

# Bleach: A Distributed Stream Data Cleaning System

Yongchao Tian  
Eurecom  
yongchao.tian@eurecom.fr

Pietro Michiardi  
Eurecom  
pietro.michiardi@eurecom.fr

Marko Vukolić  
IBM Research - Zurich  
mvu@zurich.ibm.com

## ABSTRACT

In this paper we address the problem of rule-based stream data cleaning, which sets stringent requirements on latency, rule dynamics and ability to cope with the unbounded nature of data streams.

We design a system, called Bleach, which achieves real-time violation detection and data repair on a dirty data stream. Bleach relies on efficient, compact and distributed data structures to maintain the necessary state to repair data, using an incremental version of the equivalence class algorithm. Additionally, it supports rule dynamics and uses a “cumulative” sliding window operation to improve cleaning accuracy.

We evaluate a prototype of Bleach using a TPC-DS derived dirty data stream and observe its high throughput, low latency and high cleaning accuracy, even with rule dynamics. Experimental results indicate superior performance of Bleach compared to a baseline system built on the micro-batch streaming paradigm.

## 1. INTRODUCTION

Today, we live in a world where decisions are often based on analytics applications that process continuous streams of data. Typically, data streams are combined and summarized to obtain a succinct representation thereof: analytics applications rely on such representations to make predictions, and to create reports, dashboards and visualizations [15, 23, 26]. All these applications expect the data, and their representation, to meet certain quality criteria. Data quality issues interfere with these representations and distort the data, leading to misleading analysis outcomes and potentially bad decisions.

As such, a range of data cleaning techniques were proposed recently [28, 22, 19]. However, most of them focus on “batch” data cleaning, by processing static data stored in data warehouses, thus neglecting the important class of streaming data. In this paper, we address this gap and focus on *stream data cleaning*. The challenge in stream cleaning

is that it requires both *real-time* guarantees as well as high *accuracy*, requirements that are often at odds.

A naïve approach to stream data cleaning could simply extend existing batch techniques, by buffering data records in a temporary data store and cleaning it periodically before feeding it into downstream components. Although likely to achieve high accuracy, such a method clearly violates real-time requirements of streaming applications. The problem is exacerbated by the volume of data cleaning systems need to process, which prohibits centralized solutions. Therefore, our goal is to design a *distributed stream data cleaning system*, which achieves efficient and accurate cleaning in real-time.

In this paper, we focus on rule-based data cleaning, whereby a set of domain-specific rules define how data should be cleaned: in particular, we consider functional dependencies (FDs) and conditional functional dependencies (CFDs). Our system, called Bleach, proceeds in two phases: *violation detection*, to find rule violations, and *violation repair*, to repair data based on such violations. Bleach relies on efficient, compact and distributed data structures to maintain the necessary state (e.g., summaries of past data) to repair data, using an incremental equivalence class algorithm.

We further address the complications due to the long-term and dynamic nature of data streams: the definition of dirty data could change to follow such dynamics. Bleach supports dynamic rules, which can be added and deleted without requiring idle time. Additionally, Bleach implements a sliding window operation that trades modest additional storage requirements to temporarily store cumulative statistics, for increasing cleaning accuracy.

Our experimental performance evaluation of Bleach is two-fold. First, we study the performance, in terms of throughput, latency and accuracy, of our prototype and focus on the impact of its parameters. Then we compare Bleach to an alternative baseline system, which we implement using a micro-batch streaming architecture. Our results indicate the benefits of a system like Bleach, which hold even with rule dynamics. Despite extensive work on rule-based data cleaning [1, 6, 9, 10, 13, 20, 7, 19], we are not aware of any other stream data cleaning system.

The paper is organized as follows. Section 2 introduces our problem statement. The system design of Bleach is discussed in Section 3; dynamic rule management and windowing are discussed respectively in Section 4 and Section 5. Section 6 presents our experimental results. Section 7 overviews related work. Finally, Section 8 concludes our work.

## 2. PRELIMINARIES

Next, we introduce some basic notations we use throughout the paper, then we define the problem statement we consider in this work.

### 2.1 Background and Definitions

Similar to data cleaning systems in data warehouses, that read a dirty dataset and write back a cleaned dataset, in this paper we assume that a stream data cleaning system ingests a data stream and outputs a cleaned data stream instance. We consider an input data stream instance  $D_{in}$  with schema  $S(A_1, A_2, \dots, A_m)$  where  $A_j$  is an attribute in schema  $S$ . We assume the existence of unique tuple identifiers for every tuple in  $D_{in}$ : thus given a tuple  $t_i$ ,  $id(t_i)$  is  $t_i$ 's identifier. In general we define a function  $id(e)$  which returns the identifier (ID) of  $e$  where  $e$  can be any element. A list of IDs  $[id(e_1), id(e_2), \dots, id(e_n)]$  is expressed as  $id(e_1, e_2, \dots, e_n)$  for brevity. The output data stream instance  $D_{out}$  complies to schema  $S$  and has the same tuple identifiers as in  $D_{in}$ , without any loss or duplication. The basic unit, a cell  $c_{i,j}$ , is the concatenation of a tuple id, an attribute and the projection of the tuple on the attribute:  $c_{i,j} = (id(t_i), A_j, t_i(A_j))$ . Note that  $t_i(A_j)$  is the value of  $c_{i,j}$ , which can also be expressed as  $v(c_{i,j})$ . Sometimes, we may simply express  $c_{i,j}$  as  $c_i$  when the cell attribute is not relevant to the discussion. In our work, when we point at a specific tuple  $t_i$ , we also refer to this tuple as the *current* tuple. Tuples appearing earlier than  $t_i$  in the data stream are referred to as *earlier* tuples and those appearing after  $t_i$  are referred to as *later* tuples.

To perform data cleaning, we define a set of rules  $\Sigma = [r_1, \dots, r_n]$ , in which  $r_k$  is either a functional dependency (FD) rule or a conditional FD rule (CFD). Each rule has a unique rule identifier  $id(r_k)$ . A CFD rule  $r_k$  is represented by  $(X \rightarrow A, cond(Y))$ , in which  $cond(Y)$  is a boolean function on a set of attributes  $Y$  where  $Y \subseteq S$ .  $X$  and  $A$  are respectively referred to as a set of left-hand side (LHS) attributes and right-hand side (RHS) attribute:  $LHS(r_k) = X$ ,  $RHS(r_k) = A$ . When the rule is clear in the context, we omit  $r_k$  so that  $LHS = X$ ,  $RHS = A$ . Cells of LHS (RHS) attributes are also referred to as LHS (RHS) cells.  $Y$  is referred to as a set of conditional attributes. For a pair of tuples  $t_1$  and  $t_2$  satisfying condition  $cond(t_1(Y)) = cond(t_2(Y)) = true$ , if  $t_1(B) = t_2(B)$  for all  $B \in X$  but  $t_1(A) \neq t_2(A)$ , then it is a *violation* for  $r_k$ . A data stream instance  $D$  satisfies  $r_k$ , denoted as  $D \models r_k$ , when there are no violations for  $r_k$  exist in  $D$ . A FD rule can be seen as a special case of CFD rule where  $cond(Y)$  is always true and  $Y$  is  $\emptyset$ . We refer to an attribute as an *intersecting* attribute if it is involved in multiple rules.

If  $D$  satisfies a set of rules  $\Sigma$ , denoted  $D \models \Sigma$ , then  $D \models r_k$  for  $\forall r_k \in \Sigma$ . If  $D$  does not satisfy  $\Sigma$ ,  $D$  is a dirty data stream instance.

### 2.2 Challenges and Goals

An ideal stream data cleaning system should accept a dirty input stream  $D_{in}$  and output a clean stream  $D_{out}$ , in which all *rule violations* in  $D_{in}$  are repaired ( $D_{out} \models \Sigma$ ). However, this is not possible in reality due to:

- **Real-time constraint:** As the data cleaning is incremental, the cleaning decision for a tuple (repair or not repair) can only be made based on itself and earlier

time	item	category	clientid	city	zipcode
...	...	...	...	...	...
$t_1$	MacBook	computer	11111	France	75001
$t_2$	bike	sports	33333	Lyon	null
$t_3$	Interstellar	movies	22222	Paris	75001
$t_4$	bike	toys	44444	Nice	06000
$t_5$	Titanic	movies	11111	Paris	null
...	...	...	...	...	...

Figure 1: Illustrative example of a data stream consisting of on-line transactions.

tuples in the data stream, which is different from data cleaning in data warehouses where the entire dataset is available. In other words, if a dirty tuple only has violations with later tuples in the data stream, it can not be cleaned. A late update for a tuple in the output data stream is not accepted.

