From Rational Number Reconstruction to Set Reconciliation

Antoine Amarilli, Fabrice Ben Hamouda, Florian Bourse,
Robin Morisset, David Naccache, and Pablo Rauzy

École normale supérieure, Département d'informatique
45, rue d'Ulm, F-75230, Paris Cedex 05, France.
`surname.name@ens.fr` (except for `fabrice.ben.hamouda@ens.fr`)

Si cet abstract vous convient, il faudrait en reprendre des bouts dans l'introduction.

Abstract. This work revisits *set reconciliation*, a problem consisting in synchronizing two multisets of fixed-size values while minimizing the amount of data transmitted. We propose a new number theoretic reconciliation protocol called "Divide & Factor" (D&F) which achieves optimal asymptotic transmission complexity like prior proposals. We then study the problem of synchronizing sets of variable-size files, and describe how constant-factor improvements can be achieved through the use of hashing with a carefully chosen hash size (balancing the quantity of data transferred and the risk of collisions). We show how this process can be applied to synchronize file hierarchies, taking into account the location of files. We describe `btrsync`, our open-source implementation of the protocol, and benchmark it against the popular software `rsync` to demonstrate that `btrsync` uses more CPU time but transmits less data.

Fabrice: j'écrirai quelque chose de plus positif comme: "`btrsync` transmits much less data at the expense of a small computationnal overhead."

1 Introduction

This work revisits *set reconciliation*, a problem consisting in synchronizing two multisets while minimizing the amount of data transmitted. Set reconciliation arises in many practical situations, the most typical of which is certainly incremental backups performed over a slow network link.

Several efficient and elegant solutions are known to achieve set reconciliation of multisets containing atomic elements of a fixed size. For instance, [8] manages to perform set reconciliation using a bandwidth which is linear in the size of the symmetric difference of the multisets multiplied by the size of the elements, which is optimal in this setting. We refer the reader to [8,9,5] (to quote a few references) for more on this problem's history and its existing solutions.

Est-ce que c'est vrai qu'on apporte quelque chose de nouveau dans le cas où c'est des fichiers de taille non fixe ? — Fabrice: je ne sais pas, il faudrait que l'on réanalyse la complexité propre de la dernière variante.

However, in the case where the elements to be synchronized can be very large (e.g., files during a backup), we must use checksums to identify the differing files before transferring them, and the question of the size of the checksum to use is non-trivial. In this article, we propose a new reconciliation protocol called "Divide & Factor" (D&F) based on number theory. In terms of asymptotic transmission complexity, the proposed procedure reaches optimality as well. In addition, the new protocols offer a very interesting gamut of parameter trade-offs. We provide an analysis of the protocol's complexity in terms of transmission and computation, as well as a probabilistic analysis of the possible choices of checksum sizes; we also provide an implementation of the protocol and experimental results.

Fabrice@Antoine: Je ne suis pas tout à fait d'accord avec ce paragraphe, ton abstract est beaucoup mieux...

This paper is structured as follows: Section 2.2 presents a basic version of the proposed protocol. This basic version suffers from two limitations: it works only if the number of differences to reconcile is bound and it may fail leave the synchronized party in an erroneous state. Failure avoidance is overcome in section 2.3 and an extension to arbitrary numbers of differences is given in section 2.4. The protocol's transmission complexity is treated in section 3. Section 3 also introduces two transmission optimizations and analyzes them in detail. Section 4 analyzes the computational complexities of the proposed protocols and 6 reports practical experiments and benchmarks against the popular software `rsync`.

2 "Divide & Factor" Set Reconciliation

2.1 Problem Definition and Notations

Oscar possesses an old version of a directory $\mathfrak{D}$ that he wishes to update. Neil has the new, up-to-date version $\mathfrak{D}'$: $\mathfrak{D}$ and $\mathfrak{D}'$ can differ both in their files and in their tree structures. Oscar wishes to obtain $\mathfrak{D}'$ but *exchange as little data as possible* during the synchronization process.

To tackle this problem we separate the *"what"* from the *"where"* by considering files as a tuple of their location and content. In other words, we will first synchronize all the file contents and then move files to the adequate location. We consider that $\mathfrak{D}$ is a multiset of files which we denote as $\mathfrak{F} = \{F_0, \ldots, F_n\}$, and likewise represent $\mathfrak{D}'$ as $\mathfrak{F}' = \{F'_0, \ldots, F'_{n'}\}$.

Fabrice: NON, ce n'est pas vrai: on considère vraiment qu'un fichier est "path + content".

Let t_0 be the number of discrepancies between $\mathfrak{F}$ and $\mathfrak{F}'$ that Oscar wishes to learn, *i.e.* the symmetric difference of $\mathfrak{F}$ and $\mathfrak{F}'$:

Fabrice: Qu'est-ce que $\mathfrak{D}$? Est-ce $\mathfrak{F}$? Dans tous les cas, il s'agit d'un set, vu la représentation indiquée

$$t_0 = \#\mathfrak{F} + \#\mathfrak{F}' - 2\#\left(\mathfrak{F}\bigcap\mathfrak{F}'\right) = \#\left(\mathfrak{F}\bigcup\mathfrak{F}'\right) - \#\left(\mathfrak{F}\bigcap\mathfrak{F}'\right)$$

Given a file F, we denote by $\texttt{Hash}(F)$ its image by a collision-resistant hash function such as SHA-1. Let $\texttt{HashPrime}(F)$[1] be a function hashing files (uniformly) into primes smaller than 2^u for some $u \in \mathbb{N}$. Define the shorthand notations: $h_i = \texttt{HashPrime}(F_i)$ and $h'_i = \texttt{HashPrime}(F'_i)$.

Fabrice: `Hash` can be introduced latter when needed (section 2.3)

2.2 Description of the Basic Exchanges

The number of differences t_0 is unknown to Oscar and Neil. However, for the time being, we will assume that t_0 is smaller than some t and attempt to perform synchronization. If $t_0 \leq t$, synchronization will succeed; if $t_0 > t$ the parties will transmit more information later to complete the synchronization, as explained in section 2.4.

We generate a prime p such that:

$$2^{2ut} \leq p < 2^{2ut+1} \tag{1}$$

[1] The design of `HashPrime` is addressed in Appendix C.

Given $\mathfrak{F}$, Oscar generates and sends to Neil the redundancy:

$$c = \prod_{F_i \in \mathfrak{F}} \texttt{HashPrime}(F_i) = \prod_{i=1}^{n} h_i \bmod p$$

Neil computes:

$$c' = \prod_{F'_i \in \mathfrak{F}'} \texttt{HashPrime}(F'_i) = \prod_{i=1}^{n'} h'_i \bmod p \quad \text{and} \quad s = \frac{c'}{c} \bmod p$$

Using [13] the integer s can be written as:

Fabrice: Quel est l'intérêt de cette citation ???

$$s = \frac{a}{b} \bmod p \text{ where the } G_i \text{ denote files and } \begin{cases} a = \prod_{G_i \in \mathfrak{F}' \wedge G_i \notin \mathfrak{F}} \texttt{HashPrime}(G_i) \\ b = \prod_{G_i \notin \mathfrak{F}' \wedge G_i \in \mathfrak{F}} \texttt{HashPrime}(G_i) \end{cases}$$

Note that if our assumption $t_0 \leq t$ is correct, $\mathfrak{F}$ and $\mathfrak{F}'$ differ by at most t elements and a and b are strictly less than 2^{ut}. The problem of recovering a and b from s efficiently is known as *Rational Number Reconstruction* [10,14]. theorem 1 (see [6]) guarantees that it can be solved in this setting. The following theorem is a slightly modified version of Theorem 1 in [6]:

Fabrice: je ne comprends pas le "where the G_i denote files..."

Theorem 1. *Let $a, b \in \mathbb{Z}$ two co-prime integers such that $0 \leq a \leq A$ and $0 < b \leq B$. Let $p > 2AB$ be a prime and $s = ab^{-1} \bmod p$. Then a, b are uniquely defined given s and p, and can be recovered from A, B, s, p in polynomial time.*

Taking $A = B = 2^{ut} - 1$, Equation (1) implies that $AB < p$. Moreover, $0 \leq a \leq A$ and $0 < b \leq B$. Thus Oscar can recover a and b from s in polynomial time: a possible option is to use Gauss algorithm for finding the shortest vector in a bi-dimensional lattice [13]. By testing the divisibility of a and b by the h_i and the h'_i, Neil and Oscar can attempt to identify the discrepancies between $\mathfrak{F}$ and $\mathfrak{F}'$ and settle them.

Fabrice: certes, mais on utilise directement un Euclide étendu tronqué. Et pourquoi citer Vallée qui est un peu incompréhensible dans notre cas... TODO: check that in our program we ensure a and b co-prime !! otherwise, it may fail !!!!!

