Efficient Black-box Checking of Snapshot Isolation in Databases

Anonymous Author(s)

ABSTRACT

Snapshot isolation (SI) is a prevalent weak isolation level that avoids the performance penalty imposed by serializability and simultaneously prevents various undesired data anomalies. Violations have however recently been found in production cloud databases that claim to provide the SI guarantee. Given the complex and often unavailable internals of such databases, a black-box SI checker is highly desirable.

In this paper we present PolySI, a novel black-box checker designed to efficiently check SI and provide understandable counterexamples upon detecting violations. PolySI builds on a novel characterization of SI using generalized polygraphs (GPs), for which we establish its soundness and completeness. PolySI employs an SMT solver and also accelerates SMT solving by utilizing the compact constraint encoding of GPs and domain-specific optimizations for pruning constraints. As demonstrated by our extensive assessment, PolySI successfully reproduces all of 2477 known SI anomalies, detects novel SI violations in three production cloud databases, identifies their causes, and outperforms the state-of-the-art black-box checkers under a wide range of workloads.

PVLDB Reference Format:
Anonymous Author(s). Efficient Black-box Checking of Snapshot Isolation in Databases. PVLDB, 14(1): XXX-XXX, 2020.
doi:XX.XX/XXX.XX

PVLDB Artifact Availability:
The source code, data, and/or other artifacts have been made available at https://github.com/anonymous-hippo/PolySI.

1 INTRODUCTION

Database systems are an essential building block of many software systems and applications. Transactional access to databases simplifies concurrent programming by providing an abstraction for executing concurrent computations on shared data in isolation [7]. The gold-standard isolation level, *serializability* (SER) [36], ensures that all transactions appear to execute serially, one after another. However, providing SER, especially in geo-replicated environments like modern cloud databases, is computationally expensive [3, 34].

Many databases provide weaker guarantees for transactions to balance the trade-off between data consistency and system performance. *Snapshot isolation* (SI) [6] is one of the prevalent weaker isolation levels used in practice, which avoids the performance penalty imposed by SER and simultaneously prevents undesired data anomalies such as fractured reads, causality violations, and lost updates [13]. In addition to classic centralized databases such as Microsoft SQL Server [40] and Oracle Database [22], SI is supported by numerous *production cloud* database systems, e.g., Google's Percolator [37], MongoDB [35], TiDB [46], YugabyteDB [49], Galera [17], and Dgraph [24].

Unfortunately, as recently reported in [30, 44, 45], violations have been found in several production cloud databases that claim to provide SI. This raises the question of whether such databases actually deliver the promised SI guarantee for transactions in practice. Given that their internals (e.g., source code and logs) are often unavailable to the outsiders or hard to digest, a black-box SI checker is highly desirable.

A natural question then to ask is "What should an ideal black-box SI checker look like?" The SIEGE principle [30] has already provided a strong baseline: an ideal checker would be *sound* (returning no false positives), *informative* (reporting understandable counterexamples), *effective* (detecting violations in real-world databases), *general* (compatible with different patterns of transactions), and *efficient* (with modest checking time even for workloads of high concurrency). Additionally, (i) we expect an ideal checker to be *complete*, thus missing no violations; and (ii) we also augment the *generality* criterion by requiring a black-box checker to be compatible not only with general (read-only, write-only, and read-write) transaction workloads but also with standard key-value/SQL APIs. We call this extended principle SIEGE+.

None of existing SI checkers, to the best of our knowledge, satisfy SIEGE+ (see Section 6.2 for the detailed comparison). For example, dbcop [8] is incomplete, incurs exponentially increasing overhead under higher concurrency (Section 5.4), and returns no counterexamples upon a violation; Elle [30] relies on specific database APIs such as lists and the (internal) timestamps of transactions to infer isolation anomalies, thus not conforming to our black-box setting.

The PolySI Checker. We present PolySI, a novel, black-box, SI checker designed to achieve all the SIEGE+ criteria. PolySI builds on three key ideas with respect to three major challenges.

First, despite previous attempts to characterize SI [1, 6, 48], its semantics is usually explained in terms of low-level implementation choices invisible to the database outsiders. Consequently, one must *guess* the dependencies (aka uncertain/unknown dependencies) between client-observable data, e.g., which of the two writes was first recorded in the database.

We introduce a novel dependency graph, called *generalized polygraph* (GP), based on which, we present a new *sound* and *complete* characterization of SI. There are two main advantages of a GP: (i) it naturally models the guesses by capturing *all* possible dependencies between transactions in a single compacted data structure; and (ii) it enables the acceleration of SMT solving by compacting constraints (see below) as demonstrated by our experiments.

Second, there have been recent advances in SAT/SMT solving for checking *graph properties* such as the MonoSAT solver [5] and its successful application to the black-box checking of SER [42].

Proceedings of the VLDB Endowment, Vol. 14, No. 1 ISSN 2150-8097.
doi:XX.XX/XXX.XX

The idea is to *search* for an acyclic graph where the nodes are transactions in the history[1] and the edges meet certain constraints. We show that SMT techniques can also be applied to build an effective SI checker. This application is nontrivial as a brute-force approach would be inefficient due to the high computational complexity of checking SI [8]: the problem is NP-complete in general and $O(n^c)$ with c (resp. n) a fixed, yet in practice large, number of clients (resp. transactions), even for a single history of transactions. In fact, checking SI has been proved to be asymptomatically more complex than checking SER [8]. In the context of SMT solving over graphs, SI leads to much larger search space due to its specific anomaly patterns [13] while checking SER simply requires finding a cycle.

Thanks to our GP-based characterization of SI, we leverage its compact encoding of constraints on transaction dependencies to accelerate MonoSAT solving. Moreover, we develop domain-specific optimizations to further prune constraints, thereby reducing the search space. For example, PolySI prunes a constraint if an associated uncertain dependency would result in an SI violation with known dependencies.

Finally, although MonoSAT outputs cycles upon detecting a violation, they are still *uninformative* with respect to understanding how the violation actually occurred. Locating the actual causes of violations would facilitate debugging and repairing the defective implementations. For example, if an SI checker were to identify a *lost update* anomaly from the returned counterexample, developers could then focus on investigating the write-write conflict resolution mechanism. Hence, we design and integrate into PolySI a novel interpretation algorithm that explains the counterexamples returned by MonoSAT. More specifically, PolySI (i) recovers the violating scenario by bringing back any potentially involved transactions and dependencies eliminated during pruning and solving and (ii) finalizes the core participants to highlight the violation cause.

Main Contributions. In summary, we provide:

1. a new GP-based characterization of SI that both facilitates the modeling of uncertain transaction dependencies inherent to black-box testing and also enables the acceleration of constraint solving (Section 3);
2. a sound and complete GP-based checking algorithm for SI with the domain-specific optimizations for pruning constraints (Section 4);
3. the PolySI tool comprising both our new checking algorithm and the interpretation algorithm for debugging; and
4. an extensive assessment of PolySI that demonstrates its fulfilment of SIEGE+ (Section 5). In particular, PolySI successfully reproduces all of 2477 known SI anomalies, detects novel SI violations in three production cloud databases, identifies their causes, and outperforms the state-of-the-art black-box checkers under a wide range of workloads.

2 PRELIMINARIES

2.1 Snapshot Isolation in a Nutshell

Snapshot isolation (SI) [6] is one of the most prominent weaker isolation levels that modern (cloud) databases usually provide to avoid the performance penalty imposed by *serializability* (SER). Figure 1 shows a hierarchy of isolation levels where SI sits inbetween *transactional causal consistency* [33] and SER, and is not comparable to *repeatable read* [1].

