.057	131.072	9.442
Non-Match	BRCA1 Var1	r19	46,650	43,054,295	1	62.941	542.441	12.997	0.908	34.369	262.144	15.131
	BRCA1 Var2	r18	72,343	43,054,295	1	71.860	556.006	12.963	1.075	35.121	262.144	15.214
	BRCA2 Primary	r20	107,184	32,325,508	1	78.083	321.560	7.988	0.971	35.121	131.072	8.411

FIGURE 30—Summary of all costs for all applications evaluated using safa+nlookup. R1CS Constraints are for the step function in Nova. Times are averaged across 10 runs, standard deviation was less than 5% for all components and applications.

Application	Document ID	Regex ID	R1CS Constraints	Doc. Length	# Steps	Compile Time (s)	Solving Time (s)	Proving Time (s)	Verifying Time (s)	Proof Size (KB)	Commit Size (KB)	Max Memory Usage (GB)
Redactions												
	Small Email	r1	292,053	415	83	3,183.479	7,993.773	69.256	1.160	25.904	0.032	1.721
	Large Email	r2	1,320,713	1,000	100	29,984.607	101,133.269	200.144	4.219	27.280	0.032	5.848
ODoH												
	5f558	r3	110,532	128	16	146.789	118.867	9.130	0.443	24.528	0.032	0.925
	25424	r4	47,561	128	16	65.722	12.513	6.353	0.302	23.840	0.032	0.640
	55824	r5	39,301	128	16	59.451	7.123	6.007	0.304	23.840	0.032	0.639
	21d97	r6	24,841	128	16	51.159	1.512	5.173	0.223	23.152	0.032	0.637
	49b9a	r7	27,937	128	16	52.823	2.222	5.317	0.222	23.152	0.032	0.639
	b8f74	r8	33,097	128	16	55.699	3.772	5.736	0.297	23.840	0.032	0.639
	3b4ed	r9	28,969	128	16	53.648	2.500	5.302	0.224	23.152	0.032	0.639
	24448	r10	25,873	128	16	51.625	1.730	5.243	0.220	23.152	0.032	0.638
	b329c	r11	23,809	128	16	50.669	1.313	5.162	0.220	23.152	0.032	0.637
	6f74a	r12	23,809	128	16	50.710	1.314	5.186	0.218	23.152	0.032	0.637
	83a9c	r13	24,841	128	16	51.194	1.513	5.173	0.221	23.152	0.032	0.637
	5410f	r14	20,713	128	16	49.586	0.818	4.999	0.223	23.152	0.032	0.634
	a0514	r15	22,777	128	16	50.244	1.133	5.125	0.217	23.152	0.032	0.636
	b4ebd	r16	26,905	128	16	52.752	1.966	5.301	0.219	23.152	0.032	0.639

FIGURE 31—Summary of all costs for all applications evaluated with DFA+Recursion. R1CS Constraints are for the step function in Nova.

Application	Document ID	Regex ID	R1CS Constraints	Doc. Length	# Steps	Compile Time (s)	Solving Time (s)	Proving Time (s)	Verifying Time (s)	Proof Size (KB)	Commit Size (KB)	Max Memory Usage (GB)
Redactions												
	Small Email	r1	23,041,771.0	415	1	10,330.135	8,181.845	194.572	51.036	18.440	0.032	76.300
ODoH												
	5f558	r3	1,564,274.0	128	1	490.940	110.792	13.089	3.572	15.688	0.032	5.048
	25424	r4	557,263.0	128	1	193.364	11.508	7.202	1.778	15.000	0.032	2.064
	55824	r5	425,163.0	128	1	157.852	6.010	3.958	1.065	14.312	0.032	1.261
	21d97	r6	193,983.0	128	1	100.478	1.230	2.450	0.645	13.624	0.032	0.745
	49b9a	r7	243,519.0	128	1	114.146	1.960	2.523	0.663	13.624	0.032	0.871
	b8f74	r8	326,079.0	128	1	134.130	3.504	3.931	1.035	14.312	0.032	1.086
	3b4ed	r9	260,031.0	128	1	116.638	2.230	2.536	0.690	13.624	0.032	0.910
	24448	r10	210,495.0	128	1	104.763	1.435	2.514	0.649	13.624	0.032	0.746
	b329c	r11	177,471.0	128	1	95.964	1.035	2.426	0.640	13.624	0.032	0.741
	6f74a	r12	177,471.0	128	1	96.223	1.032	2.420	0.639	13.624	0.032	0.741
	83a9c	r13	193,983.0	128	1	100.632	1.229	2.456	0.637	13.624	0.032	0.745
	5410f	r14	127,935.0	128	1	84.745	0.549	1.728	0.425	12.936	0.032	0.738
	a0514	r15	156,009.0	128	1	89.062	0.821	2.434	0.628	13.624	0.032	0.692
	b4ebd	r16	227,007.0	128	1	108.855	1.670	2.478	0.639	13.624	0.032	0.747