The formal description of the protocol is given in Figure 1. The "output $\perp_{\mathsf{bandwidth},\square}$ " protocol interruptions will:

- never occur if the assumption $t_0 \leq t$ holds.
- occur with high probability if $t_0 > t$. Indeed, for a potential $\perp_{\mathsf{bandwidth},1}$ to be overlooked, the ut-bit number a must perfectly factor over a set of n primes of size u. If we assume that a is "random", the probability γ that a is divisible by some h_i is essentially $\gamma \sim 1/h_i \sim 2^{-u}$, the probability that a is divisible by exactly t digests is:

$$\alpha = \binom{n}{t} \gamma^t (1-\gamma)^{n-t} \sim \binom{n}{t} 2^{-ut} (1 - 2^{-u})^{n-t}$$

Oscar		**Neil**
compute c		
	$\xrightarrow{c}$	
		compute a, b
		if a doesn't factor as a product of h'_is
		then output $\perp_{\mathsf{bandwidth},1}$ and halt
		$\mathfrak{S} \leftarrow \{F'_i \text{ s.t. } a \bmod h'_i = 0\}$
	$\xleftarrow{\mathfrak{S},b}$	
if b doesn't factor as a product of h_i's		
then output $\perp_{\mathsf{bandwidth},2}$ and halt		
delete files s.t. $b \bmod h_i = 0$		
add $\mathfrak{S}$ to the disk		

Fig. 1. Basic Protocol.

and the probability that the protocol does not terminate by a $\perp_{\mathsf{bandwidth},\Box}$ when $t_0 > t$ is $\sim \alpha^2$.

However, we are not interested in the fact a is divisible by "exactly" t digests but the fact that a can be factorized over of the basis of h_i...

The very existence of $\perp_{\mathsf{bandwidth},\Box}$'s is annoying for two reasons:

– A file synchronization procedure that works *only* for a limited number of differences is not really useful in practice. Thus, section 2.4 explains how to extend the protocol to perform the synchronization even when the number of differences t_0 exceeds the initial estimation t.
– If, by sheer bad luck, both $\perp_{\mathsf{bandwidth},\Box}$'s went undetected (double accidental factorization) the Basic Protocol (Fig. 1) may leave Oscar in an inconsistent state.

Double accidental factorization is not only possible source of inconsistent states: as we did not specifically require `HashPrime` to be collision-resistant, the events

$$\perp_{\mathsf{collision},1} = \begin{cases} h'_i = h'_j \text{ for } i \neq j \\ \\ a \bmod h_i = 0 \end{cases} \quad \text{and/or} \quad \perp_{\mathsf{collision},2} = \begin{cases} h_i = h_j \text{ for } i \neq j \\ \\ b \bmod h'_i = 0 \end{cases}$$

will cause Neil to send wrong files in $\mathfrak{S}$ ($\perp_{\mathsf{collision},1}$) and/or have Oscar unduely delete files owned by Neil ($\perp_{\mathsf{collision},2}$).

Inconsistent states may hence stem from three events:

- accidental double factorization of a and/or b when $t_0 > t$ (probability α^2)
- $\perp_{\mathsf{collision},1}$ = collisions within the set $\{h'_i\}$
- $\perp_{\mathsf{collision},2}$ = collisions within the set $\{h_i\}$

Fabrice: ce qui est un peu tordu, c'est que les collisions qui nous intéressent sont les collisions entre $\mathfrak{S}$ et $\mathfrak{F} \cup \mathfrak{F}'$... il faut qu'on en discute

Section 2.3 explains how protect the protocol from all inconsistent events at once.

2.3 Avoiding Inconsistency

The Basic Protocol of Figure 1 is fully deterministic. Hence if any sort of trouble occurs, repeating the protocol will be of no help. We modify the protocol as follows:

- Let $H \leftarrow \texttt{Hash}(\mathfrak{F}')$
- Replace $\texttt{HashPrime}(F)$ by a diversified $\hbar_k(F) = \texttt{HashPrime}(k|F)$.
- Define the shorthand notations: $\hbar_{k,i} = \hbar_k(F_i)$ and $\hbar'_{k,i} = \hbar_k(F'_i)$.
- Let $\mathsf{StepProtocol}(k)$ denote the sub-protocol shown in Figure 2.
- Use the protocol of Figure 3 as a fully functional reconciliation protocol for $t_0 \leq t$.

Fabrice: This notation (hash of a set) is not defined... and it may be useful to recall the definition of `Hash` *here (collision-resistant) if the reader has forgotten it...*

Oscar		Neil
		if a doesn't factor as a product of $\hbar'_{k,i}$s then output $\perp_{\mathsf{bandwidth},1}$ and halt $\mathfrak{S} \leftarrow \{F'_i \text{ s.t. } a \bmod \hbar'_{k,i} = 0\}$ if there are collisions in $\mathfrak{S}$ then output $\perp_{\mathsf{collision},1}$ and halt
	$\xleftarrow{\mathfrak{S},b}$	
if b doesn't factor as a product of $\hbar_{k,i}$'s then output $\perp_{\mathsf{bandwidth},2}$ and halt $\mathfrak{A} \leftarrow \{F_i \text{ s.t. } b \bmod \hbar_{k,i} = 0\}$ if there are collisions in $\mathfrak{A}$ then output $\perp_{\mathsf{collision},2}$ and halt if $H \neq \texttt{Hash}(\mathfrak{F} \bigcup \mathfrak{S} - \mathfrak{A})$ then output $\perp_{\mathsf{bandwidth},3}$ and halt add $\mathfrak{S}$ to the disk and erase $\mathfrak{A}$ from the disk return $\mathsf{success}$		

Fig. 2. $\mathsf{StepProtocol}(k)$.

Note: To avoid transmitting the (potentially very voluminous) $\mathfrak{S}$ during $\mathsf{StepProtocol}$ before knowing if one of the errors $\perp_{\mathsf{bandwidth},2}, \perp_{\mathsf{bandwidth},3}, \perp_{\mathsf{collision},2}$ will occur, Neil may transmit

$$\mathfrak{S}' = \{\texttt{Hash}(F'_i),\ F'_i \in \mathfrak{S}\}$$

instead of $\mathfrak{S}$ and send $\mathfrak{S}$ only after successfully passing the $\perp_{\mathsf{bandwidth},3}$ test. The definition of H must be changed accordingly to

$$H = \texttt{Hash}(\{\texttt{Hash}(F'_i),\ F'_i \in \mathfrak{F}'\})$$

Fabrice: en même temps, le Hash d'un ensemble, que l'on n'a pas définit, a tout intérêt à déjà être de cette forme, sinon, on a des problèmes pour avoir une concaténation

Oscar		Neil
		compute $H \leftarrow \mathtt{Hash}(\mathfrak{F}')$
	$\xleftarrow{H}$	
compute c		
	$\xrightarrow{c}$	
		compute a, b $k \leftarrow 1$
	$\xleftarrow{\mathsf{StepProtocol}(k)}$	while $\mathsf{StepProtocol}(k) = \perp_{\mathsf{collision},\square}$ $k \leftarrow k + 1$

Fig. 3. Fully Functional Protocol for $t_0 \leq t$.

2.4 Handling a High Number of Differences

To extend the protocol to an arbitrary t_0, Oscar and Neil agree on an infinite set of primes $p_1, p_2, \ldots$ As long as the protocol fails with a $\perp_{\mathsf{bandwidth},\square}$ status, Neil and Oscar redo the protocol with a new p_i and Neil will keep accumulating information about the difference between $\mathfrak{F}$ and $\mathfrak{F}'$ as shown in Appendix A. Each of this repetition is called a round. Note that no information is lost and that the transmitted modular knowledge about the difference adds up until it reaches a threshold sufficient to reconcile $\mathfrak{F}$ and $\mathfrak{F}'$.

More precisely, let us suppose $2^{2ut_i} \leq p_i < 2^{2ut_i+1}$. Let us write $P_i = p_1 \ldots p_i$ and $T_i = u(t_1 + \ldots t_i)$. After receiving the redundancies $c_1, \ldots, c_i$ corresponding to $p_1, \ldots, p_i$, Neil has as many information as if Oscar had transmitted a redundancy C_i corresponding to the modulo P_i, and can compute $S_i = C'_i/C_i$ from $s_i = c'_i/c_i$ and S_{i-1} using the CRT (TODO ref ?). Therefore, the number λ of rounds used is the minimum number λ such that $T_i \geq t_0$. If $t_1 = t_2 = \cdots = t$, then $\lambda = \lceil t/t_0 \rceil$.

All $\perp$ treatments were removed from Appendix A for the sake of clarity (these can be very easily added by modifying Appendix A *mutatis mutandis*). In essence, the rules are: add information modulo a new p_i whenever the protocol fails with a $\perp_{\mathsf{bandwidth},\square}$ and increment k whenever the protocol fails with a $\perp_{\mathsf{collision},\square}$.