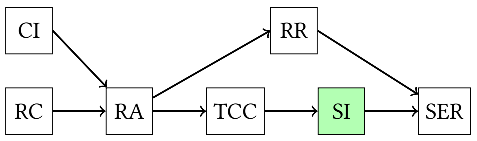


Figure 1: A hierarchy of isolation levels. $A \longrightarrow B$: A is strictly weaker than B. CI: cut isolation [3]; RC: read committed [6]; RA: read atomicity [4]; RR: repeatable read [1]; TCC: transactional causal consistency [33]; SI: snapshot isolation [13]; SER: serializability [6].

A transaction with SI always reads from a snapshot that reflects a single commit ordering of transactions and is allowed to commit if no concurrent transaction has updated the data that it intends to write. SI prevents various undesired data anomalies such as fractured reads, causality violations, lost updates, and long fork [6, 13]. The following examples illustrate two kinds of anomalies disallowed by SI. As we will see in Section 5, both anomalies have been detected by our PolySI checker in production cloud databases.

Example 1 (Causality Violation). Alice posts a photo of her birthday party. Bob writes a comment to her post. Later, Carol sees Bob's comment but not Alice's post.

Example 2 (Lost Update). Dan and Emma share a banking account with the current balance of 10 dollars. Both simultaneously deposit 50 dollars. The resulting balance is 60, instead of 110, as one of the deposits is lost.

In this paper we focus on the prevalent *strong session* variant of SI [13, 23], which additionally requires a transaction to observe all the effects of the preceding transactions in the same *session* [43]. Many production databases, including DGraph [24], Galera [17], and CockroachDB [18], provide this isolation level in practice.

2.2 Snapshot Isolation: Formal Definition

We recall the formalization of SI over dependency graphs, which serves as the theoretical foundation of PolySI. The following account is standard; see for example [13]. We consider a distributed key-value store managing a set of keys $\mathsf{Key} = \{x, y, z, \dots\}$, which take on a set of values Val.[2] We denote by Op the set of possible read or write operations on keys: $\mathsf{Op} = \{\mathsf{R}_\iota(x, v), \mathsf{W}_\iota(x, v) \mid \iota \in \mathsf{OpId}, x \in \mathsf{Key}, v \in \mathsf{Val}\}$, where OpId is the set of operation identifiers. We omit operation identifiers when they are not important.

2.2.1 Relations, Orderings, Graphs, and Logics. A binary relation R over a given set A is a subset of $A \times A$, i.e., $R \subseteq A \times A$. For $a, b \in A$, we use $(a, b) \in R$ and $a \xrightarrow{R} b$ interchangeably. We use $R?$ and R^+ to denote the reflexive closure and the transitive closure of R, respectively. A relation $R \subseteq A \times A$ is *acyclic* if $R^+ \cap I_A = \emptyset$, where $I_A \triangleq \{(a, a) \mid a \in A\}$ is the identity relation on A. Given two

[1]A history collected from dynamically executing a system records the transactional requests to and responses from the database. See Section 2.2 for its formal definition.

[2]We discuss how to support SQL queries in Section 5.5. However, we do not support predicates in this work.

binary relations R and S over set A, we define their composition as $R\ ;\ S = \{(a,c) \mid \exists b \in A : a \xrightarrow{R} b \xrightarrow{S} c\}$. A strict partial order is an irreflexive and transitive relation. A strict total order is a relation which is a strict partial order and total.

For a directed labeled graph $G = (V, E)$, we use V_G and E_G to denote the set of vertices and the set of edges in G, respectively. For a set F of edges, we use $G|_F$ to denote the directed labeled graph that has the set V of vertices and the set F of edges.

In logical formulas, we write _ for components that are irrelevant and implicitly existentially quantified. We use $\exists!$ to mean "unique existence".

2.2.2 *Transactions and Histories.*

Definition 3. A ***transaction*** is a pair (O, po), where $O \subseteq \mathrm{Op}$ is a finite, non-empty set of operations and $\mathrm{po} \subseteq O \times O$ is a strict total order called the ***program order***.

For a transaction T, we let $T \vdash \mathsf{W}(x, v)$ if T writes to x and the last value written is v, and $T \vdash \mathsf{R}(x, v)$ if T reads from x before writing to it and v is the value returned by the first such read. We also use $\mathsf{WriteTx}_x = \{T \mid T \vdash \mathsf{W}(x, \_)\}$.

Clients interact with the store by issuing transactions via *sessions*. We use a *history* to record the client-visible results of such interactions. For conciseness, we consider only committed transactions in the formalism [13]; see further discussions in Section 4.5.

Definition 4. A ***history*** is a pair $\mathcal{H} = (\mathcal{T}, \mathrm{SO})$, where $\mathcal{T}$ is a set of transactions with disjoint sets of operations and the ***session order*** $\mathrm{SO} \subseteq \mathcal{T} \times \mathcal{T}$ is the union of strict total orders defined on disjoint sets of $\mathcal{T}$, which correspond to transactions in different sessions.

2.2.3 Dependency Graph-based Characterization of SI. A dependency graph extends a history with three relations (or typed edges in terms of graphs), i.e., WR, WW, and RW, representing one possibility of dependencies between transactions in this history [13]. The WR relation associates a transaction that reads some value with the one that writes this value. The WW relation stipulates a strict total order (aka the version order [1]) among the transactions on the same key. The RW relation is derived from WR and WW, relating a transaction that reads some value to the one that overwrites this value, in terms of the version orders specified by the WW relation.

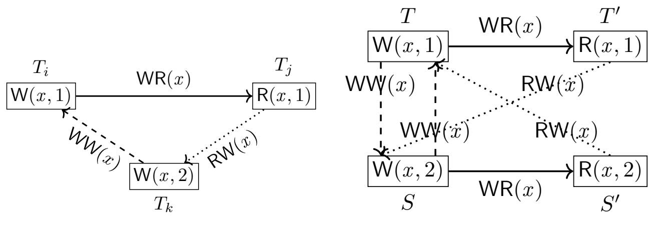
Definition 5. A ***dependency graph*** is a tuple $\mathcal{G} = (\mathcal{T}, \mathrm{SO}, \mathrm{WR}, \mathrm{WW}, \mathrm{RW})$, where $(\mathcal{T}, \mathrm{SO})$ is a history and

(1) $\mathrm{WR} : \mathsf{Key} \to 2^{\mathcal{T}\times\mathcal{T}}$ is such that
- $\forall x \in \mathsf{Key}.\ \forall S \in \mathcal{T}.\ S \vdash \mathsf{R}(x, \_) \implies \exists! T \in \mathcal{T}.\ T \xrightarrow{\mathrm{WR}(x)} S$.
- $\forall x \in \mathsf{Key}.\ \forall T, S \in \mathcal{T}.\ T \xrightarrow{\mathrm{WR}(x)} S \implies (T \neq S) \wedge (\exists v \in \mathsf{Val}.\ T \vdash \mathsf{W}(x, v) \wedge S \vdash \mathsf{R}(x, v))$

(2) $\mathrm{WW} : \mathsf{Key} \to 2^{\mathcal{T}\times\mathcal{T}}$ is such that for every $x \in \mathsf{Key}$, $\mathrm{WW}(x)$ is a strict total order on the set $\mathsf{WriteTx}_x$;

(3) $\mathrm{RW} : \mathsf{Key} \to 2^{\mathcal{T}\times\mathcal{T}}$ is such that $\forall T, S \in \mathcal{T}.\ \forall x \in \mathsf{Key}.\ T \xrightarrow{\mathrm{RW}(x)} S \iff T \neq S \wedge \exists T' \in \mathcal{T}.\ T' \xrightarrow{\mathrm{WR}(x)} T \wedge T' \xrightarrow{\mathrm{WW}(x)} S$.