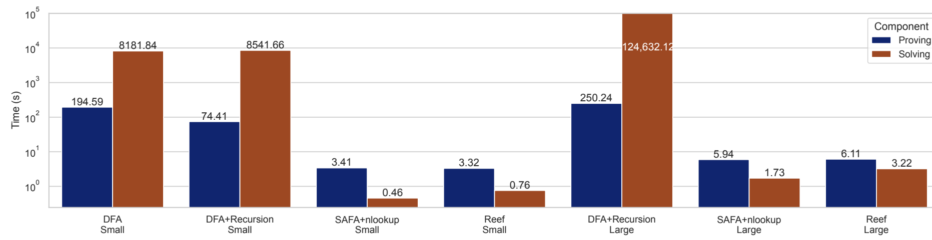
FIGURE 32—Summary of all costs for all applications evaluated in Reef using a DFA and no recursion. R1CS Constraints for the entire circuit.

Application	Regex ID	SAFA States	SAFA Transitions	DFA States	DFA Transitions
Redactions					
	r1	331	42,318	433	55,424
	r2	908	116,751	1,013	129,664
ODoH					
	r3	28	3,232	94	12,032
	r4	36	4,012	33	4,224
	r5	30	3,238	25	3,200
	r6	12	1,421	11	1,408
	r7	15	1,808	14	1,792
	r8	20	2,453	19	2,432
	r9	16	1,937	15	1,920
	r10	13	1,550	12	1,536
	r11	11	1,292	10	1,280
	r12	11	1,292	10	1,280
	r13	12	1,421	11	1,408
	r14	8	905	7	896
	r15	10	1,163	9	1,152
	r16	14	1,679	13	1,664
Passwords					
	r17	21	1,188	—	—
DNA					
	r18	331	42,318	43,052,484*	172,209,936*
	r19	331	42,318	43,050,383*	172,201,532*
	r20	976	4,861	32,318,453*	129,273,812*

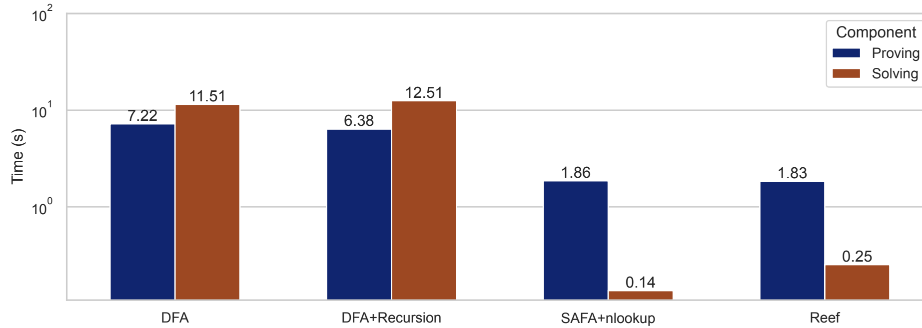
FIGURE 33—SAFA size vs DFA size for all evaluated regex
* are estimates