Fabrice: plus clair si on le met dans une deuxième figure en appendix quand même, je pense...

A typical execution sequence is thus expected to be something like:

$$\perp_{\mathsf{bandwidth},1}, \perp_{\mathsf{bandwidth},1}, \perp_{\mathsf{bandwidth},1}, \perp_{\mathsf{bandwidth},1}, \perp_{\mathsf{collision},1}, \perp_{\mathsf{collision},1}, \mathsf{success}$$

3 Transmission Complexity

This section explores two strategies for reducing the size of p and hence improving transmission by *constant factors* (from an asymptotic communication standpoint, improvements cannot be expected as the protocol already transmits information proportional to t_0, the difference to settle). Excluding the core information $\mathfrak{S}$ and assuming that no $\perp_{\mathsf{collision},\square}$ events occurred, the transmission complexity of the protocol of Appendix A is:

$$\lambda \log(\max_{i=1}^{\lambda} c_i) + \log b \leq \lambda \log(\max_{i=1}^{\lambda} p_i) + \frac{1}{2} \log \prod_{i=1}^{\lambda} p_i \leq 3\lambda(ut_0 + 1) = O(\lambda ut_0) = O(ut),$$

As we have no control over t, decreasing u is the main natural optimization option. We will get back to this later on in this paper (section 3.2).

3.1 Probabilistic Decoding: Reducing p

Generate a prime p about twice shorter than the p recommended in section 2.2, namely:

$$2^{ut} < p \leq 2^{ut+1} \tag{2}$$

Let $\eta = \max(n, n')$. The new redundancy c is calculated as previously and is hence also approximately twice smaller. Namely:

$$s = \frac{a}{b} \bmod p \text{ and } \begin{cases} a = \prod\limits_{G_i \in \mathfrak{F}' \wedge G_i \notin \mathfrak{F}} \texttt{HashPrime}(G_i) \\ b = \prod\limits_{G_i \notin \mathfrak{F}' \wedge G_i \in \mathfrak{F}} \texttt{HashPrime}(G_i) \end{cases}$$

and since there are at most t differences, we must have:

$$ab \leq 2^{ut} \tag{3}$$

By opposition to section 2.2 we do not have a fixed bound for a and b anymore; Equation (3) only provides a bound for the *product* ab. Therefore, we define a sequence of $t+1$ couples of bounds:

$$(A_i, B_i) = \left(2^{ui}, 2^{u(t-i)}\right) \forall i \in \{0, \ldots, t\}$$

Equations (2) and (3) imply that there must exist at least one index i such that $0 \leq a \leq A_i$ and $0 < b \leq B_i$. Then using Theorem 1, since $A_i B_i = 2^{ut} < p$, given (A_i, B_i, p, s) one can recover (a, b), and hence the difference between $\mathfrak{F}$ and $\mathfrak{F}'$.

The problem is that (unlike section 2.2) we have no guarantee that such an (a, b) is unique. Namely, we could (in theory) stumble over an $(a', b') \neq (a, b)$ satisfying (3) for some index $i' \neq i$. We conjecture that such failures happen with negligible probability (that we do not try to estimate here) when u is large enough, but this makes the modified protocol heuristic only. If failures never occur, this variant will roughly halve the quantity of transmitted bits with respect to section 2.2.

Fabrice: "heuristic" $\longrightarrow$ not really, because of the final hash verification, though we cannot compute exactly the complexity...

3.2 The File Laundry: Reducing u

What happens if we brutally shorten u in the basic Divide & Factor protocol? As expected by the birthday paradox, we should start seeing collisions. Let us analyze the statistics governing the appearance of collisions.

Fabrice: What is η exactly ? n, n', $n+n'$, $|\mathfrak{F} \cup \mathfrak{F}'|$? I prefer the last one actually, i.e., the total number of files...

Consider HashPrime as a random function from $\{0,1\}^*$ to $\{0,\ldots,2^u-1\}$. Let X_i be the random variable:

Fabrice: maybe say it in another way, since we already do not suppose HashPrime is collision-resistant

$$X_i = \begin{cases} 1 & \text{if file } F_i \text{ collides with another file.} \\ 0 & \text{otherwise.} \end{cases}$$

Clearly, we have $\Pr[\,X_i = 1\,] \le \frac{\eta-1}{2^u}$. The average number of colliding files is hence:

$$\mathbb{E}\left[\sum_{i=0}^{\eta-1} X_i\right] \le \sum_{i=0}^{\eta-1} \frac{\eta-1}{2^u} = \frac{\eta(\eta-1)}{2^u}$$

For instance, for $\eta = 10^6$ files and 32-bit digests, the expected number of colliding files is less than 233.

However, it is important to note that a collision can only yield a *false positive*, and never a *false negative*. In other words, while a collision may obliviate a difference[2] a collision will never create a nonexistent difference *ex nihilo*.

Fabrice: difference or discrepancy ?

Thus, it suffices to replace $\mathtt{HashPrime}(F)$ by a diversified $\hbar_\ell(F) = \mathtt{HashPrime}(\ell|F)$ to quickly filter-out file differences by repeating the protocol for $k = 1, 2, \ldots$ At each iteration the parties will detect new files and new deletions, fix these and "launder" again the remaining multisets.

Fabrice: already said in Section 2.3... and maybe it is better to say that, in this first analysis, we suppose we do not filter out files, because this only improves the algo...

Assume that the diversified $\hbar_k(F)$'s are random and independent. To understand why the probability that a stubborn file persists colliding decreases exponentially with the number of iterations k, assume that η remains invariant between iterations and define the following random variables:

$$X_i^\ell = \begin{cases} 1 & \text{if file } F_i \text{ collides with another file during iteration } \ell. \\ 0 & \text{otherwise.} \end{cases}$$

$$Y_i = \prod_{\ell=1}^{k} X_i^\ell = \begin{cases} 1 & \text{if file } F_i \text{ collides with another file during all the } k \text{ first protocol iterations.} \\ 0 & \text{otherwise.} \end{cases}$$

Fabrice je ne vois pas la difference entre X_i^ℓ dans cette section et Z_i^ℓ dans la suivante. Peux-tu preciser STP

[2] *e.g.* make the parties blind to the difference between `index.htm` and `iexplore.exe`.

By independence, we have:

$$\Pr\left[\,Y_i = 1\,\right] = \prod_{\ell=1}^{k} \Pr\left[\,X_i^\ell = 1\,\right] = \Pr\left[\,X_i^1 = 1\,\right] \ldots \Pr\left[\,X_i^k = 1\,\right] \leq \left(\frac{\eta - 1}{2^u}\right)^k$$

Therefore the average number of colliding files is:

$$\mathbb{E}\left[\sum_{i=0}^{\eta-1} Y_i\right] \leq \sum_{i=0}^{\eta-1} \left(\frac{\eta - 1}{2^u}\right)^k = \eta \left(\frac{\eta - 1}{2^u}\right)^k$$

And the probability that at least one false positive will survive k rounds is:

$$\epsilon_k \leq \eta \left(\frac{\eta - 1}{2^u}\right)^k$$

For the previously considered instance[3] we get $\epsilon_2 \leq 5.43\%$ and $\epsilon_3 \leq 2 \cdot 10^{-3}\%$.

A more refined (but somewhat technical) analysis. As mentioned previously, the parties can remove the files confirmed as different during iteration k and work during iteration $k+1$ only with common and colliding files. Now, the only collisions that can fool round k, are the collisions of file-pairs (F_i, F_j) such that F_i and F_j have both already collided during *all the previous iterations*[4]. We call such collisions "masquerade balls" (*cf.* Figure 4). Define the two random variables:

$$Z_i^\ell = \begin{cases} 1 & \text{if } F_i \text{ participated in masquerade balls during all the } \ell \\ & \text{first protocol iterations.} \\ 0 & \text{otherwise.} \end{cases}$$

$$X_{i,j}^\ell = \begin{cases} 1 & \text{if files } F_i \text{ and } F_j \text{ collide during iteration } \ell. \\ 0 & \text{otherwise.} \end{cases}$$

[3] $\eta = 10^6$,$u = 32$.

[4] Note that we do not require that F_i and F_j repeatedly collide *which each other*. *e.g.* we may witness during the first round $\hbar_1(F_1) = \hbar_1(F_2)$, $\hbar_1(F_3) = \hbar_1(F_4)$ and $\hbar_1(F_5) = \hbar_1(F_6)$ while during the second round $\hbar_2(F_1) = \hbar_2(F_2)$, $\hbar_2(F_3) = \hbar_1(F_6)$ and $\hbar_2(F_5) = \hbar_2(F_4)$ as shown in Figure 4.