We denote a component of $\mathcal{G}$, such as WW, by $\mathrm{WW}_{\mathcal{G}}$. We also write $T \xrightarrow{\mathrm{WR/WW/RW}} S$ when the key x in $T \xrightarrow{\mathrm{WR}(x)/\mathrm{WW}(x)/\mathrm{RW}(x)}$ is irrelevant or the context is clear.



(a) A polygraph **(b) A generalized polygraph**

Figure 2: Examples of polygraphs and generalized polygraphs. WR, WW, and RW relations are represented by solid arrows, dashed arrows, and dotted arrows, respectively.

Intuitively, a history satisfies SI if and only if it can be extended to a dependency graph that contains only cycles (if any) with at least two adjacent RW edges. Formally,

Theorem 6 (Dependency Graph-based Characterization of SI (Theorem 4.1 of [13])). *For a history* $\mathcal{H} = (\mathcal{T}, \mathrm{SO})$,

$$\mathcal{H} \models \mathrm{SI} \iff (\mathcal{H} \models \textsc{Int}) \wedge \big(\exists \mathrm{WR}, \mathrm{WW}, \mathrm{RW}.\ \mathcal{G} = (\mathcal{H}, \mathrm{WR}, \mathrm{WW}, \mathrm{RW}) \wedge (((\mathrm{SO}_{\mathcal{G}} \cup \mathrm{WR}_{\mathcal{G}} \cup \mathrm{WW}_{\mathcal{G}})\ ;\ \mathrm{RW}_{\mathcal{G}}?)\ \textit{is acyclic})\big).$$

The *internal consistency axiom* Int ensures that, within a transaction, a read from a key returns the same value as the last write to or read from this key in the transaction.

2.3 The SI Checking Problem

Definition 7. The ***SI checking problem*** is the decision problem of determining whether a given history $\mathcal{H}$ satisfies SI, i.e., is $\mathcal{H} \models \mathrm{SI}$.

We take the common "UniqueValue" assumption on histories [1, 8, 10, 20, 42]: for each key, every write to the key assigns a unique value. In implementations, this can be achieved by tagging values with unique identifiers, e.g., ETags in S3 [16] and timestamps in Cassandra [11]. Under this assumption, each read can be associated with the transaction that issues the corresponding write [13].

Theorem 6 provides a brute-force approach to the SI checking problem: enumerate all possible WW relations and check whether any of them results in a dependency graph that contains only cycles with at least two adjacent RW edges. However, this approach is prohibitively expensive. Actually the SI checking problem has been shown NP-complete in general [8].

2.4 Polygraphs

A dependency graph extending a history represents *one* possibility of dependencies between transactions in this history. To capture *all* possible dependencies between transactions in a single structure, we rely on polygraphs [36]. Intuitively, a polygraph can be viewed as a family of dependency graphs.

Definition 8. A ***polygraph*** $G = (V, E, C)$ associated with a history $\mathcal{H} = (\mathcal{T}, \mathrm{SO})$ is a directed labeled graph (V, E) called the *known graph*, together with a set C of *constraints* such that

- V corresponds to all transactions in the history $\mathcal{H}$;

- $E = \{(T, S, \mathsf{SO}) \mid T \xrightarrow{\mathsf{SO}} S\} \cup \{(T, S, \mathsf{WR}) \mid T \xrightarrow{\mathsf{WR}} S\}$, where SO and WR, when used as the third component of an edge, are edge labels (i.e., types); and
- $C = \{\langle (T_k, T_i, \mathsf{WW}), (T_j, T_k, \mathsf{RW})\rangle \mid (T_i \xrightarrow{\mathsf{WR}(x)} T_j) \land T_k \in \mathsf{WriteTx}_x \land T_k \neq T_i \land T_k \neq T_j\}$.

For a pair of transactions T_i and T_j such that $T_i \xrightarrow{\mathsf{WR}(x)} T_j$ and a third transaction T_k that also writes x, the constraint $\langle (T_k, T_i, \mathsf{WW}), (T_j, T_k, \mathsf{RW})\rangle$ captures the unknown dependencies of "either T_k happened before T_i or T_k happened after T_j"; see Figure 2a.

3 CHARACTERIZING SI USING GENERALIZED POLYGRAPHS

In this section we introduce generalized polygraphs with generalized constraints and use them to characterize SI. By compacting several constraints together, using generalized constraints leads to a compact encoding which in turn helps accelerate the solving process (see Section 5.4.3).

3.1 Generalized Polygraphs

In a polygraph, a constraint involves only a single pair of transactions related by WR, e.g., T_i and T_j on x in Figure 2a. Thus, several constraints are needed when there are multiple transactions reading the value of x from T_i. To *compact* these constraints, we introduce *generalized polygraphs* with generalized constraints.

Definition 9. A ***generalized polygraph*** $G = (V, E, C)$ associated with a history $\mathcal{H} = (\mathcal{T}, \mathsf{SO})$ is a directed labeled graph (V, E) called the *known graph*, together with a set C of *generalized constraints* such that

- V corresponds to all transactions in the history $\mathcal{H}$;
- $E \subseteq V \times V \times \mathcal{L}$ is a set of edges with labels (i.e., types) from $\mathcal{L} = \{\mathsf{SO}, \mathsf{WR}, \mathsf{WW}, \mathsf{RW}\}$; and
- $C = \Big\{\Big\langle \mathit{either} \triangleq \{(T, S, \mathsf{WW})\} \cup \bigcup_{T' \in \mathsf{WR}(x)(T)} \{(T', S, \mathsf{RW})\},\ \mathit{or} \triangleq \{(S, T, \mathsf{WW})\} \cup \bigcup_{S' \in \mathsf{WR}(x)(S)} \{(S', T, \mathsf{RW})\}\Big\rangle \mid T \in \mathsf{WriteTx}_x \land S \in \mathsf{WriteTx}_x \land T \neq S\Big\}$.

A generalized constraint is a pair of sets of edges of the form $\langle \mathit{either}, \mathit{or}\rangle$. The *either* part handles the possibility of T being ordered before S via a WW edge. This forces each transaction T' that reads the value of x from T to be ordered before S via a RW edge. Symmetrically, the *or* part handles the possibility of S being ordered before T via a WW edge. This forces each transaction S' that reads the value of x from S to be ordered before T via a RW edge.

Example 10 (Generalized Polygraphs vs. Polygraphs). In Figure 2b, both transactions T and S write to x, and T' and S' read the values of x from T and S, respectively. The possible dependencies between these transactions can be compactly expressed as a single generalized constraint $\langle \{(T, S, \mathsf{WW}), (T', S, \mathsf{RW})\}, \{(S, T, \mathsf{WW}), (S', T, \mathsf{RW})\}\rangle$, which corresponds to two constraints: $\langle (T, S, \mathsf{WW}), (S', T, \mathsf{RW})\rangle$ and $\langle (S, T, \mathsf{WW}), (T', S, \mathsf{RW})\rangle$.

Note that a generalized polygraph may contain edges of any type in E such that a "pruned" generalized polygraph (Section 4.3) is still a generalized polygraph. For a generalized polygraph $G = (V, E, C)$ and a label $L \in \mathcal{L}$, we use V_G, E_G, C_G, and L_G to denote the set V of vertices, the set E of known edges, the set C of constraints, and the set of known edges with label L in E, respectively. For a generalized polygraph $G = (V, E, C)$ and a set F of edges, we use $G|_F$ to denote the directed labeled graph that has the set V of vertices and the set F of edges. For any two subsets $R, S \subseteq E$, we define their composition as $R\ ;\ S = \{(a, c, L_1\ ;\ L_2) \mid \exists b \in V : (a, b, L_1) \in R \land (b, c, L_2) \in S\}$, where $L_1\ ;\ L_2$ is a newly introduced dummy label when composing edges.[3] In the sequel, we use generalized polygraphs, but sometimes we still refer to them as polygraphs.