Application	Document ID	Regex ID	DFA	DFA # Steps	DFA + Recursion	DFA + Recursion # Steps	SAFA + nlookup	SAFA + nlookup # Steps	Reef	Reef # Steps
Redactions										
	Small Email	r1	23,041,771	1	292,053	83	49,144	4	46,655	4
	Large Email	r2	—	—	1,320,713	100	75,812	6	65,727	7
ODoH										
	5f558	r3	1,564,274	1	110,532	16	22,573	3	18,437	3
	25424	r4	557,263	1	47,561	16	25,129	2	22,692	2
	55824	r5	425,163	1	39,301	16	25,576	1	23,148	1
	21d97	r6	193,983	1	24,841	16	22,193	2	18,409	2
	49b9a	r7	243,519	1	27,937	16	22,217	2	18,433	2
	b8f74	r8	326,079	1	33,097	16	22,094	2	18,263	2
	3b4ed	r9	260,031	1	28,969	16	21,009	2	17,177	2
	24448	r10	210,495	1	25,873	16	22,020	2	18,865	2
	b329c	r11	177,471	1	23,809	16	21,138	2	18,241	2
	6f74a	r12	177,471	1	23,809	16	21,749	2	18,241	2
	83a9c	r13	193,983	1	24,841	16	21,305	2	17,785	2
	5410f	r14	127,935	1	20,713	16	20,515	1	17,617	1
	a0514	r15	156,009	1	22,777	16	20,589	1	17,365	1
	b4ebd	r16	227,007	1	26,905	16	21,149	3	17,617	3
Passwords										
Match	dcdc9	r17	—	—	—	—	21,002	5	19,982	5
	43db4	r17	—	—	—	—	21,002	5	19,982	5
	91edc	r17	—	—	—	—	21,002	5	19,982	5
	2bcf2	r17	—	—	—	—	21,002	5	19,982	5
	10bf0	r17	—	—	—	—	21,002	5	19,982	5
	aff42	r17	—	—	—	—	21,002	5	19,982	5
	edde7	r17	—	—	—	—	21,002	5	19,982	5
	1539c	r17	—	—	—	—	21,002	5	19,982	5
	7bfcc	r17	—	—	—	—	21,002	5	19,982	5
	dfa02	r17	—	—	—	—	21,002	5	19,982	5
Non-Match	e73ee	r17	—	—	—	—	21,721	7	20,728	7
	b5f3a	r17	—	—	—	—	21,721	6	20,728	6
	fd1e7	r17	—	—	—	—	21,401	6	20,725	6
	db267	r17	—	—	—	—	21,401	5	20,725	5
	40867	r17	—	—	—	—	21,721	6	20,728	6
	f4a98	r17	—	—	—	—	21,401	6	20,725	6
	7474f	r17	—	—	—	—	21,721	6	20,728	6
	b20ef	r17	—	—	—	—	21,401	6	20,725	6
	27ba9	r17	—	—	—	—	21,721	6	20,728	6
	304b5	r17	—	—	—	—	21,721	6	20,728	6
DNA										
Match	BRCA1 Var1	r18	—	—	—	—	44,698	2	35,306	2
	BRCA1 Var2	r19	—	—	—	—	71,818	4	50,783	4
	BRCA2 Var1	r20	—	—	—	—	96,296	8	81,722	8
Non-Match	BRCA1 Var1	r19	—	—	—	—	46,650	1	34,940	1
	BRCA1 Var2	r18	—	—	—	—	72,343	1	50,783	1
	BRCA2 Primary	r20	—	—	—	—	107,184	1	81,722	1

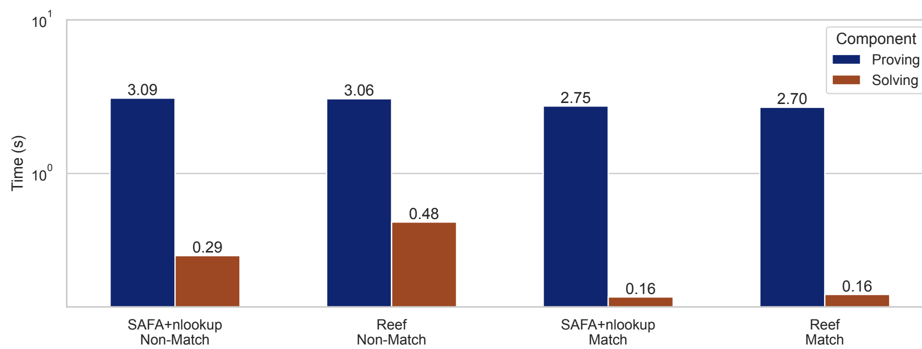
FIGURE 34—Total number of R1CS constraints for DFA, number for step function for DFA+Recursion, SAFA+nlookup, and Reef



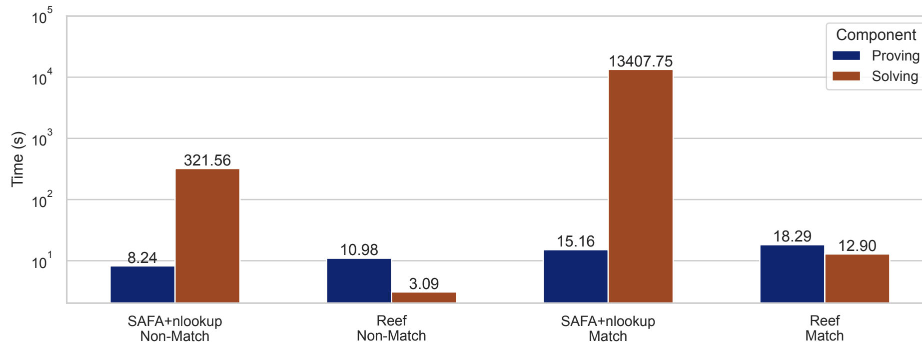
(a) Proof that a (small / large) committed email matches a redaction regex. DFA was unable to finish within 12 hours for the large email.