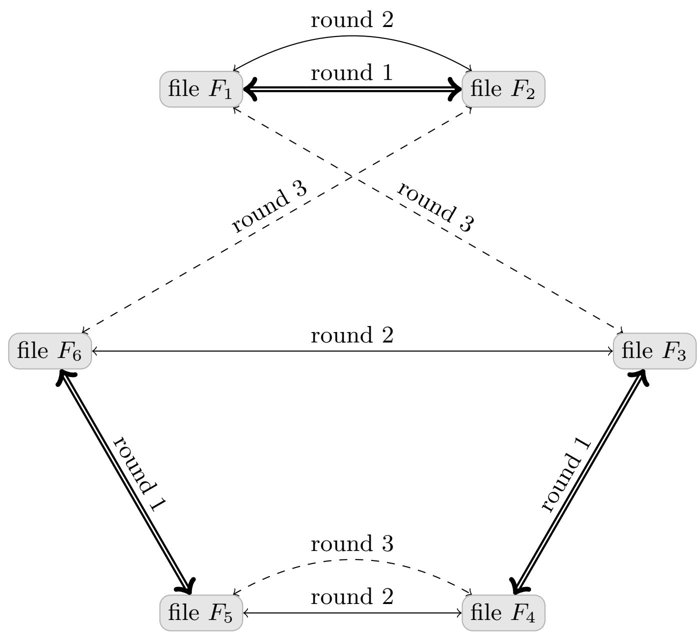

Fig. 4. Illustration of three masquerade balls. Each protocol round is materialized by a different type of arrow. Arrows denotes collisions.

Set $Z_i^0 = 1$ and write $p_\ell = \Pr\left[\, Z_i^\ell = 1 \text{ and } Z_j^\ell = 1 \,\right]$ for all ℓ and $i \neq j$. For $k \geq 1$, we have:

$$
\begin{aligned}
\Pr\left[\, Z_i^k = 1 \,\right] &= \Pr\left[\, \exists j \neq i,\ X_{i,j}^k = 1,\ Z_i^{k-1} = 1 \text{ and } Z_j^{\ell-1} = 1 \,\right] \\
&\leq \sum_{j=0, j\neq i}^{\eta-1} \Pr\left[\, X_{i,j}^k = 1 \,\right] \Pr\left[\, Z_i^{k-1} = 1 \text{ and } Z_j^{k-1} = 1 \,\right] \\
&\leq \frac{\eta - 1}{2^u} p_{k-1}
\end{aligned}
$$

Fabrice: maybe put this in appendix and add a few comments...

Furthermore $p_0 = 1$ and

$$
\begin{aligned}
p_\ell &= \Pr\left[X^\ell_{0,1}=1,\ Z^\ell_0=1 \text{ and } Z^\ell_1=1\right] + \Pr\left[X^\ell_{0,1}=0,\ Z^\ell_0=1 \text{ and } Z^\ell_1=1\right] \\
&\le \Pr\left[X^\ell_{0,1}=1,\ Z^{\ell-1}_0=1 \text{ and } Z^{\ell-1}_1=1\right] \\
&\quad + \sum_{i\ge 2, j\ge 2} \Pr\left[X^\ell_{0,i}=1,\ X^\ell_{1,j}=1,\ Z^{\ell-1}_0=1 \text{ and } Z^{\ell-1}_1=1\right] \\
&= \Pr\left[X^\ell_0=X^\ell_1\right]\Pr\left[Z^{\ell-1}_0=1 \text{ and } Z^{\ell-1}_1=1\right] \\
&\quad + \sum_{i\ge 2, j\ge 2} \Pr\left[X^\ell_{0,i}=1\right]\Pr\left[X^\ell_{1,j}=1\right]\Pr\left[Z^{\ell-1}_0=1 \text{ and } Z^{\ell-1}_1=1\right] \\
&\le \frac{1}{2^u}p_{\ell-1} + \frac{(\eta-2)^2}{2^{2u}}p_{\ell-1} = p_{\ell-1}\left(\frac{1}{2^u}+\frac{(\eta-2)^2}{2^{2u}}\right)
\end{aligned}
$$

hence:

$$
p_\ell \le \left(\frac{1}{2^u}+\frac{(\eta-2)^2}{2^{2u}}\right)^\ell,
$$

and

$$
\Pr\left[Z^\ell_i=1\right] \le \frac{\eta-1}{2^u}\left(\frac{1}{2^u}+\frac{(\eta-2)^2}{2^{2u}}\right)^{k-1}
$$

And finally, the survival probability of at least one false positive after k iterations satisfies:

$$
\epsilon'_k \le \frac{\eta(\eta-1)}{2^u}\left(\frac{1}{2^u}+\frac{(\eta-2)^2}{2^{2u}}\right)^{k-1}
$$

For $(\eta=10^6, u=32, k=2)$, we get $\epsilon'_2 \le 0.013\%$.

How to select u? For a fixed k, ϵ'_k decreases as u grows. For a fixed u, ϵ'_k also decreases as k grows. Transmission, however, grows with both u (bigger digests) and k (more iterations). We write for the sake of clarity: $\epsilon'_k=\epsilon'_{k,u,\eta}$.

Fix η. Note that the number of bits transmitted per iteration ($\simeq 3ut$), is proportional to u. This yields an expected transmission complexity bound $T_{u,\eta}$ such that:

$$
T_{u,\eta} \propto u\sum_{k=1}^{\infty} k\cdot\epsilon'_{k,u,\eta} = \frac{u\eta\,(\eta-1)}{2^u}\sum_{k=1}^{\infty} k\left(\frac{1}{2^u}+\frac{(\eta-2)^2}{2^{2u}}\right)^{k-1} = \frac{u\eta\,(\eta-1)\,8^u}{\left(2^u-4^u+(\eta-2)^2\right)^2}
$$

Dropping the proportionality factor $\eta\,(\eta-1)$, neglecting $2^u \ll 2^{2u}$ and approximating $(\eta-2)\simeq\eta$, we can optimize the function:

$$
\phi_\eta(u) = \frac{u\cdot 8^u}{(4^u-\eta^2)^2}
$$

$\phi_{10^6}(u)$ admits an optimum for $u=19$.

Note: The previous analysis is incomplete because of the following approximations:

- We consider u-bit prime digests while u-bit strings contain only about $2^u/u$ primes.
- We used a fixed u in all rounds. Nothing forbids using a different u_k at each iteration, or even fine-tuning the u_ks adaptively as a function of the laundry's effect on the progressively reconciliated multisets..
- Our analysis treats t as a constant, but large t values increase p and hence the number of potential files detected as different per iteration - an effect disregarded *supra*.

A different approach is to optimize t and u experimentally, *e.g.* using the open source D&F program `btrsync` developed by the authors (*cf.* section 6).

3.3 How to Stop a Probabilistic Washing Machine?

We now combine both optimizations and assume that ℓ laundry rounds are necessary for completing some given reconciliation task using a half-sized p. By opposition to section 2.2, confirming correct protocol termination is now non-trivial.

Fabrice: oups, pas sûr de tout comprendre ici, il faudrait que l'on en rediscute... En particulier, le CRT n'a pas d'erreur a priori, sauf erreur de ma part...

~~La définition~~ de $\zeta(u)$ est assez gratuite vu que ça n'intervient pas dans la suite.

We say that a *round failure* occurs whenever a round results in an $(a', b') \neq (a, b)$ satisfying Equation (3). Let the round failure probability be some function $\zeta(u)$ (that we did not estimate). If u is kept small (for efficiency reasons), the probability $(1 - \zeta(u))^\ell$ that the protocol will properly terminate may dangerously drift away from one.

If v of $\ell + v$ rounds fail, Oscar needs to solve a problem called *Chinese Remaindering With Errors* [2]:

Problem 1. (Chinese Remaindering With Errors Problem: CRWEP*).* Given as input integers v, B and $\ell + v$ points $(s_1, p_1), \ldots, (s_{\ell+v}, p_{\ell+v}) \in \mathbb{N}^2$ where the p_i's are coprime, output all numbers $0 \leq s < B$ such that $s \equiv s_i \bmod p_i$ for at least ℓ values of i.

We refer the reader to [2] for more on this problem, which is beyond our scope. Boneh [3] provides a polynomial-time algorithm for solving the CRWEP under certain conditions satisfied in our setting.

But how can we confirm the solution? As mentioned in section 2.3, Neil will send to Oscar $H = \mathtt{Hash}(\mathfrak{F}')$ as the interaction starts. As long as Oscar's CRWEP resolution will not yield a state matching H, the parties will continue the interaction.

4 Computational Complexity

In this section, we are interested in computing the computational complexity of our protocol, when there is no collision, to simplify the analysis. We first present some variants of the protocol we described above, and then we analyse the complexity of all these variants and propose some algorithmic optimizations to speed up the file reconciliation. A summary of all costs can be found in Table 1.