3.2 Characterizing SI

According to Theorem 6, we are interested in the *induced* graph of a generalized polygraph G, obtained by composing the edges of G according to the rule $((\mathsf{SO} \cup \mathsf{WR} \cup \mathsf{WW})\ ;\ \mathsf{RW}?)$.

Definition 11. The ***induced SI graph*** of a polygraph $G = (V, E, C)$ is the graph $G' = (V, E, C, \mathcal{R})$, where $\mathcal{R} = (\mathsf{SO} \cup \mathsf{WR} \cup \mathsf{WW})\ ;\ \mathsf{RW}?$ is the induce rule.

The concept of *compatible* graphs gives a meaning to polygraphs and the induced SI graphs. A graph is compatible with a polygraph when it is a resolution of the constraints of the polygraph. Thus, a polygraph corresponds to a family of its compatible graphs.

Definition 12. A directed labeled graph $G' = (V', E')$ is ***compatible with a generalized polygraph*** $G = (V, E, C)$ if

- $V' = V$;
- $E' \supseteq E$; and
- $\forall \langle \mathit{either}, \mathit{or}\rangle \in C.\ (\mathit{either} \subseteq E' \land \mathit{or} \cap E' = \emptyset) \lor (\mathit{or} \subseteq E' \land \mathit{either} \cap E' = \emptyset)$.

By applying the induce rule $\mathcal{R}$ to a compatible graph of a polygraph, we obtain a compatible graph with the induced SI graph of this polygraph.

Definition 13. Let $G' = (V', E')$ be a compatible graph with a polygraph G. Then $G'|_{(\mathsf{SO}_{G'} \cup \mathsf{WR}_{G'} \cup \mathsf{WW}_{G'})\ ;\ \mathsf{RW}_{G'}?}$ is a ***compatible graph with the induced SI graph*** of G.

Example 14 (Compatible Graphs). There are two compatible graphs with the generalized polygraph of Figure 2b: one is with the edge set $\{(T, T', \mathsf{WR}), (S, S', \mathsf{WR}), (T, S, \mathsf{WW}), (T', S, \mathsf{RW})\}$, and the other is with $\{(T, T', \mathsf{WR}), (S, S', \mathsf{WR}), (S, T, \mathsf{WW}), (S', T, \mathsf{RW})\}$.

Accordingly, there are also two compatible graphs with the induced SI graph of the polygraph of Figure 2b: one is with the edge set $\{(T, T', \mathsf{WR}), (S, S', \mathsf{WR}), (T, S, \mathsf{WW}), (T, S, \mathsf{WR}\ ;\ \mathsf{RW})\}$. The edge $(T, S, \mathsf{WR}\ ;\ \mathsf{RW})$ is obtained from $(T, T', \mathsf{WR})\ ;\ (T', S, \mathsf{RW})$. It is duplicate with (T, S, WW) if the edge types are ignored. The other is with $\{(T, T', \mathsf{WR}), (S, S', \mathsf{WR}), (S, T, \mathsf{WW}), (S, T, \mathsf{WR}\ ;\ \mathsf{RW})\}$. Similarly, $(S, T, \mathsf{WR}\ ;\ \mathsf{RW})$ is duplicate with (S, T, WW) if the edge types are ignored.

We are concerned with the acyclicity of polygraphs and their induced SI graphs.

Definition 15. An induced SI graph is ***acyclic*** if there exists an acyclic compatible graph with it, *when the edge types are ignored.* A polygraph is ***SI-acyclic*** if its induced SI graph is acyclic.

[3]We will not use such labels in our formalism.

Finally, we present the generalized polygraph-based characterization of SI. Its proof can be found in [2, Appendix A]. The key lies in the correspondence between compatible graphs of polygraphs and dependency graphs.

Theorem 16 (Generalized Polygraph-based Characterization of SI). *A history $\mathcal{H}$ satisfies* SI *if and only if $\mathcal{H} \models$* Int *and the generalized polygraph of $\mathcal{H}$ is SI-acyclic.*

4 THE CHECKING ALGORITHM FOR SI

Given a history $\mathcal{H}$, PolySI encodes the induced SI graph of the generalized polygraph of $\mathcal{H}$ into an SMT formula and utilizes the MonoSAT solver [5] to test its acyclicity. We choose MonoSAT mainly because, compared to conventional SMT solvers such as Z3, it is more efficient in checking graph properties [5].

The main challenge for efficient MonoSAT solving is that the size (measured as the number of variables and clauses) of the resulting SMT formula may be too large for MonoSAT to solve in reasonable time. Hence, PolySI prunes constraints of the generalized polygraph of $\mathcal{H}$ before encoding. As we will see in Section 5.4, this pruning process is crucial to PolySI's high performance. Additionally, PolySI is accelerated in solving by utilizing, instead of original polygraphs, generalized polygraphs with compact generalized constraints.

4.1 Overview

The procedure CheckSI (line 1:65) outlines the checking algorithm. First, if $\mathcal{H}$ does not satisfy the Int axiom, the checking algorithm terminates and returns false (line 1:66; see Section 4.5 for the predicates AbortedReads and IntermediateReads). The algorithm proceeds otherwise in the following three steps:

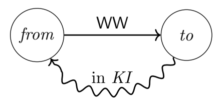
(1) construct the generalized polygraph G of $\mathcal{H}$ (lines 1:68 and 1:69; Section 4.2);
(2) prune constraints in the generalized polygraph G (line 1:70; Section 4.3); and
(3) encode the induced SI graph, denoted I, of G after pruning into an SMT formula (line 1:72), and call MonoSAT to test whether I is acyclic (line 1:73); see Section 4.4.

4.2 Constructing the Generalized Polygraph

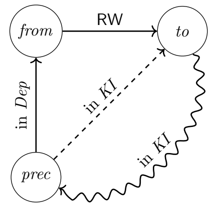
We construct the generalized polygraph G of the history $\mathcal{H}$ in two steps. First, we create the known graph of G by adding the known edges of types SO and WR to E_G (line 1:1). Second, we generate the generalized constraints of G on possible dependencies between transactions (line 1:6). Specifically, for each key x and each pair of transactions T and S that both write x, we generate a generalized constraint of the form $\langle either, or \rangle$ according to Definition 9 (lines 1:9 – 1:11).

4.3 Pruning Constraints

To accelerate MonoSAT solving, we prune as many constraints as possible before encoding (line 1:28). The key idea is to eliminate one of the two possibilities captured by a constraint and to insert the other remaining possibility into the known graph of the polygraph. A possibility can be eliminated only if we are sure that it cannot happen, i.e., it, together with other known edges, would create an undesired cycle. If both possibilities in a constraint are eliminated,



(a) (*from*, *to*) **is a WW edge.**



(b) (*from*, *to*) **is a RW edge.**

Figure 3: Two cases for pruning constraints.

PolySI immediately returns False. The pruning process repeats until no more constraints can be eliminated this way (line 1:63).

In each iteration, we first construct the *currently known part* of the induced SI graph, denoted *KI*, of *G*. That is, every compatible graph with the induced SI graph of *G* contains *KI* as a subgraph. We define two auxiliary graphs, namely $Dep \leftarrow G|_{\mathsf{SO}_G \cup \mathsf{WR}_G \cup \mathsf{WW}_G}$ and $AntiDep \leftarrow G|_{\mathsf{RW}_G}$. By Definition 13, *KI* is $Dep \cup (Dep \ ; \ AntiDep)$ (line 1:32). Then, we compute the reachability relation of *KI* using, e.g., the classic Floyd-Warshall algorithm [19, Section 25.2] (line 1:33). We next examine the current set C_G of constraints. For each constraint *cons* of the form $\langle either, or \rangle$, we check if any of the two possibilities of *either* and *or* can be eliminated (line 1:34). These two possibilities are similar, and in the following we explain in detail how the possibility of *either* is handled.