- **Dynamic rules:** In a stream data cleaning system, the rule set is not static. A new rule may be added or an obsolete rule may be deleted at any time. A processed data tuple can not be cleaned again with an updated rule set. Reprocessing the whole data stream whenever the rule set is updated is not realistic.
- **Unbounded data:** A data stream produces an unbounded amount of data, that cannot be stored completely. Thus, stream data cleaning can not afford to perform cleaning on the full data history. Namely, if a dirty tuple only has violations with tuples that appear much earlier in the data stream, it is likely that it will not be cleaned.

Consider the example in Figure 1, which is a data stream of on-line shopping transactions. Each tuple represents a purchase record, which contains a purchased item (*item*), the category of that item (*category*), a client identifier (*clientid*), the city of the client (*city*) and the zip code of that city (*zipcode*). In the example, we show an extract of five data tuples of the data stream, from  $t_1$  to  $t_5$ .

Now, assume we are given two FD rules and one CFD rule stating how a clean data stream should look like: ( $r_1$ ) the same items can only belong to the same category; ( $r_2$ ) two records with the same *clientid* must have the same city; ( $r_3$ ) two records with the same non-null zip code must have the same city:

$$(r_1) \text{ item} \rightarrow \text{category}$$

$$(r_2) \text{ clientid} \rightarrow \text{city}$$

$$(r_3) \text{ zipcode} \rightarrow \text{city}, \text{zipcode} \neq \text{null}$$

If we focus on the detection of tuples that *violate* such rules, we recognize three violations among the five tuples: ( $v_1$ )  $t_1$  and  $t_3$  have the same non-null zip code ( $t_1(\text{zipcode}) = t_3(\text{zipcode}) \neq \text{null}$ ) but different city names ( $t_1(\text{city}) \neq t_3(\text{city})$ ); ( $v_2$ )  $t_2$  claims bikes belong to category sports while  $t_4$  classifies bikes as toys ( $t_2(\text{item}) = t_4(\text{item}), t_2(\text{category}) \neq t_4(\text{category})$ ); and ( $v_3$ )  $t_1$  and  $t_5$  have the same *clientid* but

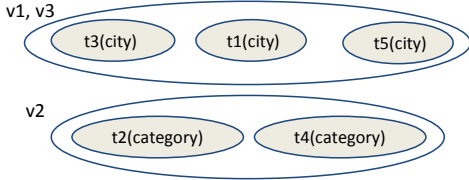


Figure 2: An example of a violation graph, derived from our running example.

different city names ( $t_1(clientid) = t_5(clientid), t_1(city) \neq t_5(city)$ ).

Note that when a stream data cleaning system receives tuple  $t_1$ , no violation can be detected as in our example  $t_1$  only has violations with later tuples  $t_3$  and  $t_5$ . Thus, no modification can be made to  $t_1$ . Furthermore, delaying the cleaning process for  $t_1$  is not an option, not only because of real-time constraints, but also because it is difficult to predict for how long this tuple should be buffered for it to be cleaned.

Although performing incremental violation detection seems straightforward, incremental violation repair is much more complex to achieve. Coming back to the example in Figure 1, assume that the stream cleaning system receives tuple  $t_5$  and successfully detects the violation  $v_3$  between  $t_5$  and  $t_1$ . Such detection is not sufficient to make the correct repair decision, as the tuple  $t_1$  also conflicts with another tuple,  $t_3$ . An incremental repair in stream data cleaning system should also take the violations among earlier tuples into account.

To account for the intricacies of the violation repair process, we introduce the concept of *violation graph* [19]. A violation graph is a data structure containing the detected violations, in which each node represents a cell. If some violations share a common cell, they will be grouped into a single *subgraph* (sg). Therefore, the violation graph is partitioned into smaller independent subgraphs. A single cell can only be in one subgraph. If two subgraphs share a common cell, they need to merge.

The repair decision of a tuple is only relevant to the subgraphs in which its cells are involved. A violation graph for our example can be seen in Figure 2. Given this violation graph, to make the repair decision for tuple  $t_5$ , the cleaning system can only rely on the upper subgraph which consists of violation  $v_1$  and  $v_3$  with the common cell  $t_1(city)$ .

We now give our problem statement as following.

**Problem statement:** Given an unbounded data stream with an associated schema<sup>1</sup> and a *dynamic* set of rules, how can we design an *incremental* and *real-time* data cleaning system, including violation detection and violation repair mechanisms, using bounded computing and storage resources, to output a cleaned data stream? In the next three sections, we give a detailed description of our distributed stream data cleaning system, that we call Bleach.

### 3. BLEACH DESIGN AND ALGORITHMS

In this section, we overview the Bleach architecture and provide details about its components. As shown in Figure 3,

<sup>1</sup>Note that although we restrict the data stream to have a fixed schema in this work, it is easy to extend our work to support a dynamic schema.

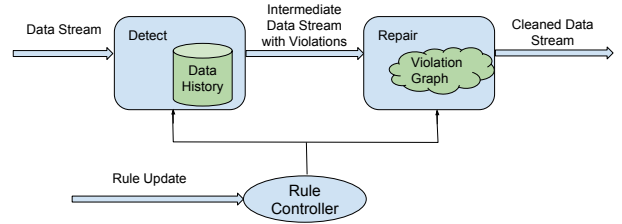


Figure 3: Stream data cleaning Overview

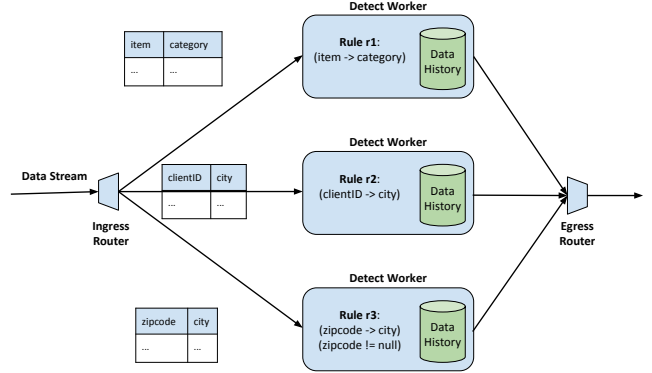


Figure 4: The Detect Module

Bleach consists in two main blocks, namely the *detect* and *repair* modules, and a rule controller module, which is discussed in Section 4.

The input data stream first enters the detect module, which reveals violations against defined rules. The intermediate data stream, output from the first module, is enriched with violation information, which the repair module uses to make repair decisions. Finally, the system outputs a cleaned data stream.

Next, we delve into the details of the first two modules and outline several optimizations that aim at achieving efficiency and performance.

#### 3.1 Violation Detection

The violation detection module aims at finding input tuples that violate rules. To do so, it stores them in an in-memory, efficient and compact data structure that we call the *data history*. Input tuples are thus compared to those in the data history to detect violations.

Figure 4 illustrates the internals of the detect module: it consists of an ingress router, an egress router and multiple detect workers (DW). Bleach maps violation rules to such DW: each worker is in charge of finding violations for a specific rule.

##### 3.1.1 The Ingress Router

The goal of the ingress router is to partition and distribute incoming tuples to DWs. Now, as discussed in Section 2, only a subset of the attributes of an input tuple are relevant when verifying data validity against a given rule. For example, a FD rule only requires its LHS and RHS attributes to be verified, ignoring the rest of the input tuple attributes.

Therefore, when the ingress router receives an input tuple, it partitions the tuple based on the current rule set, and only sends the relevant information to each DW in charge

of each specific rule. As such, an input tuple is broken into multiple sub-tuples, which all share the same identifier of the corresponding input tuple. Note that some attributes of an input tuple might be required by multiple rules: in this case, sub-tuples will contain redundant information, allowing each DW to work independently.

An example of tuple partitioning can be found in Figure 4, where we reuse the input data schema and the rules from Section 2.

### 3.1.2 The Detect Worker

Each DW is assigned a rule, and receives the relevant sub-tuples stemming from the input stream. For each sub-tuple, a DW needs to perform a lookup operation in the data history, and eventually emit a message (that is part of an intermediate data stream) to downstream components when a rule violation is detected.

To achieve efficiency and performance, lookup operations on the data history need to be fast, and the intermediate data stream should avoid redundant information. Next, we first describe how the data history is represented and materializes in memory; then, we describe the output messages a DW generates, and finally outline the DW algorithm.

**Data history representation.** A DW accumulates relevant input sub-tuples in a compact data structure that enables an efficient lookup process, which makes it similar to a traditional indexing mechanism.