4.1 Basic Complexity

Let $\mu(k)$ be the time required to multiply two k-bit numbers[5]. For naïve (*i.e.* convolutive) algorithms $\mu(k) = O(k^2)$, but using FFT multiplication [11], $\mu(k) = \tilde{O}(k)$. FFT is experimentally faster than convolutive methods starting at $k \sim 10^6$. The modular division of two k-bit numbers and the reduction of $2k$-bit number modulo a k-bit number are also known to cost $\tilde{O}(\mu(k))$ [4]. Indeed, in packages such as `gmp`, division and modular reduction run in $\tilde{O}(k)$.for sufficiently large k.

As proven in TODO APPENDIX, the complexity of `HashPrime` is $u^2\mu(u)$. Hence, we have the costs depicted in the third column of Table 1.

4.2 Adapting p_i

Using p_i ut_i-primes is not very practical, because generating a big prime number is slow, and storing a list of such primes require to be able to bound the number of rounds, and to fix all the parameters (u and $t_1, t_2, \dots$). That is why, in this section we show that we can adapt p_i to be able to generate them easily and also, to speed up the computations with a constant factor.

Let $\texttt{Prime}[i]$ denote the i-th prime[6]. Besides conditions on size, the *only* property required from p is to be co-prime with all the h_i and all the h'_i. We can hence consider the following variants:

Variant 1: Smooth p_i:

$$p_i = \prod_{j=r_i}^{r_{i+1}-1} \texttt{Prime}[j],$$

where the bounds r_i are chosen to ensure that each p_i has the proper size. Generating such a prime is much faster than generating a big prime p_i.

Variant 2: $p_i = \texttt{Prime}[i]^{r_i}$ where the exponents r_i are chosen to ensure that each p_i has the proper size. This variant is even faster than the previous one, but require to choose h_i bigger than all $\texttt{Prime}[i]^{r_i}$.

Variant 3: $p_i = 2^{ut_i}$. This variant, is probably the most efficient of all, but somewhat complex to explain. We hence describe it in detail in Appendix B. This variant is compatible with the use of FFT-multiplication, hence asymptotic complexity is preserved. In addition, it avoids all modular reductions and all CRT re-combinations and hence offers considerable constant-factor accelerations.

Actually it is easier to explain since there is no CRT, isn't it ???

[5] We assume that $\forall k, k', \mu(k+k') \geq \mu(k) + \mu(k')$.

[6] with $\texttt{Prime}[1] = 2$

```
Have this written by Fabrice...

(les modulos etant $2^{ut}, 2^{2ut}, ...$):
- calcul des $H_j$ comme dans 4.1 par product tree
- produit des $H_j$ de la facon suivante pour round 1:
  - $\mbox{prod} = 1$ (de taille $ut$)
  - pour $j=1,...$
      $\mbox{carry}_j | \mbox{prod} = \mbox{prod} \times H_j$
  - retourner prod, et garder carry
- produit des $H_j$ pour round suivants:
  - prod = 0
  - pour $j=1,...$
      $\mbox{carry}_j | \mbox{prod} = \mbox{prod} \times H_j + \mbox{carry}_j$

L'idee est la suivante (invariant): au round i et au tour j, prod contient les bits $(i-
$\mbox{carry}_j$ contient le carry genere par ces bits...
```

4.3 Algorithmic Optimizations using Product Trees

Indiquer clairement à quelles lignes du tableau ça se réfère, et si la complexité donnée dans le tableau prend ces optims en compte.

The non-overwhelming (but nonetheless important) complexities of the computations of (c, c') and of the factorizations can be even reduced to $\tilde{O}(\frac{n}{t}\mu(ut))$ using convolutive methods. These complexities can be reduced to $\tilde{O}(nu)$ with FFT [11]. To simplify the presentation, assume that $t = 2^\tau$ is a power of two dividing n.

The idea is the following: group h_i's by subsets of t elements and compute the product of each such subset in $\mathbb{N}$.

$$H_j = \prod_{i=jt}^{jt+t-1} h_i \in \mathbb{N}$$

Each H_j can be computed in $\tilde{O}(\mu(ut))$ using the standard product tree method described in Algorithm 1 (for $j = 0$) illustrated in Figure 5. Thus, all these $\frac{n}{t}$ products can be computed in $\tilde{O}(\frac{n}{t}\mu(ut))$. We can then compute c by multiplying the H_j modulo p, which costs $\tilde{O}(\frac{n}{t}\mu(ut))$.

What is the caveat?

The same technique applies to factorization[7], but with a slight *caveat*.

After computing the tree product, we can compute the residues of a modulo H_0. Then we can compute the residues of a mod H_0 modulo the two children π_2 and π_3 of $H_0 = \pi_1$ in the product tree (depicted in Figure 5), and so on. Intuitively, we descend the product tree doing modulo reduction. At the end (*i.e.*, as we reach the leaves), we obtain the residues of a modulo each of the h_i ($i \in \{0, \ldots, t-1\}$). This is described in Algorithm 6 and illustrated in Figure 6. We can use the same method for the tree product associated to any H_j, and

[7] We explain the process with a, this is applicable *ne variatur* to b as well.

Algorithm 1 Product Tree Algorithm

Require: the set h_i
Ensure: $\pi = \pi_1 = \prod_0^{t-1} h_i$, and π_i for $i \in \{1, \ldots, 2t-1\}$ as in Figure 5

$\pi \leftarrow$ array of size t
function $\textsc{prodTree}(i,\text{start},\text{end})$
  if start $=$ end **then**
    return 1
  else if start $+ 1 =$ end **then**
    return h_{start}
  else
    mid $\leftarrow \lfloor \frac{\text{start}+\text{end}}{2} \rfloor$
    $\pi_{2i} \leftarrow \textsc{prodTree}(2i,\text{start},\text{mid})$
    $\pi_{2i+1} \leftarrow \textsc{prodTree}(2i+1,\text{mid},\text{end})$
    return $\pi_{2i} \times \pi_{2i+1}$
$\pi_1 \leftarrow \textsc{prodTree}(1, 0, t)$

the residues of a modulo each of the h_i ($i \in \{jt, \ldots, jt+t-1\}$) for any j, *i.e.*, a modulo each of the h_i for any i. Complexity is $\tilde{O}(\mu(ut))$ for each j, which amounts to a total complexity of $\tilde{O}(\frac{n}{t}\tilde{O}(\mu(ut)))$.

Does it make sense to nest $\tilde{O}$'s?

Algorithm 2 Division Using a Product Tree

Require: $a \in \mathbb{N}$, π the product tree of Algorithm 1
Ensure: $A[i] = a \bmod \pi_i$ for $i \in \{1, \ldots, 2t-1\}$, computed as in Figure 2

$A \leftarrow$ array of size t
function $\textsc{modTree}(i)$
  if $i < 2t$ **then**
    $A[i] \leftarrow A[\lfloor i/2 \rfloor] \bmod \pi_i$
    $\textsc{modTree}(2i)$
    $\textsc{modTree}(2i+1)$
$A[1] \leftarrow a \bmod \pi_1$
$\textsc{modTree}(2)$
$\textsc{modTree}(3)$

4.4 Summary

Summary of costs is depicted in Table 1.

4.5 Remove Hashing to Prime Numbers

In most cases, the most costly operation is the hashing to prime number. That is why, a further optimization can consist in removing the need to hash to prime numbers by hashing to any integer coprime with all the p_i. The problem is that, even if there are no collision

Entity	Computation	Complexity expressed in $\tilde{O}$ of			
		Basic algo.		**Opt. algo.**	
		p_i prime	$p_i = 2^{ut_i}$	p_i prime	$p_i = 2^{ut_i}$
Both	computation of h_i and h'_i	$nu^2\mu(u)$			
	for round i				
Both	compute redundancies c_i and c'_i	$n \cdot \mu(ut_i)$		$\frac{n}{t_i} \cdot \mu(ut_i)$	
Neil	compute $s_i = c'_i/c_i$[a] or $S_i = C'_i/C_i$	$\mu(uT_i)$			
Neil	compute S_i from S_{i-1} and s_i (CRT)[a]	$\mu(uT_i)$	n/a	$\mu(uT_i)$	n/a
Neil	find a_i, b_i such that $S_i = a_i/b_i \bmod p_i$	$\mu(uT_i)^3$			
Neil	factor a_i	$n \cdot \mu(uT_i)$		$\frac{n}{T_i} \cdot \mu(uT_i)$	
	last round				
Oscar	factor b_i	$n \cdot \mu(ut_i)$		$\frac{n}{t_i} \cdot \mu(ut_i)$	
	Overwhelming complexity	TODO	TODO	TODO	TODO

[a] only for p_i prime or equivalent TODO;
[b] only for $p_i = 2^{ut_i}$.
[c] using advanced algorihms in [10,14] — naive extended GCD leads $(uT_i)^2$.

Table 1. Global Protocol Complexity

and a and b are correctly recovered, the fact that h_i divides a does not mean F_i should be a file in $\mathcal{S}$, because, for example, we can have $a = 150$, $h_1 = 10$, $h_2 = 15$, $h_3 = 6$, and $a = h_1 \times h_2$, but h_3 divides a. Therefore, we need a slightly more complex method for the factorization and a careful probability analysis to ensure this case does not come too often. This is left for future work.