Consider an edge (*from*, *to*, *type*) in *either* (line 1:35). By construction, it must be of type WW or RW. Suppose first that (*from*, *to*) is a WW edge (line 1:36; see Figure 3(a)). If there is already a path from *to* to *from* in *KI* (line 1:37), adding the edge (*from*, *to*) would create a cycle in *KI*. Thus, we can eliminate this constraint *cons* from C_G and insert the edges corresponding to the other possibility *or* to the known graph E_G.

Now suppose that (*from*, *to*) is a RW edge; see Figure 3(b). Note that the edge (*from*, *to*, RW) does not belong to *KI*. Therefore, we cannot eliminate the constraint *cons* due only to a path from *to* to *from*. Instead, we check if there is a path in *KI* from *to* to any immediate predecessor *prec* of *from* in *Dep* (line 1:42). If so, by the construction of *KI*, there is an edge from *prec* to *to* in *KI*, which, together with the path from *to* to *prec*, creates a cycle in *KI*. Thus, the constraint *cons* can be resolved as discussed before.

The pruning process of the *or* possibility is same with that for *either* except that (1) We replace "*or*" with "*either*" at lines 53 and 61; and (2) It returns False if *cons* has already been pruned at lines 50 and 58. This means that both *either* and *or* possibilities cause undesired cycles and thus the input history does not satisfy SI.

Example 17 (Pruning Constraints). Consider the history in Figure 4. For key *x*, we generate the constraint $\langle either = \{(T_1, T_2, \mathsf{WW}), T_4, T_2, \mathsf{RW}\}, or = \{(T_2, T_1, \mathsf{WW})\} \rangle$; see Figure 4a. However, the *or* case could be pruned due to the cycle $T_2 \xrightarrow{\mathsf{WW}(x)} T_1 \xrightarrow{\mathsf{SO/WR}(y)} T_2$.

For key *y*, we generate the constraints $\langle either = \{(T_3, T_1, \mathsf{WW}), (T_4, T_1, \mathsf{RW})\}, or = \{(T_1, T_3, \mathsf{WW}), (T_2, T_3, \mathsf{RW})\} \rangle$. Consider the edge $T_4 \xrightarrow{\mathsf{RW}(y)} T_1$ in the *either* case; see Figure 4b. To decide whether it is in an undesired cycle, we first identify the immediate predecessors of T_4 along non-RW edges, i.e., T_1 and T_3. Since T_1 is trivially

Algorithm 1 The PolySI algorithm for checking SI

```
    $\mathcal{H} = (\mathcal{T}, \mathrm{SO})$: the history to check
    $G = (V, E, C)$: the polygraph, initially $(\mathcal{T}, \emptyset, \emptyset)$
    $I$: the induced SI graph, initially an empty graph
    BV: the set of Boolean variables, initially $\emptyset$
    CL: the set of clauses, initially $\emptyset$

 1: procedure CreateKnownGraph($\mathcal{H}$)
 2:     for all $T, S \in \mathcal{T}$ such that $T \xrightarrow{\mathrm{SO}} S$
 3:         $E_G \leftarrow E_G \cup (T, S, \mathrm{SO})$
 4:     for all $T, S \in \mathcal{T}$ such that $T \xrightarrow{\mathrm{WR}} S$
 5:         $E_G \leftarrow E_G \cup (T, S, \mathrm{WR})$

 6: procedure GenerateConstraints($\mathcal{H}$)
 7:     for all $x \in \mathsf{Key}$
 8:         for all $T, S \in \mathsf{WriteTx}_x$ such that $T \neq S$
 9:             $either \leftarrow \{(T, S, \mathrm{WW})\} \cup \bigcup_{T' \in \mathrm{WR}(x)(T)} \{(T', S, \mathrm{RW})\}$
10:             $or \leftarrow \{(S, T, \mathrm{WW})\} \cup \bigcup_{S' \in \mathrm{WR}(x)(S)} \{(S', T, \mathrm{RW})\}$
11:             $C_G \leftarrow C_G \cup \{\langle either, or \rangle\}$

12: procedure SMT-Encode()
13:     for all $v_i, v_j \in V_G$ such that $i \neq j$
14:         Create two Boolean variables $\mathsf{BV}_{i,j}$ and $\mathsf{BV}^I_{i,j}$
15:         $\mathsf{BV} \leftarrow \mathsf{BV} \cup \{\mathsf{BV}_{i,j}, \mathsf{BV}^I_{i,j}\}$
16:     for all $(v_i, v_j) \in E_G$                              ▷ encode the known graph
17:         $\mathsf{CL} \leftarrow \mathsf{CL} \cup \{\mathsf{BV}_{i,j} = \mathsf{True}\}$
18:     for all $\langle either, or \rangle \in C_G$                     ▷ encode the constraints
19:         $\mathsf{CL} \leftarrow \mathsf{CL} \cup \Big\{\big(\bigwedge_{(v_i,v_j,\_) \in either} \mathsf{BV}_{i,j} \land \bigwedge_{(v_i,v_j,\_) \in or} \neg\mathsf{BV}_{i,j}\big) \lor$
            $\big(\bigwedge_{(v_i,v_j,\_) \in or} \mathsf{BV}_{i,j} \land \bigwedge_{(v_i,v_j,\_) \in either} \neg\mathsf{BV}_{i,j}\big)\Big\}$
20:     $Dep \leftarrow G|_{\mathrm{SO}_G \cup \mathrm{WR}_G \cup \mathrm{WW}_G}$
21:     $E_{Dep} \leftarrow E_{Dep} \cup \{(\_, \_, \mathrm{WW}) \in either \cup or \mid \langle either, or \rangle \in C_G\}$
22:     $AntiDep \leftarrow G|_{\mathrm{RW}_G}$
23:     $E_{AntiDep} \leftarrow E_{AntiDep} \cup \{(\_, \_, \mathrm{RW}) \in either \cup or \mid \langle either, or \rangle \in C_G\}$
24:     $\mathsf{CL} \leftarrow \mathsf{CL} \cup \Big\{\mathsf{BV}^I_{i,j} = \big(\mathsf{BV}_{i,j} \land (v_i, v_j, \_) \in E_{Dep}\big) \lor \big(\bigvee_{\substack{(v_i,v_k,\_) \in E_{Dep} \\ (v_k,v_j,\_) \in E_{AntiDep}}} \mathsf{BV}_{i,k} \land$
        $\mathsf{BV}_{k,j}\big) \mid v_i, v_j \in V_G\Big\}$                ▷ encode the induced SI graph $I$ of $G$

25: procedure MonoSAT-Solve()
26:     $solver \leftarrow$ MonoSAT-Solver(BV, CL)
27:     return Solve($solver$, $I$ is acyclic)

28: procedure PruneConstraints()
29:     repeat
30:         $Dep \leftarrow G|_{\mathrm{SO}_G \cup \mathrm{WR}_G \cup \mathrm{WW}_G}$
31:         $AntiDep \leftarrow G|_{\mathrm{RW}_G}$
32:         $KI \leftarrow Dep \cup (Dep \,;\, AntiDep)$
33:         $reachability \leftarrow$ Reachability($KI$)
34:         for all $cons \leftarrow \langle either, or \rangle \in C_G$
35:             for all $(from, to, type) \in either$          ▷ for the "either" possibility
36:                 if $type = \mathrm{WW}$
37:                     if $(to, from) \in reachability$
38:                         $C_G \leftarrow C_G \setminus \{cons\}$
39:                         $E_G \leftarrow E_G \cup or$
40:                         break the "for all $(from, to, type) \in either$" loop
41:                 else                                                ▷ $type = \mathrm{RW}$
42:                     for all $prec \in V_{Dep}$ such that $(prec, from, \_) \in E_{Dep}$
43:                         if $(to, prec) \in reachability$
44:                             $C_G \leftarrow C_G \setminus \{cons\}$
45:                             $E_G \leftarrow E_G \cup or$
46:                             break the "for all $(from, to, type) \in either$" loop
47:             for all $(from, to, type) \in or$              ▷ for the "or" possibility
48:                 if $type = \mathrm{WW}$
49:                     if $(to, from) \in reachability$
50:                         if $cons \notin C_G$        ▷ neither "either" nor "or" is possible
51:                             return False
52:                         $C_G \leftarrow C_G \setminus \{cons\}$
53:                         $E_G \leftarrow E_G \cup either$
54:                         break the "for all $(from, to, type) \in or$" loop
55:                 else                                                ▷ $type = \mathrm{RW}$
56:                     for all $prec \in V_{Dep}$ such that $(prec, from, \_) \in E_{Dep}$
57:                         if $(to, prec) \in reachability$
58:                             if $cons \notin C_G$  ▷ neither "either" nor "or" is possible
59:                                 return False
60:                             $C_G \leftarrow C_G \setminus \{cons\}$
61:                             $E_G \leftarrow E_G \cup either$
62:                             break the "for all $(from, to, type) \in or$" loop
63:     until $C_G$ remains unchanged
64:     return True

65: procedure CheckSI($\mathcal{H}$)
66:     if $\mathcal{H} \not\models$ Int $\lor$ AbortedReads $\lor$ IntermediateReads
67:         return false
68:     CreateKnownGraph($\mathcal{H}$)
69:     GenerateConstraints($\mathcal{H}$)
70:     if $\neg$PruneConstraints()
71:         return false
72:     SMT-Encode()
73:     return MonoSAT-Solve()
```