The structure<sup>2</sup> of the data history is illustrated in Figure 5. First, to speed-up the lookup process, sub-tuples are grouped by the value of the LHS attribute used by a given rule: we call such group a *cell group* (*cg*). Thus, A *cg* stores all RHS cells whose sub-tuples share the same LHS value. The identifier of a cell group  $cg_l$  is the combination of the rule assigned to the DW, and the value of LHS attributes, expressed as  $id(cg_l) = (id(r_k), t(LHS))$  where  $r_k$  is the rule assigned to the DW.

Next, to achieve a compact data representation, all cells in a *cg* sharing the same RHS value are grouped into a *super cell* (*sc*):  $sc_m = [c_{1,j}, c_{2,j}, \dots, c_{n,j}]$ . From Section 2, recall that a cell is made of a tuple ID, an attribute and a value:  $(id(t_i), A_j, t_i(A_j))$ . Therefore, a super cell can be *compressed* as a list of tuple IDs, an attribute and their common value:  $sc_m = (id(t_1, t_2, \dots, t_n), A_j, t(A_j))$  where  $t(A_j) = t_1(A_j) = \dots = t_n(A_j)$ . Hence, within an individual DW, sub-tuples whose cells are compressed in the same *sc* are equivalent, as they have the same LHS attributes value (the identity of the cell group) and the same RHS attribute value (the value of super cell). A cell group  $cg_l$  now can be expressed as:  $cg_l = ((id(r_k), t(LHS)), [sc_1, sc_2, \dots])$  including an identifier and a list of super cells.

In summary, the lookup process for a given input sub-tuple is as follows. Cell groups are stored in a hash-map using their identifier as keys: therefore the DW first finds the *cg* corresponding to the current sub-tuple. Cells in the corresponding *cg* are the only cells that might be in conflict with the current cell. Overall, the complexity of the lookup process for a sub-tuple is  $O(1)$ .

**Violation messages.** DWs generate an intermediate data stream of *violation messages*, which help downstream components to eventually repair input tuples. The goal of the

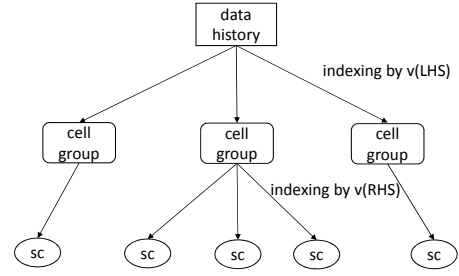


Figure 5: The structure of the data history in a detect worker

DW is to generate as few messages as possible, while allowing effective data repair.

When the lookup process reveals the current tuple does not violate a rule, DWs emit a non-violation message ( $msg_{nvio}$ ). Instead, when a violation is detected, a DW constructs a message with all the necessary information to repair it, including: the ID of the cell group corresponding to the current tuple and the RHS cells of the current and earlier tuples in data history:  $msg_{vio} = (id(cg_l), c_{cur}, c_{old})$ .

Now, to reduce the number of violation messages, the DW can use a super cell in place of a single cell ( $c_{old}$ ) in conflict with the current tuple. In addition, recall that a single *cg* can contain multiple super cells, thus possibly requiring multiple messages for each group. However, we observe that two cells in the same *cg* must also conflict with each other, as long as their values are different. Since the data repair module in Bleach is stateful, it is safe to omit multiple violation messages for such cells.

**Algorithm details.** Next, we present the DW violation algorithm details, as illustrated in Algorithm 1.

---

#### Algorithm 1 Violation Detection

---

```

1: given rule  $r = (X \rightarrow A_j, cond(Y))$ 
2: procedure RECEIVE(sub-tuple  $t_i$ ) ▷
    $cond(t_i(Y)) = true$ 
3:   if  $\exists id(cg_l) = (id(r), t_i(X))$  then
4:     if  $|cg_l| = 1$  then ▷ cg_l contains  $sc_{old}$ 
5:       if  $v(sc_{old}) = t_i(A_j)$  then
6:         Emit  $msg_{nvio}$ 
7:       else
8:         Emit  $msg_{vio}(id(cg_l), c_{cur}, sc_{old})$ 
9:       end if
10:    else
11:      Emit  $msg_{vio}(id(cg_l), c_{cur}, null)$ 
12:    end if
13:  else
14:    Create  $cg_l$  ▷ Create a new cell group
15:    Emit  $msg_{nvio}$ 
16:  end if
17:  Add  $c_{cur}$  to  $cg_l$ 
18: end procedure

```

---

The algorithm starts by treating FD rules as a special case of CFD rules (line 1).

Then, when a DW receives a sub-tuple  $t_i$  satisfying the rule condition (line 2), it performs a lookup in the data history to check if the corresponding cell group  $cg_l$  exists

<sup>2</sup>The techniques we use are similar to the notion of *partitions* and *compression* introduced in Nadeef [10].

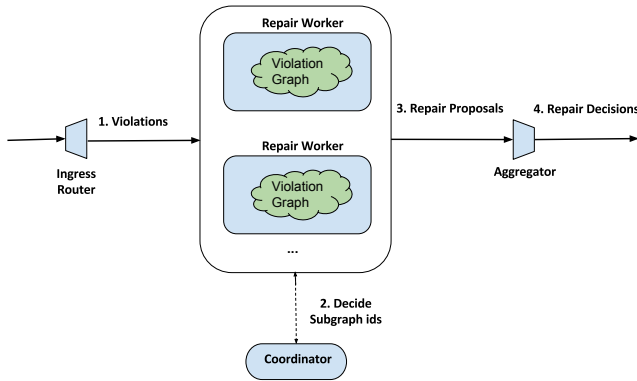


Figure 6: Violation Repair

(line 3). If yes, it determines the number of  $sc$  contained in the  $cg_l$  (line 4).

If there is only one  $sc$   $sc_{old}$ , violation detection works as follows. If the RHS cell of the current sub-tuple,  $c_{cur}$ , has the same value as  $sc_{old}$ , it emits a non-violation message (line 5-6). Otherwise, a violation has been detected: the DW emits a *complete* violation message, containing both the current cell and the old cell (line 8).

If the  $cg$  contains more than one  $sc$ , the DW emits a single *append-only* violation message, which only contains the cell of the current sub-tuple (line 11). Such compact messages omits the  $sc$  from the data history, since they must be contained in earlier violation messages.

Finally, if the lookup procedure (line 3) fails, the DW creates a new cell group and emits a non-violation message (line 14-15).

At this point, the current cell  $c_{cur}$  is added to the corresponding group  $cg_l$  (line 17), either in an existing  $sc$ , or as a new distinct cell.

It is worth noticing that, following Algorithm 1, a DW emits a single message for each input sub-tuple, no matter how many tuples in the data history it conflicts with.

### 3.1.3 The Egress Router

The egress router gathers (violation or non-violation) messages for a given data tuple, as received from all DWs. Such messages are then sent together downstream towards the repair module.

## 3.2 Violation Repair

The goal of this module is to take the repair decisions for dirty data tuples, based on an intermediate stream of violation messages generated by the detect module. To achieve data repair, Bleach uses a data structure called *violation graph*, as outlined in Section 2. Violation messages contribute to the creation and dynamics of the violation graph, which essentially groups those cells that, together, are used to perform data repair.

Figure 6 sketches the internals of the repair module: it consists of an ingress router, the repair workers (RW), and an aggregation component that emits clean data. An additional component, called the coordinator, is used to steer violation graph management, with the contribution of RWs.

### 3.2.1 The Ingress Router

The ingress router is a simple component that actually broadcasts all incoming violation messages to all RWs. As opposed to its counterpart in the detection module, the ingress router does not perform data partitioning: instead, RWs are in charge of using only relevant information contained in the violation messages they receive, with the goal of creating and maintaining the violation graph.

### 3.2.2 The Repair Worker

Next, we delve into the details of the operation of a RW. First, we focus on the violation graph and the data repair algorithm we implement in Bleach. Then, we move to the key challenge that RWs address, that is how to maintain a *distributed* violation graph. As such, we focus on graph partitioning and maintenance. Due to violation graph dynamics, coordination issues might arise in a distributed setting: such problems are addressed by the coordinator component. **The repair algorithm.** Current data repair algorithms use the concept of a violation graph to repair dirty data based on user-defined rules. As outlined in Section 2, a violation graph is a succinct representation of cells (both current and historical) that are in conflict according to some rules. A violation graph is composed of subgraphs. As incoming data streams in, the violation graph evolves: specifically, its subgraphs might merge or split, depending on the contents of violation messages.