Remarks Fabrice:

- we should separate move resolution algo from implementation
- we should implement the non-prime version stuff... but difficult to find the correct subset of hi which factorizes...

5 From Efficient Set Reconciliation to Efficient Files Synchronization

TODO: expliquer que l'on utilise d'abord le hashé du path+content, puis ensuite on fait le move reconciliation sur le hashé des content (avant d'envoyer tous les fichiers) et enfin, on fait une copie intelligente des fichiers restants (via rsync)...

5.1 TODO

TODO: rewrite this as explained in the previous TODO !!!!

1. Hash of files' contents concatenated with their paths, types (folder/file), and permissions (not supported yet).

2. Implement the protocol proposed in Section **??** with input data coming from `stdin` and output data going to `stdout`.

*encore une fois "Implement the protocol proposed in Section **??**" qu'a t on implemente au juste?*

More precisely:

– Oscar sends it product of hashes modulo a first prime number p_1.
– Neil receives the product, divides by its own product of hashes, reconstructs the fraction modulo p_1 and checks if he can factor the denominator using his hashes base. If he can, he stops and sends the numerator and the list of tuples (path, type, hash of content of the file) corresponding to the denominator's factors. Otherwise he sends "None" [is this the ASCII string "None"? if not what does he send precisely?].
– If Neil sent "None", Oscar computes the product of hashes modulo another prime p_2, sends it... CRT mechanism... [can we elaborate more on what happens here? which functions in GMP are used to do the CRT?]
– If Neil sent the numerator and a list of tuples, then Oscar factors the numerator over his own hash values. Now each party (Neil, Oscar) knows precisely the list of files (path + type + hash of content) that differs from the other party.

~~sur~~ "a first prime number p_1", l'implem marche avec p_i ou des 2^u?

[please structure the following:]

2. synchronize all the files. This part is not completely optimized.

We just remove all folders Oscar should not have and create new folders.

Then we remove all files Oscar should not have and synchronize using rsync the last files.

We could check for move (since we have the list of hash of contents of files) and do moves locally.

We can even try to detect moves of complete subtrees...

5.2 Move Resolution Algorithm

To reproduce the structure of Oscar on Neil's disk, we need to perform a sequence of file moves. Sadly, it is not straightforward to apply the moves, because, if we take a file to move, its destination might be blocked, either because a file already exists (we want to move a to b, but b already exists), or because a folder cannot be created (we want to move a to b/c, but b already exists as a file and not as a folder). Note that for a move operation $a \to b$, there is at most one file blocking the location b: we will call it the *blocker*.

If the blocker is absent on Oscar, then we can just delete the blocker. However, if a blocker exists, then we might need to move it somewhere else before we solve the move we are interested in. This move itself might have a blocker, and so on. It seems that we just need to continue until we reach a move which has no blocker or whose blocker can be deleted, but we can get caught in a cycle: if we must move a to b, b to c and c to a, then we will not be able to perform the operations without using a temporary location.

How can we perform the moves? A simple way would be to move each file to a unique temporary location and then rearrange files to our liking: however, this performs many unnecessary moves and could lead to problems if the program is interrupted. We can do something more clever by performing a decomposition in Strongly Connected Components (SCC) of the *move graph* (with one vertex per file and one edge per move operation going from to the file to its blocker or to its destination if no blocker exists). The computation of the SCC decomposition is simplified by the observation that because two files being moved to the same destination must be equal, we can only keep one arbitrary in-edge per node, and look at the graph pruned in this fashion: its nodes have in-degree at most one, so the strongly connected components are either single nodes or cycles. Once the SCC decomposition is known, the moves can be applied by applying each SCC in a bottom-up fashion, an SCC's moves being solved either trivially (for single files) or using one intermediate location (for cycles).

The detailed algorithm is implemented as two mutually recursive functions and presented as Algorithm 3.

6 Implementation

We implemented and benchmarked the Divide & Factor protocol described in the previous sections. The implementation is called `btrsync`, its source code is available from [1].

on parle de "in the previous sections" mais de quelle variate s'agit il?

Fabrice: in the paper, it is unidirectionnal too !!!!!!

TODO REMOVE THIS ! The program performs unidirectional synchronization, which is simpler to understand. The code is divided into two subprograms: a shell script and a Python program:

6.1 Organization

The program is composed of three parts: shell script, python program cmd, python program... TODO

The Shell Script sets up two instances of the Python program on Oscar and Neil and establishes a bidirectional communication channel between them using two Unix pipes between their standard inputs and outputs.

The Python Program uses `gmp` to perform all the number theory operations and performs the actual synchronization. It proceeds in two phases:

6.2 What is Implemented ?

- variant 2 of adapting p_i (power of small primes)

Algorithm 3 Perform Moves

Require: $\mathfrak{D}$ is a dictionary where $\mathfrak{D}[f]$ denotes the intended destinations of f

```
M ← []
T ← []
for f in 𝔇's keys do
    M[f] ← not_done
function UNBLOCK_COPY(f, t)
    if t is blocked by some b then
        if b is not in 𝔇's keys then
            unlink(b)                              ▷ We don't need b
        else
            RESOLVE(b)                             ▷ Take care of b and make it go away
    if T[f] was set then
        f ← T[f]
    copy(f, d)
function RESOLVE(f)
    if M[f] = done then
        return                                     ▷ Already managed by another in-edge
    if M[f] = doing then
        T[f] ← mktemp()
        move(f, T[f])
        M[f] ← done
        return                                     ▷ We found a loop, moved f out of the way
    M[f] ← doing
    for d ∈ 𝔇[f] do
        if d ≠ f then
            unblock_copy(f, d)                     ▷ Perform all the moves
    if f ∉ 𝔇[f] and T[f] was not set then
        unlink(f)
    if T[f] was set then
        unlink(T[f])
for f in 𝔇's keys do
    RESOLVE(f)
```

- hash to prime (remove use of hash to prime, not supported yet, though it WOULD be VERY interesting)
- $t_i = t' \times 2^i$ with t' a constant, i.e., doubling t at each round, for complexity reasons...

Small difference with theory:

- move reconciliation: An optimization implemented by `btrsync` over the algorithm described here is to move files instead of copying them and then remove the original file. Moves are faster than copies on most filesystems as the OS does not need to copy the actual file contents to perform moves.
- hash to prime: slightly different, I should correct this and use the fast version described in appendix if I have time (Fabrice)

6.3 Experimental Comparison to `rsync`

We compared `rsync`[8] and our Divide & Factor implementation (called `btrsync`) under the following experimental conditions:

Test Directories: The directories used for transmission and time comparisons are described in Table 3.

Command-Line Options: `rsync` was called with the following options, for the reasons below:

▶ `--delete` to delete existing files on Oscar which do not exist on Neil like `btrsync` does.

▶ `-I` to ensure that `rsync` did not cheat by looking at file modification times (which `btrsync` does not do).

▶ `--chmod="a=rx,u+w"` in an attempt to disable the transfer of file permissions (which `btrsync` does not transfer). Although these settings ensure that `rsync` does not need to transfer permissions, verbose logging suggests that it does transfer them anyway, so `rsync` must lose a few bytes per file as compared to `btrsync` for this reason.

▶ `-v` Transmission accounting was performed by calling `rsync` with the `-v` flag (which reports the number of sent and received bytes). For `btrsync` we added a piece of code counting the amount of data transmitted during `btrsync`'s own negotiations.

Network Configuration: Experiments were performed without any network transfer, by synchronizing two folders on the same host. Hence, time measurements should mostly represent the CPU cost of the synchronization.

[8] `rsync` version 3.0.9, used both as a competitor to benchmark against and as an underlying call in our own code.

Results: Results are given in Table 2. In general, `btrsync` spent more time than `rsync` on computation (especially when the number of files is large, which is typically seen in the experiments involving `synthetic`). Transmission results, however, turn out to be favorable to `btrsync`.

In the trivial experiments where either Oscar or Neil have no data at all, `rsync` outperforms `btrsync`. This is especially visible when Neil has no data: `rsync` immediately notices that there is nothing to transfer, but `btrsync` engages in information transfers to determine the symmetric difference.

On non-trivial tasks, however, `btrsync` outperforms `rsync`. This is the case of the `synthetic` datasets, where `btrsync` does not have to transfer information about all unmodified files, and even more so in the case where there are no modifications at all. For Firefox source code datasets, `btrsync` saves a very small amount of bandwidth, presumably because of unmodified files. For the `btrsync` source code dataset, we notice that `btrsync`, unlike `rsync`, was able to detect the move and avoid retransferring the moved folder.

7 Conclusion and Further Improvements

Vérifier les claims de cette liste, et en parler dès l'intro.