reachable to itself, we find a cycle $T_4 \xrightarrow{\mathrm{RW}(y)} T_1 \xrightarrow{\mathrm{WR}(x)} T_4$. Hence, the *either* case could be pruned.

We next check the *or* case; see Figure 4c. When checking the edge $T_2 \xrightarrow{\mathrm{RW}(y)} T_3$, T_1 the only immediate predecessor of T_2 along a non-RW edge. However, T_1 is unreachable from T_3. Therefore, the *or* case is kept. Note that the cycle $T_4 \xrightarrow{\mathrm{RW}(x)} T_2 \xrightarrow{\mathrm{RW}(y)} T_3 \xrightarrow{\mathrm{WR}(y)} T_4$ that contains two adjacent RW edges is allowed by SI.

The following theorem shows that PruneConstraints is correct in that (1) it preserves the SI-(a)cyclicity of generalized polygraphs; and (2) it does not introduce new undesired cycles, which ensures that any violation found in the pruned polygraph using MonoSAT later also exists in the original polygraph (and thus in the history). This property is crucial to the informativeness of PolySI. The proof of the theorem can be found in [2, Appendix A].

Theorem 18 (Correctness of PruneConstraints). *Let G and G_p be the generalized polygraphs before and after* PruneConstraints*, respectively. Then,*

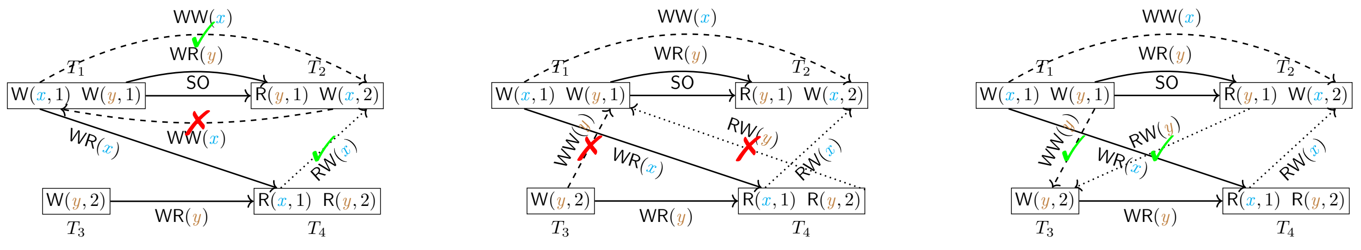
(1) G is SI-acyclic if and only if PruneConstraints *returns* True *and G_p is SI-acyclic.*
(2) Suppose that G_p is not SI-acyclic. Let C be a cycle in a compatible graph with the induced SI graph of G_p. Then there is a compatible graph with the induced SI graph of G that contains C.

Combining Theorems 16 and 18, we prove PolySI's soundness.

Theorem 19 (Soundness of PolySI). PolySI *is sound, i.e., if* PolySI *returns* False*, then the input history indeed violates SI.*

4.4 SMT Encoding

In this step we encode the induced SI graph, denoted I, of the pruned generalized polygraph G into an SMT formula. MonoSAT will try to find an acyclic compatible graph with I. We use BV and



(a) $T_2 \xrightarrow{\textbf{WW}(x)} T_1$ **is pruned due to the cycle** $T_2 \xrightarrow{\textbf{WW}(x)} T_1 \xrightarrow{\textbf{SO/WR}(y)} T_2$.

(b) $T_3 \xrightarrow{\textbf{WW}(y)} T_1$ **is pruned due to the cycle** $T_4 \xrightarrow{\textbf{RW}(y)} T_1 \xrightarrow{\textbf{WR}(x)} T_4$.

(c) $T_1 \xrightarrow{\textbf{WW}(y)} T_3$ **stays as cycle** $T_4 \xrightarrow{\textbf{RW}(x)} T_2 \xrightarrow{\textbf{RW}(y)} T_3 \xrightarrow{\textbf{WR}(y)} T_4$ **is allowed by SI.**

Figure 4: Illustration of pruning constraints. SO and WR relations are represented by solid arrows. WW and RW relations are indicated by dashed arrows and dotted arrows, respectively.

CL to denote the set of Boolean variables and the set of clauses of the SMT formula, respectively. For each pair of vertices v_i and v_j, we create two Boolean variables $\mathsf{BV}_{i,j}$ and $\mathsf{BV}^I_{i,j}$: one for the generalized polygraph G, and the other for the induced SI graph I of G (line 1:14). The edge (v_i, v_j) is included in the resulting compatible graph with I (resp., with G) if and only if $\mathsf{BV}^I_{i,j}$ (resp., $\mathsf{BV}_{i,j}$) is assigned to True by MonoSAT.

We first encode the generalized polygraph G. For each edge (v_i, v_j) in the known graph of G, we add a clause $\mathsf{BV}_{i,j} = \mathsf{True}$ to CL (line 1:17). For each constraint $\langle \mathit{either}, \mathit{or} \rangle$, the clause

$$\Big(\bigwedge_{(v_i,v_j,\_)\in \mathit{either}} \mathsf{BV}_{i,j} \wedge \bigwedge_{(v_i,v_j,\_)\in \mathit{or}} \neg\mathsf{BV}_{i,j} \Big) \vee$$
$$\Big(\bigwedge_{(v_i,v_j,\_)\in \mathit{or}} \mathsf{BV}_{i,j} \wedge \bigwedge_{(v_i,v_j,\_)\in \mathit{either}} \neg\mathsf{BV}_{i,j} \Big)$$

expresses that exactly one of *either* or *or* happens (line 1:19).