Using the violation graph, several algorithms can perform data cleaning, such as the equivalence class algorithm [5] or the holistic data cleaning algorithm [9]. Currently, Bleach relies on an *incremental* version of the equivalence class algorithm, that supports streaming input data, although alternative approaches can be easily plugged in our system. Thus, a subgraph in the violation graph can be interpreted as an equivalence class, in which all cells are supposed to have the same value.

**Distributed violation graph.** Due to the unbounded nature of streaming data, it is reasonable to expect the violation graph to grow to sizes exceeding the capacity of a single RW. As such, in Bleach, the violation graph is a distributed data structure, partitioned across all RWs.

However, unlike for DWs, the partitioning scheme can not be simply rule based, because a cell may violate multiple rules, creating issues related to coordination and load balancing. More generally, no partitioning scheme can guarantee that cells from a single violation message to be placed in the same partition to store subgraphs in a single RW.

Next, we describe how Bleach builds and maintains a distributed violation graph. The graph is built using  $msg_{vio}$  output by the egress router, which all RWs receive. Upon receiving a violation message, RWs process it independently, according to the following rules: *i*) if any of the current or old cells encapsulated in the message are already contained in an existing subgraph, both cells are added to the existing subgraph; *ii*) if an existing subgraph has cells which are in the same cell group as any of the cells in the message, the cells in the message are both added to the existing subgraph; *iii*) if any of these two cells are contained in multiple subgraphs, these subgraphs need to merge; *iv*) if none of these two cells is already contained in any subgraph, a new subgraph will be created with these two cells.

We define a *subgraph identifier*  $id(sg_k)$  to be the list of cell group IDs comprised in  $msg_{vio}$ :  $id(cg_1, cg_2, \dots)$ . A subgraph can be expressed as  $sg_k = (id(cg_1, cg_2, \dots), [sc_1, sc_2, \dots])$ : it

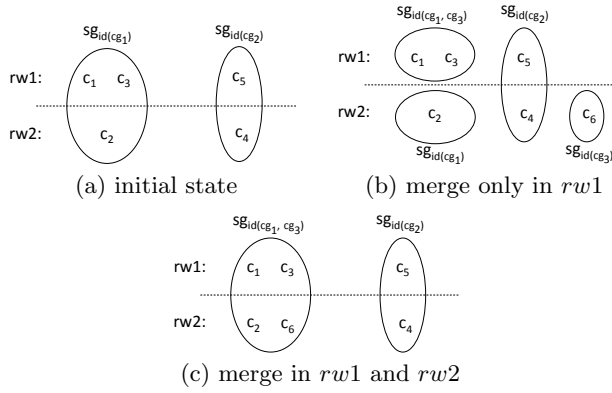


Figure 7: Violation graph build example

consists of a group of  $sc$ , stored in compressed format, as shown in Section 3.1.2. Note that when two subgraphs merge, their identifiers are also merged by concatenating both  $cg$  ID lists. To make the subgraph ID clear,  $sg_k$  can be presented as  $sg_{id(cg_1, cg_2, \dots)}$ .

Since subgraphs are collections of cells, we distribute the latter across all RWs, using the cells tuple IDs for partitioning. Then, we use the subgraph identifier to recognize partitions from the same subgraph. As a consequence, a subgraph spans several RWs, each storing a fraction of the cells comprised in the subgraph.

Finally, we note that the violation graph, and in particular the subgraph partitions stored by each RW, materializes as a data structure stored in RAM. Such data structure is organized similarly to that of the data history presented in Section 3.1, which allows an efficient execution of the repair algorithm, and a compact data representation.

An illustrative example is in order. Let's assume there are two RWs,  $rw1$  and  $rw2$ , and the current violation graph consists in two subgraphs  $sg_{id(cg_1)}$ , containing cells  $c_1, c_2, c_3$ , and  $sg_{id(cg_2)}$ , containing cells  $c_4, c_5$ . In our example, the violation graph is partitioned as in Figure 7(a): both RWs have a portion of cells of every subgraph.

### 3.2.3 The Coordinator

The problem we address now is due to the dynamics of the violation graph, which evolves as new violation messages stream into the Repair module. As each subgraph is partitioned among all RWs, subgraph partitions must be identified by the same ID: this is because a subgraph is a proxy for an equivalence class, and all its cells contribute to the correct functioning of the repair algorithm.

Continuing with the example from Figure 7(a), suppose a new violation message  $\{id(cg_3), c_6, c_1\}$  is received by all RWs. Now, in  $rw1$ , the new violation is added to subgraph  $sg_{id(cg_1)}$  since both the message and the subgraph share the same cell  $c_1$ : as such, the new subgraph becomes  $sg_{id(cg_1, cg_3)}$ . Instead, in  $rw2$ , the new violation triggers the creation of a new subgraph  $sg_{id(cg_3)}$ , since no common cells are shared between the message and existing subgraphs in  $rw2$ . The violation graph becomes *inconsistent*, as shown in Figure 7(b): this is a consequence of the independent operation of RWs. Instead, the repair algorithm requires the violation graph to be in a consistent state, as shown in Figure 7(c), where both RWs use the same subgraph identifier

for the same equivalence class.

To guarantee the consistency of the violation graph among independent RWs, Bleach uses a stateless coordinator component that helps RWs agree on subgraph identifiers. In what follows we present three variants of the simple protocol RWs use to communicate with the coordinator.

**RW-basic.** When a RW receives violation messages for a tuple, it adds the cells in the messages to the violation graph, according to its local state. Then, the RW creates a merge proposal containing the subgraph id for each conflicting attribute, and sends it to the coordinator.

Once the coordinator receives merge proposals from all RWs, it produces a merge decision, which contains a list of all  $cg$  IDs contained in the various merge proposals, and broadcasts it to all RWs. RWs merge their local subgraphs and converge to a globally consistent state.

Clearly, such a simple approach to coordination harms Bleach performance. Indeed, the RW-basic scheme requires one round-trip message for every incoming data tuple, from all RWs.

However, we note that it is not necessarily true that the coordination is always needed for every tuple. For example, when every cell violates at most one rule, every subgraph would only have a single  $cg$  ID. Thus, coordination is not necessary. More generally, given violation messages for a tuple, coordination is only necessary when there is a complete violation message containing an old cell which already exists in the violation graph because of a different violation rule.

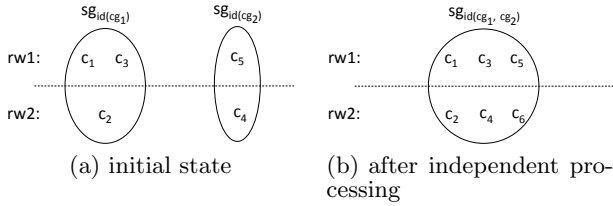
Figure 8 gives an example, where the initial state (Figure 8(a)) is the same as in Figure 7(a). Then, two violation messages,  $\{id(cg_1), c_6, null\}$  and  $\{id(cg_2), c_6, null\}$ , are received. Cell  $c_6$  is a current cell contained in the current tuple. Obviously  $sg_{id(cg_1)}$  and  $sg_{id(cg_2)}$  should merge into  $sg_{id(cg_1, cg_2)}$ . This can be accomplished without coordination by both repair workers, as shown in Figure 8(b). Indeed, each RW is aware that  $c_6$  is involved in two subgraphs, although  $c_6$  is only stored in  $rw2$  because of the partitioning scheme.

Next, we use the above observations and propose two variants of the coordination mechanism that aim at bypassing the coordinator component to improve performance. In both variants, if there exist subgraphs which can merge correctly without coordination like the example of Figure 8, they merge immediately.

**RW-dr.** In RW-dr, the coordination is only conducted if it is necessary, and the repair worker sends a merge proposal to the coordinator and waits for the merge decision. However, this approach is not exempt from drawbacks: it may cause some data tuples in the stream to be delivered out of order. This is because the repair worker wait for the merge decision in a non-blocking way. The violation messages of a tuple which do not require coordination may be processed in the coordination gap of an earlier tuple.

**RW-ir.** With this variant, no matter if the violation messages of a tuple require coordination or not, a RW immediately updates its local subgraphs, executes the repair algorithm and emits a repair proposal downstream to the aggregator component. Then, if necessary, the RW lazily executes the coordination protocol.

Clearly, this approach caters to system performance and avoids tuples to be delivered out of order, but might harm cleaning accuracy. Indeed, individual data repair proposals from a RW are based on a local view prior to finishing all



**Figure 8: Example of violation graph built without coordination**

necessary merge operations on subgraphs, which has a direct impact on equivalence classes.