The main contributions of our work are:

- We present the novel "Divide & Factor" protocol for set reconciliation, which is based on number theory and is optimal with respect to transfer size.
- We study the problem of set reconciliation of directories of files. We discuss the optimal size of message digests in this setting, as well as a move resolution algorithm to reproduce a directory structure.
- We present `btrsync`, an open source implementation of the "Divide & Factor" protocol.
- We demonstrate the usability of this implementation through benchmarks on synthetic and real-world tasks, and show that `btrsync` exchanges less data than the popular software `rsync`.
- The optimizations presented in this paper apply to [] as well.

TODO résoudre ce point.

Be more specific!

Many fine questions of the probabilistic discussions in the paper are left as future work. Another further line of research would be to pursue development of `btrsync` to make it suitable for end users.

8 Acknowledgment

The authors acknowledge Guillain Potron for his early involvement in this research work.

9 ToDo

- @Fabrice: Pour éviter le cas empty $\rightarrow$ source trop gros, on pourrait imaginer l'astuce suivante: si jamais Neil la taille de c est plus petite que la taille du produit des nombres premiers $p_1 ... p_n$ utilisés, Neil envoie un message pour l'indiquer, et on arrête là le protocole. Et Oscar peut directement factoriser ce nombre envoyé...
- @Fabrice: il faut discuter de la taille de la taille maximale des "petits" premiers utilisés pour les variantes de p et montrer que cela n'enlève pas trop d'entropie pour les h_i. Encore une fois, je m'en occupe la semaine prochaine si besoin.
- Refaire une dernière fois les expériences, vu que Fabrice a significativement amélioré les perfs.
- Faire clarifier par Fabrice l'histoire du doublement...
- Fabrice: comparer nos perfs avec http://ipsit.bu.edu/programs/reconcile/ ? ou pas ?

References

1. Amarilli, A., Ben Hamouda, F., Bourse, F., Morisset, R., Naccache, D., Rauzy, P.: `https://github.com/RobinMorisset/Btrsync`
2. Bleichenbacher, D., Nguyen, P.Q.: Noisy polynomial interpolation and noisy chinese remaindering. In: Advances in Cryptology – Proceedings of Eurocrypt 2000. pp. 53–69. Springer-Verlag (2000)
3. Boneh, D.: Finding smooth integers in short intervals using crt decoding. In: Proceedings of the 32nd Annual ACM Symposium on Theory of Computing. pp. 265–272 (2000)
4. Burnikel, C., Ziegler, J., Stadtwald, I., D-Saarbrucken: Fast recursive division (1998)
5. Eppstein, D., Goodrich, M., Uyeda, F., Varghese, G.: What's the difference?: efficient set reconciliation without prior context. In: ACM SIGCOMM Computer Communication Review. vol. 41, pp. 218–229. ACM (2011)
6. Fouque, P.A., Stern, J., Wackers, J.G.: Cryptocomputing with rationals. In: Blaze, M. (ed.) Financial Cryptography. Lecture Notes in Computer Science, vol. 2357, pp. 136–146. Springer (2002)
7. Hofmann, M., Pierce, B.C., Wagner, D.: Edit lenses. In: Field, J., Hicks, M. (eds.) POPL. pp. 495–508. ACM (2012)
8. Minsky, Y., Trachtenberg, A.: Scalable set reconciliation. In: 40th Annual Allerton Conference on Communications, Control and Computing (October 2002), a full version can be downloaded from `http://ipsit.bu.edu/documents/BUTR2002-01.ps`
9. Minsky, Y., Trachtenberg, A., Zippel, R.: Set reconciliation with nearly optimal communication complexity. Information Theory, IEEE Transactions on 49(9), 2213–2218 (2003)
10. Pan, V., Wang, X.: On rational number reconstruction and approximation. SIAM Journal on Computing 33, 502 (2004)
11. Schönhage, A., Strassen, V.: Schnelle multiplikation grosser zahlen. Computing 7(3), 281–292 (1971)
12. Tridgell, A.: Efficient algorithms for sorting and synchronization. Ph.D. thesis, PhD thesis, The Australian National University (1999)
13. Vallée, B.: Gauss' algorithm revisited. J. Algorithms 12(4), 556–572 (1991)
14. Wang, X., Pan, V.: Acceleration of Euclidean algorithm and rational number reconstruction. SIAM Journal on Computing 32(2), 548 (2003)

A Extended Protocol

Oscar		Neil
		First phase during which Neil amasses modular information on the difference
		start the protocol with p_1
	$\xrightarrow{c_1}$	
		computes a, b using p_1
		if a factors properly then go to Final Phase
		else perform the protocol with p_2
	$\xrightarrow{c_2}$	
		computes $c \bmod p_1p_2 = \mathrm{CRT}_{p_1,p_2}(c_1, c_2)$
		computes a, b using p_1p_2
		if a factors properly then go to Final Phase
		else perform the protocol with p_3
	$\xrightarrow{c_3}$	
		computes $c \bmod p_1p_2p_3 = \mathrm{CRT}_{p_1,p_2,p_3}(c_1, c_2, c_3)$
		computes a, b using $p_1p_2p_3$
		if a factors properly then go to Final Phase
		else perform the protocol with p_4
	$\vdots$	
	Final Phase	
		Let $\mathfrak{S} = \{F'_i \text{ s.t. } a \bmod h'_i = 0\}$
	$\xleftarrow{\mathfrak{S}, b}$	
deletes files s.t. $b \bmod h_i = 0$		
adds $\mathfrak{S}$ to the disk		

Note that parties do not need to store the p_i's in full. Indeed, the p_is could be subsequent primes sharing their most significant bits. This reduces storage per prime to a very small additive constant $\cong \ln(p_i) \cong \ln(2^{2tu+2}) \cong 1.39(tu + 1)$ of about $\log_2(tu)$ bits.

B Power of Two Protocol

In this variant Oscar computes c in $\mathbb{N}$:

$$c = \prod_{F_i \in \mathfrak{F}} \texttt{HashPrime}(F_i) = \prod_{i=1}^{n} h_i \in \mathbb{N}$$

and considers $c = \bar{c}_{n-1}|\dots|\bar{c}_2|\bar{c}_0$ as the concatenation of n successive u-bit strings. Again, we omit the treatment of $\perp$s for the sake of clarity.

First phase during which Neil amasses modular information on the difference		
Oscar		**Neil**
computes $c \in \mathbb{N}$		
	$\xrightarrow{\bar{c}_0}$	
		computes a, b modulo 2^u
		if a factors properly then go to Final Phase
		else request next chunk $\bar{c}_1$
	$\xrightarrow{\bar{c}_1}$	
		construct $c \bmod 2^{2u} = \bar{c}_1\|\bar{c}_0$
		computes a, b modulo 2^{2u}
		if a factors properly then go to Final Phase
		else request next chunk $\bar{c}_2$
	$\xrightarrow{\bar{c}_2}$	
		construct $c \bmod 2^{3u} = \bar{c}_2\|\bar{c}_1\|\bar{c}_0$
		computes a, b modulo 2^{3u}
		if a factors properly then go to Final Phase
		else request next chunk $\bar{c}_3$
	$\vdots$ (for $2t$ rounds)	

Final Phase		
		Let $\mathfrak{S} = \{F'_i \text{ s.t. } a \bmod 2^{2tu} = 0\}$
	$\xleftarrow{\mathfrak{S}, b}$	
deletes files s.t. $b \bmod 2^{2tu} = 0$		
adds $\mathfrak{S}$ to the disk		

C Hashing Into Primes

analyse ci-contre

Hashing into primes is frequently needed in cryptography. A recommended implementation of `HashPrime`(F) is given in Algorithm 5. If u is large enough (*e.g.* 160) one might sacrifice uniformity to avoid repeated file hashings by defining `HashPrime`(F) = `NextPrime`(`Hash`(F)). Yet another acceleration (that further destroys uniformity) consists in replacing `NextPrime` by Algorithm 4 where $\alpha = 2 \times 3 \times 5 \times \cdots \times \texttt{Prime}[d]$ is the product of the first primes until some rank d.

Fabrice: Cela accélère un peu, car il y a environ $\frac{n}{\log(n)\varphi(\alpha)}$ nombres premiers $\leq n$ et congrus à 1 modulo α, contre $\frac{n}{\log(n)}$ nombres premiers $\leq n$ (voir `http://fr.wikipedia.org/wiki/Th%C3%A9or%C3%A8me_de_la_progression_arithm%C3%A9tique#Version_quantitative`). Dans l'algo 4, h est donc premier avec proba $\frac{\frac{n}{\log(n)\varphi(a)}}{\frac{n}{\alpha}} = \frac{\alpha}{\log(n)\varphi(\alpha)}$, tandis que dans l'algo 5, h est premier avec proba $\frac{1}{\log(n)}$. On a alors une accélération d'environ 10 si on prend $d = 60$.