Now we encode the induced SI graph I of G. The auxiliary graph *Dep* contains all the known and potential SO, WR, and WW edges of G (lines 1:20 and 1:21), while *AntiDep* contains all the known and potential RW edges of G (lines 1:22 and 1:23). The clauses on BV^I at line 1:24 state that I is the union of graph *Dep* and the composition of graph *Dep* and graph *AntiDep*.

Finally, we feed the SMT formula to MonoSAT for an acyclicity test of the graph I (line 1:27).

4.5 Completing the SI Checking

Theorem 6 assumes histories with only committed transactions and considers the WR, WW, and RW relations over transactions rather than read/write operations inside them. This would miss non-cycle anomalies. Hence, for completeness, PolySI also checks whether a history violates the following two properties [1, 30], indicated by the two predicates AbortedReads and IntermediateReads in line 1:66, respectively.

- *Aborted Reads*: a committed transaction cannot read a value from an aborted transaction.
- *Intermediate Reads*: a transaction cannot read a value that was overwritten by the transaction that wrote it.

Note that PolySI's completeness relies on a common assumption about *determinate* transactions [1, 8, 12, 13, 20, 30], i.e., the status of each transaction, whether committed or aborted, is legitimately decided. Indeterminate transactions are inherent to black-box testing: it is difficult for a client to justify the status of a transaction due to the invisibility of system internals. Together with the completeness of the dependency-graph-based characterization of SI in Theorem 6, we prove PolySI's completeness.

Theorem 20 (Completeness of PolySI). PolySI *is complete with respect to a history that contains only determinate transactions, i.e., if such a history indeed violates* SI, PolySI *returns* false.

5 EXPERIMENTS

We have presented our SI checking algorithm PolySI and established its *soundness* and *completeness*. In this section, we conduct a comprehensive assessment of PolySI to answer the following questions with respect to the remaining criteria of SIEGE+ (Section 1):

(1) Effective: Can PolySI find SI violations in (production) databases?

(2) Informative: Can PolySI provide understandable counterexamples for SI violations?

(3) Efficient: How efficient is PolySI (and its components)? Can PolySI outperform the state of the art under *various* workloads?

Our answer to (1) is twofold (Section 5.2): (i) PolySI successfully reproduces all of 2477 known SI anomalies in production databases; and (ii) we use PolySI to detect novel SI violations in three cloud databases of different kinds, namely, the graph database Dgraph [24], the relational database MariaDB-Galera [17], and YugabyteDB [49] supporting multiple data models. To answer (2) we provide an algorithm that recovers the violating scenario, highlighting the cause of the violation found (Section 5.3). Regarding (3), we (i) show that PolySI outperforms several competitive baselines including the most performant SI and serializability checkers to date and (ii) measure the contributions of its different components/optimizations to the overall performance under both general and specific transaction workloads (Section 5.4). Note that we demonstrate PolySI's **generality** along with the answers to questions (1) and (3).

5.1 Workloads, Benchmarks, and Setup

5.1.1 Workloads and Benchmarks. To evaluate PolySI on *general* read-only, write-only, and read-write transaction workloads, we have implemented a parametric workload generator. The parameters are: the number of client sessions (#sess), the number of transactions per session (#txns/sess), the number of read/write operations per transaction (#ops/txn), the percentage of reads (%reads), the total number of keys (#keys), and the key-access distribution (dist) including uniform, zipfian, and hotspot (80% operations touching 20% keys). Table 1 shows the default value of each parameter in our experiments; in particular, the default 2k transactions with 30k operations issued by 20 sessions are sufficient to distinguish PolySI from competing tools (see Section 5.4.1).

Table 1: Workload parameters and their default values.

Parameter	Default Value	Parameter	Default Value
#sess	20	#keys	10k
#txns/sess	100	%reads	50%
#ops/txn	15	dist	uniform

Among such general workloads, we also consider three representatives, each with 10k transactions and 80k operations in total (#sess=25, #txns/sess=400, and #ops/txn=8), in the comparison with Cobra and the decomposition and differential analysis of PolySI:

- GeneralRH: read-heavy workload with 95% reads;
- GeneralRW: medium workload with 50% reads; and
- GeneralWH: write-heavy workloads with 30% reads.

We also use three synthetic benchmarks with only serializable histories of at least 10k transactions (which also satisfy SI):

- RUBiS [39]: an eBay-like bidding system where users can, for example, register and bid for items. The dataset archived by [42] contains 20k users and 200k items.
- TPC-C [47]: an open standard for benchmarking online transaction processing with a mix of five different types of transactions (e.g., for orders and payment) portraying the activity of a wholesale supplier. The dataset includes one warehouse, 10 districts, and 30k customers.
- C-Twitter [29]: a Twitter clone where users can, for example, tweet and follow or unfollow other users (following the zipfian distribution).

5.1.2 Setup. We use a PostgreSQL (v15 Beta 1) instance to produce *valid* histories without isolation violations: for the performance comparison with other SI checkers and the decomposition and differential analysis of PolySI itself, we set the isolation level to *repeatable read* (implemented as SI in PostgreSQL [38]); for the runtime comparison with Cobra (Section 5.4.1), we use the *serializability* isolation level to produce serializable histories. We co-locate the client threads and PostgreSQL (or other databases for testing; see Section 5.2.2) on a local machine. Each client thread issues a stream of transactions produced by our workload generator to the database and records the execution history. All histories are saved to a file to benchmark each tool's performance.

We have implemented PolySI in 2.3k lines of Java code, and the workload generator, including the transformation from generated key-value operations to SQL queries (for the interactions with relational databases such as PostgreSQL), in 2.2k lines of Rust code. We ensure unique values written for each key using counters. We use

Table 2: Summary of tested databases. Multi-model refers to relational DBMS, document store, and wide-column store.

Database	GitHub Stars	Kind	Release
New violations found:			
Dgraph	18.2k	Graph	v21.12.0
MariaDB-Galera	4.4k	Relational	v10.7.3
YugabyteDB	6.7k	Multi-model	v2.11.1.0
Known bugs [8, 21, 28]:			
CockroachDB	25.1k	Relational	v2.1.0
			v2.1.6
MySQL-Galera	381	Relational	v25.3.26
YugabyteDB	6.7k	Multi-model	v1.1.10.0

a simple database schema of a two-column table storing keys and values, which is effective to find real violations in three production databases (see Section 5.2).

We conducted all experiments on a machine with a 4.5GHz Intel i7-9700H (6-core) CPU, 16GB memory, and an NVIDIA 1650 GPU.

5.2 Finding SI Violations

5.2.1 Reproducing Known SI Violations. PolySI successfully reproduces *all* known SI violations in an extensive collection of 2477 anomalous histories [8, 21, 28]. These histories were obtained from the earlier releases of three different production databases, i.e., CockroachDB, MySQL-Galera, and YugabyteDB; see Table 2 for details. This set of experiments provides supporting evidence for PolySI's *soundness* and *completeness*, established in Section 4.

5.2.2 Detecting New Violations. We use PolySI to examine recent releases of three well-known cloud databases (of different kinds) that claim to provide SI: Dgraph [24], MariaDB-Galera [17], and YugabyteDB [49]. See Table 2 for details. We have found and reported novel SI violations in all three databases which, as of the time of writing, are being investigated by the developers. In particular, as communicated with the developers, (i) our finding has helped the DGraph team confirm some of their suspicions about their latest release; and (ii) Galera has confirmed the incorrect claim on preventing lost updates for transactions issued on different cluster nodes and thereafter removed any claims on SI or "partially supporting SI" from the previous documentation.[4]

5.3 Understanding Violations

MonoSAT reports cycles, constructed from its output logs, upon detecting an SI violation. However, such cycles are *uninformative* with respect to understanding how the violation actually occurred. For instance, Figure 5(a) and Figure 6(a) depict the original cycles returned by MonoSAT for an SI violation found in MariaDB-Galera and Dgraph, respectively, where it is difficult to identify the cause of each violation.