### 3.2.4 The Aggregator

Using the distributed violation graph, each RW executes independently the Bleach repair algorithm and emits a data repair proposal, which includes all<sup>3</sup> candidate values and their frequency computed in a local subgraph partition. The aggregator component collects all repair proposals and selects the candidate value to repair a given cell as the one having the highest aggregate frequency. Finally, the aggregator modifies the current data tuple and outputs a clean data stream.

Note that the aggregator only modifies current tuples. Instead, more importantly, the cells stored in the violation graph are not modified regardless of the repair decision: this allows to update frequency counts as new data streams into the system, thus steering the aggregator to make different repair decisions as the violation graph evolves.

## 4. DYNAMIC RULE MANAGEMENT

In stream data cleaning the rule set is usually not immutable but dynamic. Therefore, we now introduce a new component, the rule controller, shown in Figure 3, which allows Bleach to adapt to rule dynamics. The rule controller accepts rule updates as input and guides the detect and the repair module to adapt to rule dynamics without stopping the cleaning process and without losing state. Rule updates can be of two types: one for adding a new rule and one for deleting an existing rule.

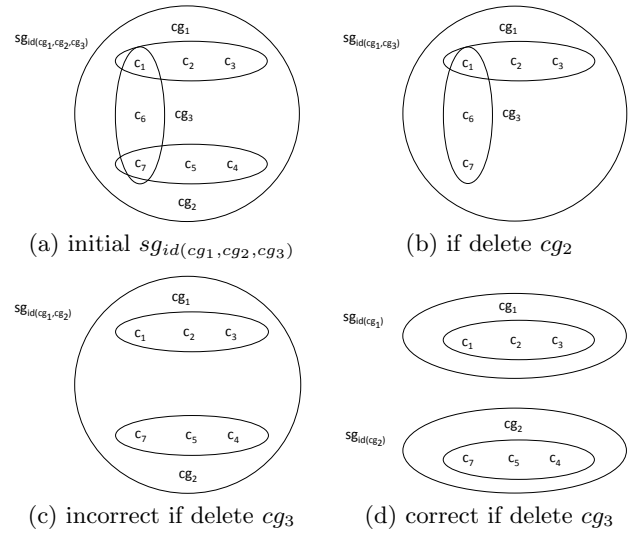
**Detect.** In the detect module, the addition of a rule triggers the instantiation of a new DW, as input tuples are partitioned by rule. The new DW starts with no state, which is built upon receiving new input tuples. As such, violation detection using past tuples cannot be achieved, which is consistent with the Bleach design goals. Instead, the deletion of an existing rule simply triggers the removal of a DW, with its own local data history.

**Repair.** In the Repair module, the addition of a new rule is not problematic with respect to violation graph maintenance operations. Instead, the removal of a rule implies violation graph dynamics (subgraphs might shrink or split) which are more challenging to address.

Thus, in a subgraph, we further group cells by cell groups. A subgraph now can also be expressed as:

$sg_k = (id(CG_1, CG_2, \dots), [CG_1, CG_2, \dots])$ , where each cell group gathers super cells. Some cells might span multiple groups, as they may violate multiple rules. We label such peculiar

<sup>3</sup>In case there are too many candidate values, we only send the top- $k$  values, where  $k = 5$ .



**Figure 9: Subgraph split example**

cells as *hinge cells*. For each hinge cell, the subgraph keeps the IDs of its connecting cell groups:  $c_{i,j}^* = (c_{i,j}, id(CG_{i_1}, CG_{i_2}, \dots))$ . Hinge cells with the same value and the same connecting cell groups are also compressed into super cells.

With the new organization of cells in subgraphs, the violation graph updates as following upon the removal of a rule. If a subgraph contains a single cell group related to the deleted rule, RWs are simply instructed to remove it. If a subgraph contains multiple cell groups, RWs remove the cell groups related to the deleted rule and update the hinge cells. With the remaining hinge cells, RWs check the connectivity of the remaining cell groups in the subgraph and decide to split the subgraph or not<sup>4</sup>.

An example of a split operation can be seen in Figure 9. The initial state of a subgraph is shown in Figure 9(a): the subgraph is  $sg_{id(CG_1, CG_2, CG_3)}$ , and its contents are three cell groups. Cell  $c_1$  and  $c_7$  are *hinge cells*, which work as bridges, connecting different cell groups together. Now, as a simple case, assume we want to remove the rule pertaining to  $CG_2$ : the subgraph should become  $sg_{id(CG_1, CG_3)}$ , as shown in Figure 9(b). Note that cell  $c_7$  loses its status of hinge cell. A more involved case arise when we delete the rule pertaining to  $CG_3$  instead of the rule pertaining to  $CG_2$ . In this case, the subgraph should not become  $sg_{id(CG_1, CG_2)}$  as shown in Figure 9(c). Indeed, removing  $CG_2$  eliminates all existing hinge cells connecting the remaining cell groups. Thus, the subgraph must split in two separate subgraphs  $sg_{id(CG_1)}$  and  $sg_{id(CG_2)}$  as shown in Figure 9(d).

## 5. WINDOWING

Bleach provides windowed computations, which allow expressing data cleaning over a sliding window of data. Despite being a common operation in most streaming systems, window-based data cleaning addresses the challenge of the unbounded nature of streaming data: without windowing, the data structures Bleach uses to detect and repair a dirty stream would grow indefinitely, which is unpractical.

<sup>4</sup>A detailed algorithm can be found in our technical report: <http://www.eurecom.fr/~tian/bleach/bleachTR.pdf>

In this section, we discuss two windowing strategies: a basic, tuple-based windowing strategy and an advanced strategy that aim at improving cleaning accuracy.

## 5.1 Basic Windowing

The underlying idea of the basic windowing strategy is to only use tuples within the sliding window to populate the data structures used by Bleach to achieve its tasks. Next, we outline the basic windowing strategy for both DWs and RWs operation.

**Windowed Detection.** We now focus on how DWs maintain their local data history. Clearly, the data history only contains cells that fall within the current window. When the window slides forward, DWs update the data history as follows: *i*) if a cell group ends up having no cells in the new window, DWs simply delete it; *ii*) for the remaining cell groups, DWs drop all cells that fall outside the new window, and update accordingly the remaining super cells.

Note that, if implemented naively, the first operation above can be costly as it involves a linear scan of all cell groups. To improve the efficiency of data history updates, Bleach uses the following approach. It creates a FIFO queue of  $k$  lists, which store cell groups. In case the sliding step is half the window size,  $k = 2$ ; more generally, we set  $k$  to be the window size divided by the sliding step. Any new cell group from the current window enters the queue in the  $k$ -th list. Any cell group updates, e.g. due to a new cell added to the cell group, “promotes” it from list  $j$  to list  $k$ . As the window slides forward, the queue drops the list (and its cell groups) in the first position and acquires a new empty list in position  $k + 1$ .

**Windowed Repair.** Now we focus on how to maintain the violation graph in RWs. Again, the violation graph only stores cells within the current window. When the window slides forward, RWs update the violation graph as follows:

- If a subgraph has no cells in the new window, RWs delete the subgraph;
- For the remaining subgraphs, if a cell group has no cells in the new window, RWs delete the cell group;
- RWs also delete hinge cells that are outside of the new window. This could require subgraphs to split, as they could miss a “bridge” cell to connect its cell groups;
- For the remaining cell groups, RWs drop all cells outside of the new window, and update the remaining super cells accordingly.

For efficiency reasons, Bleach uses the same  $k$ -list approach described for DWs to manage violation graph updates due to a sliding window.

## 5.2 Bleach Windowing

The basic windowing strategy only relies on the data within the current window to perform data cleaning, which may limit the cleaning accuracy. We begin with a motivating example, then describe the Bleach windowing strategy, that aims at improving cleaning accuracy. Note that here we only focus on the repair module and its violation graph, since Bleach windowing does not modify the operation of the detect module.

Figure 10(a) illustrates a data stream of two-attribute tuples. Assume we use a single FD rule ( $A \rightarrow B$ ), a window

	A	B	
$t_1$	a	b	window [1, 4]
$t_2$	a	b	
$t_3$	a	b	
$t_4$	a	c	
$t_5$	a	c	window [3, 6]
$t_6$	a	b	

(a) input data

	A	B
$t_1$	a	b
$t_2$	a	b
$t_3$	a	b
$t_4$	a	b
$t_5$	a	c
$t_6$	a	b

(b) output data with basic windowing

	A	B
$t_1$	a	b
$t_2$	a	b
$t_3$	a	b
$t_4$	a	b
$t_5$	a	b
$t_6$	a	b

(c) output data with Bleach windowing

**Figure 10: Motivating example: Basic vs. Bleach windowing.**

size of 4 tuples, a sliding step of 2 tuples, and the *basic* windowing strategy.