Algorithm 4 Fast Nonuniform Hashing Into Primes

$h = \alpha \left\lfloor \frac{\texttt{Hash}(F)}{\alpha} \right\rfloor + 1$
while h is composite **do**
  $h = h - \alpha$
return h

Algorithm 5 Possible Implementation of `HashPrime`(F)

$i = 0$
repeat
  $h = 2 \cdot \texttt{Hash}(F|i) + 1$
  $i = i + 1$
until h is prime
return h

Antoine sur quelle plate-formes ont ete obtenus les resultats experimentaux? quelles vitesses de processeur etc?

Entities and Datasets		Transmission (Bytes)						Time (s)	
Neil's $\mathfrak{F}'$	**Oscar's $\mathfrak{F}$**	$\mathtt{TX}_{\mathtt{rs}}$	$\mathtt{RX}_{\mathtt{rs}}$	$\mathtt{TX}_{\mathtt{bt}}$	$\mathtt{RX}_{\mathtt{bt}}$	$\delta_{\mathtt{rs}} - \delta_{\mathtt{bt}}$	$\frac{\delta_{\mathtt{bt}}}{\delta_{\mathtt{rs}}}$	$\mathtt{time}_{\mathtt{rs}}$	$\mathtt{time}_{\mathtt{bt}}$
`source`	`empty`	778311	1614	779517	10140	9732	1.0	0.1	0.4
`empty`	`source`	24	12	11927	5952	17843	496.6	0.1	0.4
`empty`	`empty`	24	12	19	30	13	1.4	0.0	0.3
`synthetic`	`synthetic_shuffled`	54799	19012	7308	3417	-63086	0.1	0.2	1.5
`synthetic_shuffled`	`synthetic`	54407	18822	6822	3042	-63365	0.1	0.2	0.8
`synthetic`	`synthetic`	54799	19012	327	30	-73454	0.0	0.1	0.7
`firefox-13.0.1`	`firefox-13.0`	40998350	1187	39604079	3305	-1392153	1.0	1.5	10.2
`source_moved`	`source`	778176	1473	2757	1966	-774926	0.0	0.1	0.6

Table 2. Experimental results. `rs` and `bt` subscripts respectively denote `rsync` and `btrsync`. The two first columns indicate the datasets. Synchronization is performed *from* Neil *to* Oscar. `RX` and `TX` denote the quantity of received and sent bytes and $\delta_{\square} = \mathtt{TX}_{\square} + \mathtt{RX}_{\square}$. $\delta_{\mathtt{rs}} - \delta_{\mathtt{bt}}$ and $\delta_{\mathtt{bt}}/\delta_{\mathtt{rs}}$ express the absolute and the relative differences in transmission between the two programs. The last two columns show timing results.

Directory	Description
`synthetic`	A directory containing 1000 very small files containing the numbers $1, 2, \ldots, 1000$.
`synthetic_shuffled`	`synthetic` with: 10 deleted files 10 renamed files 10 modified files
`source`	A snapshot of `btrsync`'s own source tree
`source_moved`	`source` with one big folder (a few megabits) renamed.
`firefox-13.0`	The source archive of Mozilla Firefox 13.0.
`firefox-13.0.1`	The source archive of Mozilla Firefox 13.0.1
`empty`	An empty folder.

Table 3. Test Directories.

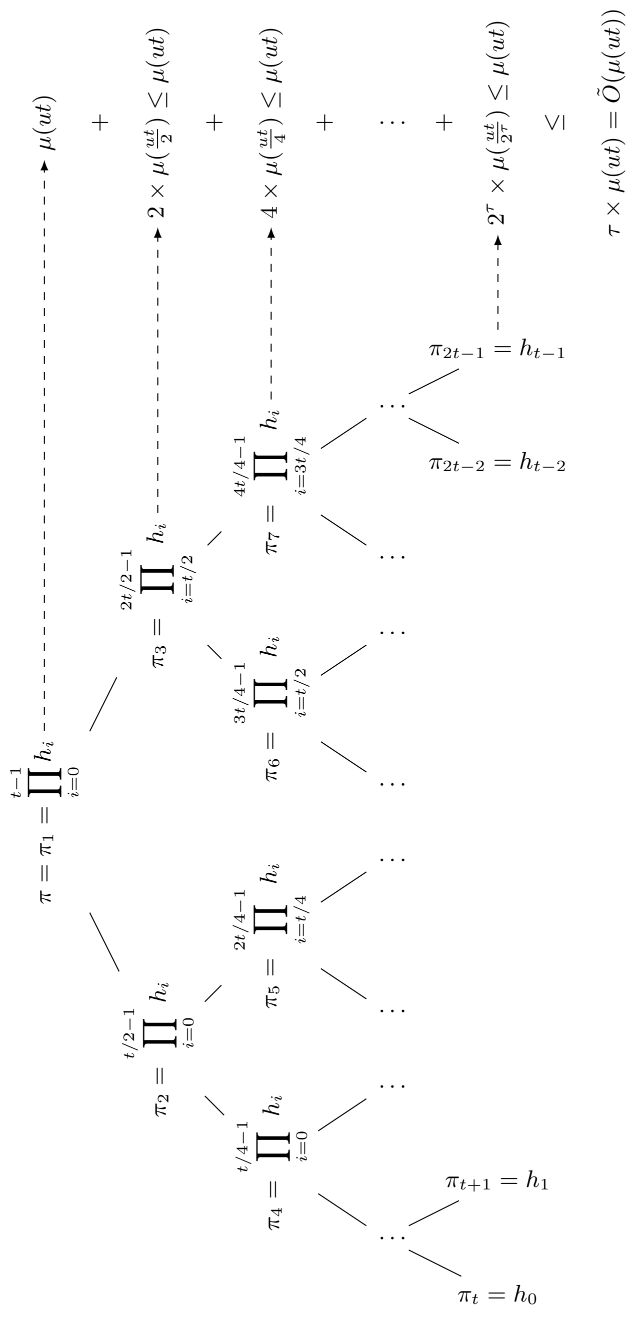

Fig. 5. Product Tree

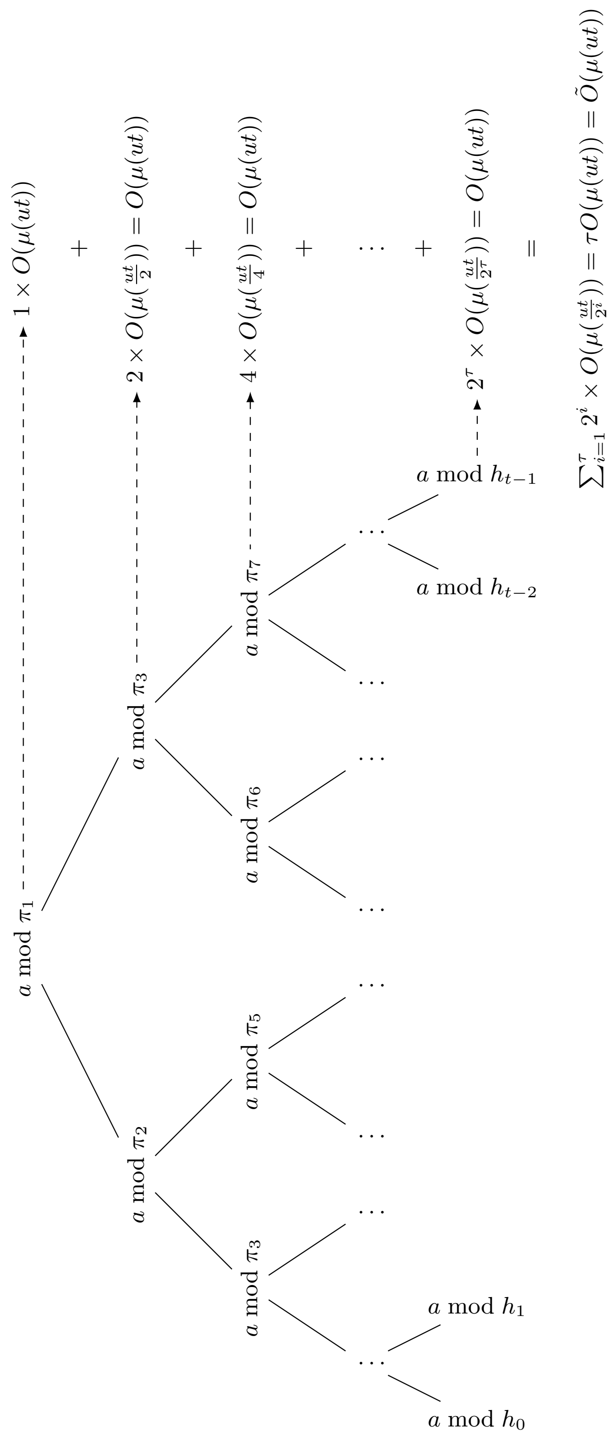

Fig. 6. Modular Reduction From Product Tree