(b) Proof that a committed document matches an ODoH regex.



(c) Proof that a committed password matches/does not match a password strength regex. Neither DFA nor DFA+recursion can handle this application.



(d) Proof that a committed DNA document matches/does not match a DNA regex. Neither DFA nor DFA+recursion can handle this application.

FIGURE 35—Mean proving and solving time across 10 runs for proving that some committed document matches/does not match a regex with Reef and various alternatives. Standard deviations were less than 5% of the mean. Each subfigure describes a different application (regex) and type of document. The corresponding document sizes are found in Figure 6.

<table>
<tr><th>Application</th><th>Regex ID</th><th>Regex</th></tr>
<tr><td>Redactions</td><td></td><td></td></tr>
<tr><td></td><td>r1</td><td>^ Message-ID: .*[[:space:]] Date: Tue, 8 May 2001 09:16:00 -0700 (PDT) [[:space:]] From: .* [[:space:]] To: .* [[:space:]] Subject: Re: [[:space:]] Mime-Version: 1.0 [[:space:]] Content-Type: text/plain; charset=us-ascii [[:space:]] Content-Transfer-Encoding:7bit [[:space:]] X-From: Mike Maggi [[:space:]] X-To: Amanda Huble[[:space:]]X-cc: [[:space:]]X-bcc: [[:space:]] X-Folder:\\ Michael_Maggi_Jun2001 \\ Notes Folders \\ Sent[[:space:]]X-Origin: Maggi-M[[:space:]]X-FileName: mmaggi.nsf[[:space:]]*at 5:00$</td></tr>
<tr><td></td><td>r2</td><td>^ Message-ID: .*[[:space:]]Date: Tue, 11 Jul 2000 11:11:00 -0700 (PDT)[[:space:]]From: .*[[:space:]]To: .*[[:space:]] Subject: Reimbursement of Individually Billed Items[[:space:]]Mime-Version: 1.0[[:space:]]Content-Type: text/plain; charset=us-ascii[[:space:]]Content-Transfer-Encoding: 7bit[[:space:]]X-From: Enron Announcements[[:space:]]X-To: All Enron Employees North America[[:space:]]X-cc: [[:space:]]X-bcc: [[:space:]]X-Folder: \\Michelle_Lokay_Dec2000_ June2001_1\\Notes Folders\\Corporate[[:space:]]X-Origin: LOKAY-M[[:space:]]X-FileName: mlokay\.nsf[[:space:]]*The memo distributed on June 27 on Reimbursement of Individually Billed Items [[:space:]]requires[[:space:]]clarification\. The intent of the memo was to give employees an alternate [[:space:]]method[[:space:]]of paying for pagers, cell phones, etc\. Employees can continue to submit[[:space:]]these[[:space:]]invoices to Accounts Payable for processing or pay these items with their [[:space:]]corporate[[:space:]]American Express card and request reimbursement through an expense report\. [[:space:]]Either[[:space:]]way is an acceptable way to process these small dollar high volume invoices\.$</td></tr>
<tr><td>ODoH</td><td></td><td></td></tr>