When  $t_4$  arrives, the window covers tuples [1, 4]. According to the repair algorithm, Bleach repairs  $t_4(B)$  and sets it to the value  $b$ . Now, when tuple  $t_5$  arrives, the window moves to cover tuples [3, 6], even though  $t_6$  has yet to arrive. With only three tuples in the current window, the algorithm determines  $t_5(B)$  is correct, because now the majority of tuples have value  $c$ . The output stream produced using basic windowing is shown in Figure 10(b). Clearly, cleaning accuracy is sacrificed, since it is easy to see that  $t_5(B)$  should have been repaired to value  $b$ , which is the most frequent value overall. Hence, the need for a different windowing strategy to overcome such problems.

Bleach windowing relies on an extension of a super cell, which we call a *cumulative super cell*. The idea is for the violation graph to accumulate past state, to complement the view Bleach builds using tuples from the current window. Hence, a cumulative super cell is represented as a super cell, with an additional field that stores the number of occurrences of cells (both hinge as well as super cells within a cell group) with the same RHS value, including those that have been dropped because they fall outside the sliding window boundaries.

When using Bleach windowing, RWs maintain the violation graph by storing cumulative super cells. When the window slides forward, RWs update the violation graph as follows. The first two steps are equivalent to those for the basic strategy. The last two steps are modified as follows:

- For the remaining subgraphs, RWs update hinge cells, and “flush” those that do not bridge cell groups anymore because of the update. Also, RWs split subgraphs according to the remaining hinge cells;
- For the remaining cell groups and hinge cells, RWs update compressed super cells, “flushing” cells which fall outside the new window while keeping their count.



The “flush” operation used above only drops the content of a super cell, but keeps its structure and its count field.

Now, going back to the example in Figure 10(a), when tuple  $t_5$  arrives, Bleach stores two cumulative super cells:  $csc_1(id(t) = 3, value = 'b', count = 3)$  and  $csc_2(id(t) = [4, 5], value = 'c', count = 2)$ . Although  $t_1$  and  $t_2$  have been deleted because they are outside the sliding window, they still contribute to the count field in  $csc_1$ . Therefore, tuple  $t_5(B)$  is correctly repaired to value  $b$ , as shown in Figure 10(c).

**Additional notes:** When using cumulative super cells, Bleach keeps tracks of candidate values to be used in the repair algorithm as long as cell groups remain. By using cumulative super cells for hinge cells, subgraphs only split if some cell groups are removed when the window moves forward. Note that the introduction of cumulative super cells does not interfere with dynamic rule management: in particular, when deleting a rule, subgraphs update correctly when hinge cells use the cumulative format. Overall, to compute the count of a candidate value in a subgraph, cumulative super cells accumulate the counts of relevant compressed super cells from all cell groups, taking into account any duplicate contributions from hinge cells.

Obviously, Bleach windowing requires more storage than basic windowing, as cumulative super cells store additional information, and because of the “flush” operation described earlier, which keeps a super cell structure, even when it has an empty cell list. Section 6 demonstrates that such additional overhead is well balanced by superior cleaning accuracy, making Bleach windowing truly desirable.

## 6. EVALUATION

Bleach prototype implementation is built using Apache Storm [27].<sup>5</sup> Input streams, including both the data stream and rule updates, are fed into Bleach using Apache Kafka [18].

Our goal is to demonstrate that Bleach achieves efficient stream data cleaning under real-time constraints. Our evaluation uses *throughput*, *latency* and *dirty ratio* as performance metrics. We express the dirty ratio as the fraction of dirty data remaining in the output data stream: the smaller the dirty ratio, the higher the cleaning accuracy. The processing latency is measured from uniformed sampled tuples (1 per 100). All our experiments are conducted in a cluster of 18 machines, with 4 cores, 8 GB RAM and 1 Gbps network interface each.

We evaluate Bleach using a synthetic dataset generated from TPC-DS (with scale factor 100 GB). To do so, we join a fact table *store\_sales* with its dimension tables in TPC-DS to build a single table (288 million tuples). By exporting this table to Kafka, we simulate an “unbounded” data stream. We manually design eight CFD rules, as shown in Table 1. Among these rules,  $r_4$  and  $r_5$  have the same RHS attribute *s\_store\_name*, while  $r_6$  and  $r_7$  have the same RHS attribute *c\_email\_addr*, as intersecting attributes.

We generate a dirty data stream, according to our rules, as follows: we modify the values of RHS attributes with probability 10% and replace the values of LHS attributes with NULL with probability 10%.<sup>6</sup> In all the experiments,

<sup>5</sup>Nothing prevents Bleach to be built using alternative systems such as Apache Flink, for example.

<sup>6</sup>In our experiments we also use BART [3], which is a well accepted dirty data generator. However, due to the sheer

**Table 1: Rule Sets**

$r_0$	$ss\_item\_sk \rightarrow i\_brand, (ss\_item\_sk \neq null)$
$r_1$	$ss\_item\_sk \rightarrow i\_category, (ss\_item\_sk \neq null)$
$r_2$	$ca\_state, ca\_city \rightarrow ca\_zip, (ca\_state, ca\_city \neq null)$
$r_3$	$ss\_promo\_sk \rightarrow p\_promo\_name, (ss\_promo\_sk \neq null)$
$r_4$	$ss\_store\_sk \rightarrow s\_store\_name, (ss\_store\_sk \neq null)$
$r_5$	$ss\_ticket\_num \rightarrow s\_store\_name, (ss\_ticket\_num \neq null)$
$r_6$	$ss\_ticket\_num \rightarrow c\_email\_addr, (ss\_ticket\_num \neq null)$
$r_7$	$ss\_customer\_sk \rightarrow c\_email\_addr, (ss\_customer\_sk \neq null)$

we set the window size to 2 M and the sliding step to 1 M tuples respectively, regardless which windowing strategy we use. If not otherwise specified, we use Bleach windowing as the default strategy.

### 6.1 Comparing Coordination Approaches

In this experiment we compare the three RW approaches discussed in Section 3.2, according to our performance metrics, as shown in Figure 11: *RW-basic* requires coordination among repair workers for each tuple; *RW-dr* omits coordination for tuples when possible; *RW-ir* is similar to *RW-dr*, but allows repair decisions to be made before finishing coordination. Next, we use our synthetic dataset and rules  $r_0$  to  $r_5$ .

Figure 11(a) shows how Bleach throughput evolves with processed tuples. The throughput with both *RW-dr* and *RW-ir* is around 15K tuples/second, whereas *RW-basic* achieves roughly 13K tuples/second. The inferior performance of *RW-basic* is due to the large number of coordination messages required to converge to global subgraph identifiers, while *RW-dr* and *RW-ir* only require 7% coordination messages in *RW-basic*.

Figure 11(b) shows the CDF of the tuple processing latency for the three RW approaches. *RW-basic* has the highest processing latency, on average 364 ms. The processing latency of *RW-ir* is on average 316 ms. *RW-dr* average latency is slightly higher, about 323 ms. This difference is due again to the additional round-trip-messages required by coordination: with *RW-ir*, RWs make their repair proposals without waiting for coordination to complete, therefore the small processing latency.

Figure 11(c) illustrates the cleaning accuracy. All three approaches lower the ratio of dirty data significantly, from the initial 10% to at most 0.5% (even 0% for rule  $r_3$  and  $r_4$ ). For the first five rules, the three approaches achieve similar cleaning accuracy. Instead, for rule  $r_5$  the *RW-ir* method suffers and the dirty ratio is larger. Indeed, for rule  $r_5$  whose cleaning accuracy is heavily linked to rule  $r_1$ , *RW-ir* fails to correctly update some of its subgraphs because it eagerly emits repair proposals without waiting for coordination to complete.

### 6.2 Comparing Windowing Strategies

In this experiment, we compare the performance of the basic and Bleach windowing strategies, and use the *RW-dr* mechanism. As for the previous experiment, we use rules  $r_0$  to  $r_5$ . Additionally, for stress testing, we increase the input dirty data ratio from 10% to 50% for data in the interval from 40 M to 42 M tuples.

size of our data stream, we present results obtained using our custom process, which mimics that of BART but scales to large data streams.