<tr><td></td><td>r3</td><td>^ad([sxv]?[0-9]*|system)[_.-]([^.[[:space:]]+.){1,}|[_.-]ad([sxv]?[0-9]*|system)[_.-]|$</td></tr>
<tr><td></td><td>r4</td><td>^(.+[_.-])?adse?rv(er?|ice)?s?[0-9]*[_.-]</td></tr>
<tr><td></td><td>r5</td><td>^(.+[_.-])?telemetry[_.-]</td></tr>
<tr><td></td><td>r6</td><td>^(adim(age|g)s?[0-9]*[_.-]</td></tr>
<tr><td></td><td>r7</td><td>^(adtrack(er|ing)?[0-9]*[_.-]</td></tr>
<tr><td></td><td>r8</td><td>^(advert(s|is(ing|ements?))?[0-9]*[_.-]</td></tr>
<tr><td></td><td>r9</td><td>^(aff(iliat(es?|ion))?[_.-]</td></tr>
<tr><td></td><td>r10</td><td>^(analytics?[_.-]</td></tr>
<tr><td></td><td>r11</td><td>^(banners?[_.-]</td></tr>
<tr><td></td><td>r12</td><td>^(beacons?[0-9]*[_.-]</td></tr>
<tr><td></td><td>r13</td><td>^(ount(ers?)?[0-9]*[_.-]</td></tr>
<tr><td></td><td>r14</td><td>^(mads.</td></tr>
<tr><td></td><td>r15</td><td>^(pixels?[-.]</td></tr>
<tr><td></td><td>r16</td><td>^(stat(s|istics)?[0-9]*[_.-]</td></tr>
<tr><td>Passwords</td><td></td><td></td></tr>
<tr><td></td><td>r17</td><td>^(?=.*[A-Z].*[A-Z])(?=.*[!%^@#$&*])(?=.*[0-9].*[0-9])(?=.*[a-z].*[a-z].*[a-z]).{12}$</td></tr>
<tr><td>DNA</td><td></td><td></td></tr>
<tr><td></td><td>r18</td><td>^.{43052424}ATGGGCTACAGAAACCGTGCCAAAAGACTTCTACAGAGTGAACCCGAAAATCCTTCCTTG</td></tr>
<tr><td></td><td>r19</td><td>^.{43050079}ATGCTGAAACTTCTCAACCAGAAGAAAGGGCCTTCACAGTGTCCTTTATGTAAGAATGATATAACCAAAAG.*AGCCTACAAGAAAGTACGAGATTTAGTCAACTTGTTGAAGAGCTATTGAAAATCATTTGTGCTTTTCAGCTTGACACAGGTTTGGAGT.*ATGCAAACAGCTATAATTTTGCAAAAAAGGAAAATAACTCTCCTGAACATCTAAAAGATGAAGTTTCTATCATCCAAAGTATGGGCTACAGAAACCGTGCCAAAAGACTTCTACAGAGTGAACCCGAAAATCCTTCCTTG</td></tr>
<tr><td></td><td>r20</td><td>^.{32317478}CACAACTAAGGAACGTCAAGAGATACAGAATCCAAATTTTACCGCACCTGGTCAAGAATTTCTGTCTAAATCTCATTTGTATGAACATCTGACTTTGGAAAAATCTTCAAGCAATTTAGCAGTTTCAGGACATCCATTTTATCAAGTTTCTGCTACAAGAAATGAAAAAATGAGACACTTGATTACTACAGGCAGACCAACCAAAGTCTTTGTTCCACCTTTTAAAACTAAATCACATTTTCACAGAGTTGAACAGTGTGTTAGGAATATTAACTTGGAGGAAAACAGACAAAAGCAAAACATTGATGGACATGGCTCTGATGATAGTAAAAATAAGATTAATGACAATGAGATTCATCAGTTTAACAAAAACAACTCCAATCAAGCAGTAGCTGTAACTTTCACAAAGTGTGAAGAAGAACCTTTAG.*ATTTAATTACAAGTCTTCAGAATGCCAGAGATATACAGGATATGCGAATTAAGAAGAAACAAAGGCAACGCGTCTTTCCACAGCCAGGCAGTCTGTATCTTGCAAAAACATCCACTCTGCCTCGAATCTCTCTGAAAGCAGCAGTAGGAGGCCAAGTTCCCTCTGCGTGTTCTCATAAACAG.*CTGTATACGTATGGCGTTTCTAAACATTGCATAAAAATTAACAGCAAAAATGCAGAGTCTTTTCAGTTTCACACTGAAGATTATTTTGGTAAGGAAAGTTTATGGACTGGAAAAGGAATACAGTTGGCTGATGGTGGATGGCTCATACCCTCCAATGATGGAAAGGCTGGAAAAGAAGAATTTTATAG.*GGCTCTGTGTGACACTCCAGGTGTGGATCCAAAGCTTATTTCTAGAATTTGGGTTTATAATCACTATAGATGGATCATATGGAAACTGGCAGCTATGGAATGTGCCTTTCCTAAGGAATTTGCTAATAGATGCCTAAGCCCAGAAAGGGTGCTTCTTCAACTAAAATACAG</td></tr>
</table>

FIGURE 36—Regexs with ID