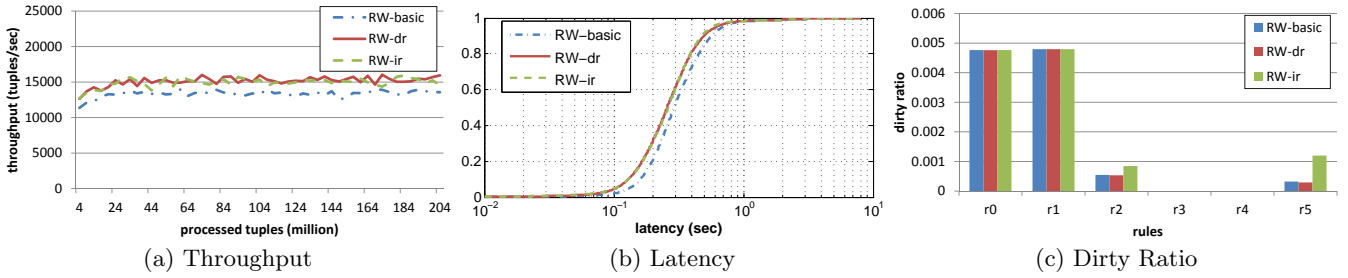


Figure 11: Comparison of coordination mechanisms: RW-basic, RW-dr and RW-ir.

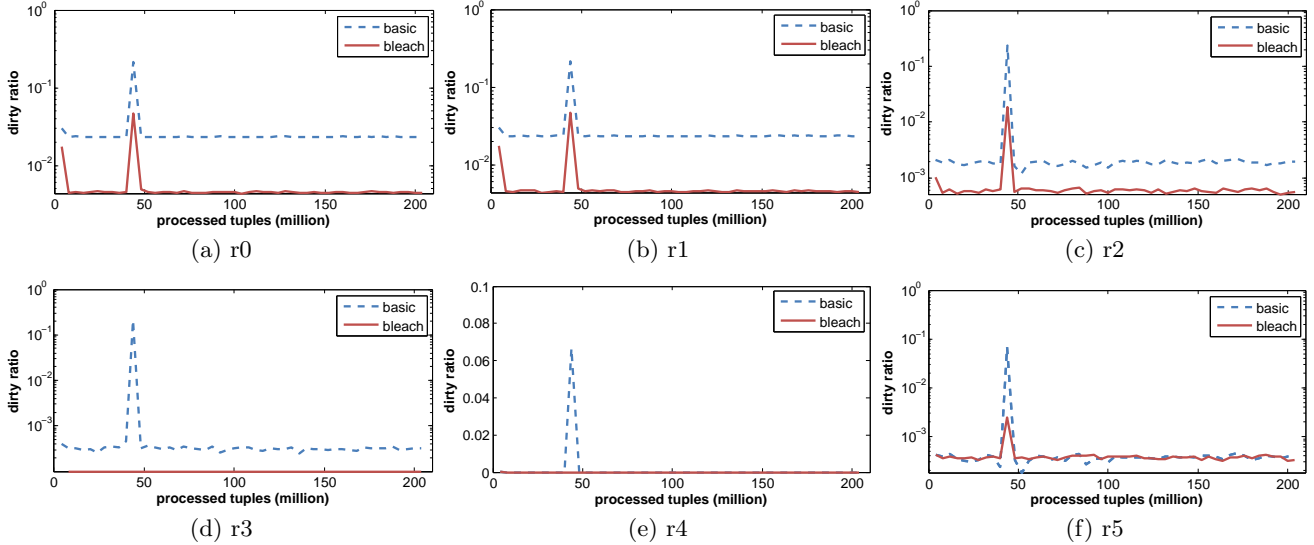


Figure 14: cleaning accuracy of two windowing strategies

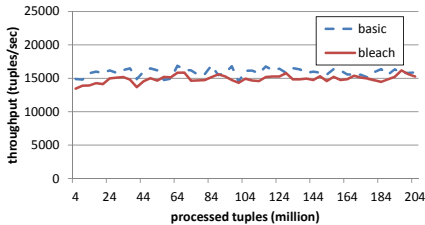


Figure 12: throughput of two windowing strategies

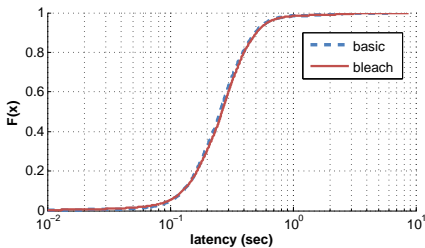


Figure 13: processing latency CDF of two windowing strategies

As shown in Figure 12 and Figure 13, the two windowing strategies are essentially equivalent in terms of throughput and latency: this is good news, as it implies the requirement for *cumulative super cells* is a negligible toll on performance.

Next, we focus on a detailed view of the cleaning accuracy, which is shown in Figure 14. Bleach windowing achieves superior cleaning accuracy: in general, the dirty ratio is one order of magnitude smaller than that of basic windowing. This advantage is kept also in presence of a 50% dirty ratio spike in the input data. In particular, for rules  $r_3$  and  $r_4$ , Bleach windowing achieves 0% dirty ratio, irrespectively of the dirty ratio spike.

Overall, Bleach windowing reveals that keeping state from past windows can indeed dramatically improve cleaning accuracy, with little to no performance overhead.

### 6.3 Dynamic Rule Management

Next, we study the performance of Bleach in presence of rule dynamics, as shown in Figure 15. To do this, we initially use the same input data stream and rule set as in Section 6.1. However, while Bleach is cleaning the input stream, we delete rule  $r_5$  and add rule  $r_6$  and  $r_7$ , as indicated in the Figure. In this experiment, we use Bleach windowing strategy and *RW-dr* coordination.

Figure 15(a) and Figure 15(b) show the evolution in time of throughput and latency, whereas Figure 15(c) gives the CDF of the processing latency.

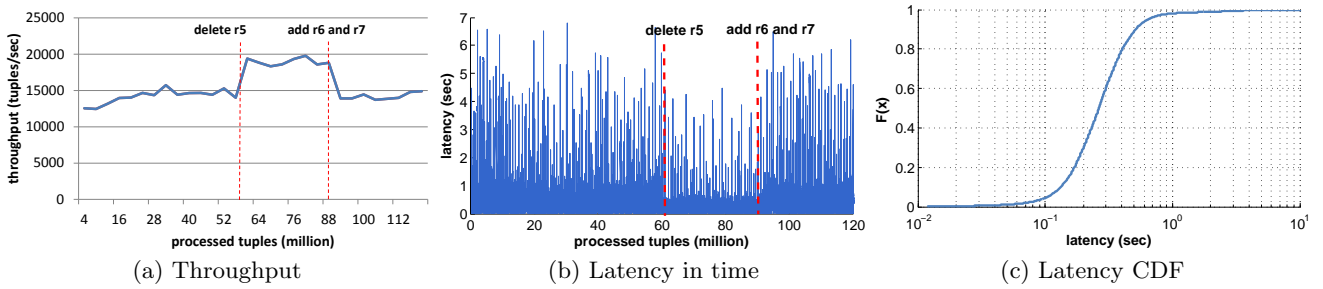


Figure 15: Bleach performance with dynamic rule management.

Figure 15(a) shows that rule dynamics can result in an increase in throughput. Indeed, removing  $r_5$  (at the 60M tuple) implies that Bleach needs to manage fewer rules; in addition,  $r_4$  becomes simpler to manage, as there are no more intersections with  $r_5$ . Similarly, Figure 15(b) shows that also latency decreases upon  $r_5$  removal. When rules  $r_6$  and  $r_7$  are added (at the 90 M tuple), the throughput drops and the latency grows, as Bleach has more rules to manage and because the new rules have intersecting attributes, requiring more work from RWs.

Figure 15(c), shows the latency distribution computed from output tuple samples. While the average latency is roughly 320 ms, we notice a tail in the distribution, indicating that some (few) tuples experience latencies up to seconds. This has been observed across all our experiments, and is due to the sliding window mechanism, which imposes computationally demanding operations when updating the violation graph, resulting also in rather low-level garbage collection problems.

Overall, we conclude that Bleach supports dynamic rule management seamlessly, with essentially no impact on performance, and no system restart required.

## 6.4 Comparing Bleach to a Baseline Approach

We conclude our evaluation with a comparative analysis of Bleach and a baseline approach, which follows the ideas discussed in Section 1.

To do so, we design and implement a new system that is based on the micro-batch streaming paradigm: essentially, such system buffers input data records and performs data cleaning periodically, as determined by a sliding window. Our implementation uses Apache Spark, and uses its Streaming API that supports all necessary operators.<sup>7</sup> We refer to the baseline approach as micro-batch cleaning.

To demonstrate the performance of micro-batch cleaning and compare it to Bleach, we perform a series of experiments whereby we increase the sliding window size. We use the same stream data input from our previous experiments, but only use a single rule,  $r_0$ . Here, we focus on performance analysis expressed in terms of latency and dirty ratio, thus we feed the input stream at a constant throughput of 15000 tuples/second.

Figure 16 illustrates the performance of both systems. As expected, for the micro-batch baseline approach, the average latency is proportional to the window size: larger sliding windows entail higher latencies. Indeed, since the data in the

<sup>7</sup>To be precise, note that window processing in Spark Streaming is time-based and not tuple-based. For our experiment, this difference is negligible.

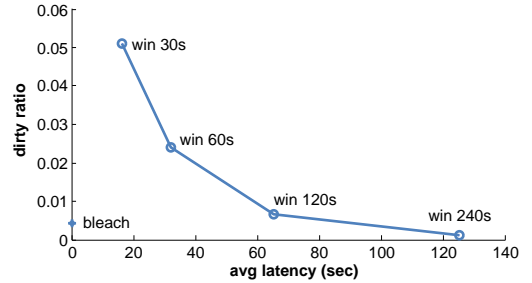


Figure 16: micro-batch cleaning result

input stream is uniformly distributed, the average latency equals the sum of half of the window size and the average execution time for cleaning data in each window.

As for the cleaning accuracy, the larger the sliding window, the more opportunities the micro-batch system has to clean the input stream, hence a smaller output stream dirty ratio. In particular, we notice that to achieve the same cleaning accuracy as Bleach, micro-batch cleaning requires the sliding window to be larger than 120 seconds, which incurs in an average latency larger than 1 minute. Instead, in Bleach, the average latency is about 320 ms.

We conclude that Bleach represents a valuable tool in the data cleaning landscape if real-time requirements must be met, while achieving high cleaning accuracy.

## 7. RELATED WORK

In recent years, data cleaning systems have flourished. Many approaches [6, 20, 4, 9, 10, 19, 16] tackle the problem of detecting and repairing dirty data based on predefined data quality rules. [9] proposes a way to combine multiple rules together and to perform data cleaning work holistically. [7] focuses on functional dependency violations in a horizontally partitioned database, aiming to minimize data shipment and parallel computation time. NADEEF[10] is an extensible and generic data cleaning system and BigDansing[19] is a large-scale version of NADEEF, which executes data cleaning job in frameworks like Hadoop and Spark. These approaches are effective when data is static.

[12, 13] provide incremental algorithms to detect errors in distributed data when data is updated, which is similar to violation detection in Bleach. [29] introduces a continuous data cleaning framework that can be applied to evolving data and rules driven by a classifier. These works focus on cleaning data stored in a data warehouse by batch processing, which achieves high accuracy but suffers high latency.

In contrast, stream processing requires to be real-time, a challenge that has drawn increasing attention from researchers [2, 14, 21, 11]. Nevertheless, stream data cleaning approaches are still in their infancy. Some works [24, 31, 17] focus on RFID or sensor stream data cleaning where the data is a sequence of numerical values. These works achieve data cleaning by operations like smoothing or deleting outliers. Instead, Bleach focuses on more general cases where data can be both numerical and categorical. In Spark Streaming [25], a data stream can be cleaned by joining it with precomputed information. However, precomputed information is not always available nor sufficient for accurate data cleaning. To the best of our knowledge, Bleach is the first stream data cleaning system based on data quality rules providing both high accuracy and low latency.

There are many other research work about data cleaning. For example, [30] and [8] are about how to perform data cleaning via knowledge base and crowdsourcing. BART [3] is an dirty data generator for evaluating data-cleaning algorithms. [1] studies the problem of temporal rules discovery for dirty web data. All such works are orthogonal to ours.

## 8. CONCLUSION

This work introduced Bleach, a novel stream data cleaning system, that aims at efficient and accurate data cleaning under real-time constraints.

First, we have introduced the design goals and the related challenges underlying Bleach, showing that stream data cleaning is far from being a trivial problem. Then we have illustrated the Bleach system design, focusing both on data quality – we have introduced dynamic rule sets, and a stateful approach to windowing – and on systems aspects – we have addressed problems related to data partitioning and coordination, which are required by the distributed nature of Bleach. We also have provided a series of optimizations to improve system performance, by using compact and efficient data structures, and by reducing the messaging overhead.

Finally, we have evaluated a prototype implementation of Bleach: our experiments showed Bleach achieves low-latency and high cleaning accuracy, while absorbing a dirty data stream, despite rule dynamics. We also have compared Bleach to a baseline system built on the micro-batch paradigm, and explained Bleach superior performance.

Our plan for future works is to support a more varied rule set and to explore alternative repair algorithms, that might require revisiting the inner data structures we use in Bleach.

## 9. REFERENCES

- [1] Abedjan et al. Temporal rules discovery for web data cleaning. In *Proc. of VLDB*, pages 336–347, 2015.
- [2] T. Akidau et al. The dataflow model: A practical approach to balancing correctness, latency, and cost in massive-scale, unbounded, out-of-order data processing. In *Proc. of VLDB*, pages 1792–1803, 2015.
- [3] Arocena et al. Messing up with bart: error generation for evaluating data-cleaning algorithms. In *Proc. of VLDB*, pages 36–47, 2015.
- [4] Beskales et al. Sampling the repairs of functional dependency violations under hard constraints. In *Proc. of VLDB*, number 1-2, pages 197–207, 2010.
- [5] Bohannon et al. A cost-based model and effective heuristic for repairing constraints by value modification. In *Proc. of SIGMOD*, pages 143–154, 2005.
- [6] P. Bohannon et al. Conditional functional dependencies for data cleaning. In *ICDE*, pages 746–755, April 2007.
- [7] Q. Chen et al. Repairing functional dependency violations in distributed data. In *DASFAA*, pages 441–457, 2015.
- [8] Chu et al. Katara: A data cleaning system powered by knowledge bases and crowdsourcing. In *Proc. of SIGMOD*, pages 1247–1261, 2015.
- [9] X. Chu et al. Holistic data cleaning: Putting violations into context. In *ICDE*, pages 458–469, 2013.
- [10] Dallachiesa et al. Nadeef: A commodity data cleaning system. In *Proc. of SIGMOD*, pages 541–552, 2013.
- [11] Elseidy et al. Scalable and adaptive online joins. In *Proc. of VLDB*, pages 441–452, 2014.
- [12] W. Fan et al. Incremental detection of inconsistencies in distributed data. In *ICDE*, pages 318–329, 2012.
- [13] W. Fan et al. Incremental detection of inconsistencies in distributed data. *IEEE Transactions on Knowledge and Data Engineering*, pages 1367–1383, 2014.
- [14] Fernandez et al. Liquid: Unifying nearline and offline big data integration. In *CIDR*, 2015.
- [15] Gdeltproject. Webpage. <http://gdeltproject.org/>.
- [16] Geerts et al. The llunatic data-cleaning framework. In *Proc. of VLDB*, pages 625–636, 2013.
- [17] Jeffery et al. A pipelined framework for online cleaning of sensor data streams. Technical report, 2005.
- [18] Kafka. Webpage. <http://kafka.apache.org/>.
- [19] Z. Khayyat et al. Bigdancing: A system for big data cleansing. In *Proc. of SIGMOD*, pages 1215–1230, 2015.
- [20] Kolahi et al. On approximating optimum repairs for functional dependency violations. In *Proc. of ICDT*, pages 53–62, 2009.
- [21] Q. Lin et al. Scalable distributed stream join processing. In *Proc. of SIGMOD*, pages 811–825.
- [22] Openrefine. Webpage. <http://openrefine.org/>.
- [23] Recordedfuture. Webpage. <https://www.recordedfuture.com/>.
- [24] S. Song et al. SCREEN: stream data cleaning under speed constraints. In *Proc. of SIGMOD*, pages 827–841, 2015.
- [25] Spark stream cleaning. Webpage. <http://spark.apache.org/docs/latest/streaming-programming-guide.html>.
- [26] Spark Summit 2015 Use Case. Webpage. <https://spark-summit.org/east-2015/streaming-machine-learning-in-spark/>.
- [27] Storm. Webpage. <https://storm.apache.org/>.
- [28] Trifacta. Webpage. <https://www.trifacta.com/>.
- [29] M. Volkovs et al. Continuous data cleaning. In *ICDE*, pages 244–255, 2014.
- [30] Wang et al. Crowd-based deduplication: An adaptive approach. In *Proc. of SIGMOD*, pages 1263–1277, 2015.
- [31] Zhao et al. A model-based approach for rfid data stream cleansing. In *Proc. of CIKM*, pages 862–871, 2012.