Hence, we have designed an algorithm to interpret the returned cycles. The key idea is to (i) bring back any potentially involved transactions and the associated dependencies, (ii) restore the violating scenario by identifying the core participants and dependencies,

[4]https://github.com/codership/documentation/commit/cc8d6125f1767493eb61e2cc82f5a365ecee6e7a and https://github.com/codership/documentation/commit/d87171b0d1b510fe59973cb7ce5892061ce67b80

and (iii) remove the "irrelevant" dependencies to simplify the scenario. We have integrated into PolySI the algorithm written in 300 lines of C++ code. The pseudocode is given in [2, Appendix B].

In the following, we present two example violations found in MariaDB-Galera and Dgraph, respectively (Section 5.2.2). In particular, we illustrate how the interpretation algorithm helps us locate the violation causes: *lost update* in MariaDB-Galera and *causality violation* in Dgraph. We defer the YugabyteDB anomaly (also a causality violation) to [2, Appendix C]. In the examples, we use T:(s, n) to denote the nth transaction issued by session s.

5.3.1 Lost Update in MariaDB-Galera. Given the original cycles returned by MonoSAT in Figure 5(a), PolySI first finds the (only) "missing" transaction T:(1,4) (colored in green) and the associated dependencies, as shown in Figure 5(b). Note that some of the dependencies are uncertain at this moment, e.g., the WW dependency between T:(1,4) and T:(1,5) (in red). PolySI then restores the violating scenario by resolving such uncertainties. For example, as depicted in Figure 5(c), PolySI determines that W(0,4) was actually installed first in the database, i.e., T:(1,4)$\xrightarrow{\text{WW}}$T:(1,5), because there would otherwise be an undesired cycle with the known dependencies, i.e., T:(1,5)$\xrightarrow{\text{WW}}$T:(1,4)$\xrightarrow{\text{WR}}$T:(1,5). The same reasoning applies to determine the WW dependency between T:(1,4) and T:(2,13) (in blue). Finally, PolySI finalizes the violating scenario by removing any remaining uncertainties including those dependencies not involved in the actual violation (the WW dependency between T:(1,5) and T:(2,13) in this case).

The violating scenario now becomes informative and explainable: transaction T:(1,4) writes value 4 on key 0, which is read by transactions T:(2,13) and T:(1,5). Both transactions subsequently commit their writes on key 0 by W(0,13) and W(0,5), respectively, which results in a *lost update* anomaly.

5.3.2 Causality Violation in Dgraph. There might be multiple "missing" transactions that together contribute to a violation. As depicted in Figure 6(b), PolySI restores the potentially involved five transactions (in green) and the associated dependencies from the original output cycle in Figure 6(a). In particular, two sub-scenarios are involved: the left subgraph concerns the RW(656) dependency on key 656 (in blue) while the right subgraph concerns the dependency WW(402) on key 402 (in red). Following the same procedure as in the MariaDB-Galera example, PolySI resolves the outstanding dependencies if no undesired cycles arise; see Figure 6(c). For example, we have T:(4,172)$\xrightarrow{\text{WW}}$T:(10,471) (in blue) as there would otherwise be a cycle T:(10,471)$\xrightarrow{\text{WW}}$T:(4,172)$\xrightarrow{\text{WR}}$T:(10,467)$\xrightarrow{\text{SO}}$T:(10,471). The finalized violating scenario is shown in Figure 6(d) where two impossible (dashed) dependencies in Figure 6(c) have been eliminated.

This violation occurs as the causality order is not preserved, which is a happens-before relationship between any two transactions in a given history [31, 33].[5] More specifically, transaction T:(9,428) causally depends on transaction T:(10,471) (via the counterclockwise path in the right subgraph) and transaction T:(10,471) causally depends on transaction T:(4,172) (via the clockwise path in the left subgraph). Hence, transaction T:(9,428) should have fetched the value 7 of key 656 written by transaction T:(10,471), instead of the value 3 of transaction T:(4,172), to respect the causality order.

[5]Intuitively, transaction T causally depends on transaction S if any of the following conditions holds: (i) T and S are issued in the same session and S is executed before T; (ii) T reads the value written by S; and (iii) there exists another transaction R such that T causally depends on R which in turn causally depends on S.

5.4 Performance Evaluation

In this section, we conduct an in-depth performance analysis of PolySI and compare it to the following black-box checkers:

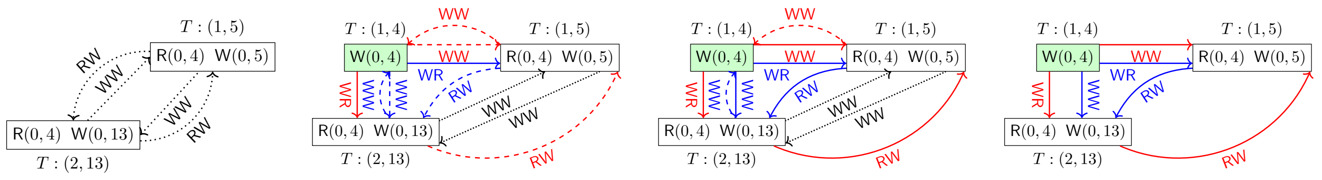
- dbcop [8] is, to the best of our knowledge, the most efficient black-box SI checker that does not use an off-the-shelf solver. Note that, unlike our PolySI tool, dbcop does not check *aborted reads* or *intermediate reads* (see Section 4.5).
- Cobra [42] is the state-of-the-art SER checker utilizing both MonoSAT and GPUs to accelerate the checking procedure. Cobra serves as a baseline because (i) checking SI is more complicated than checking SER in general [8], and constraint pruning and the MonoSAT encoding for SI are more challenging in particular due to more complex cycle patterns in dependency graphs (Theorem 6, Section 2.2.3); and (ii) Cobra is the most performant SER checker to date.
- CobraSI: We implement the incremental algorithm [8, Section 4.3] for reducing checking SI to checking serializability (in polynomial time) to leverage Cobra. We consider two variants: CobraSI without GPU for a fair comparison with other SI checkers (i.e., PolySI and dbcop) and CobraSI with GPU as a strong competitor.

5.4.1 Performance Comparison with State of the Art. Our first set of experiments compares PolySI with the competing SI checkers under a wide range of workloads. The input histories extracted from PostgreSQL (with the *repeatable read* isolation level) are all valid with respect to SI. The experimental results are shown in Figure 7: PolySI significantly surpasses not only the state-of-the-art SI checker dbcop but also CobraSI with GPU. In particular, with more concurrency, such as more sessions (a), transactions per session (b), and operations per transaction (c), CobraSI with GPU exhibits exponentially increasing checking time while PolySI incurs only moderate overhead. The result depicted in Figure 7(f) is also consistent: with the skewed key accesses representing high concurrency as in the zipfian and hotspot distributions, both dbcop and CobraSI without GPU time out. Moreover, even with the GPU acceleration, CobraSI takes 6x more time than PolySI. Finally, unlike the other SI checkers, PolySI's performance is fairly stable with respect to varying read/write proportions (d) and keys (e).

In Figure 8 we compare PolySI with the baseline serializability checker Cobra. We present the checking time on various benchmarks. PolySI outperforms Cobra (with its GPU acceleration enabled) in five of the six benchmarks with up to 3x improvement (as for GeneralRH). The only exception is TPC-C, where most of the transactions have the read-modify-write pattern,[6] for which Cobra implements a specific optimization to efficiently infer dependencies.

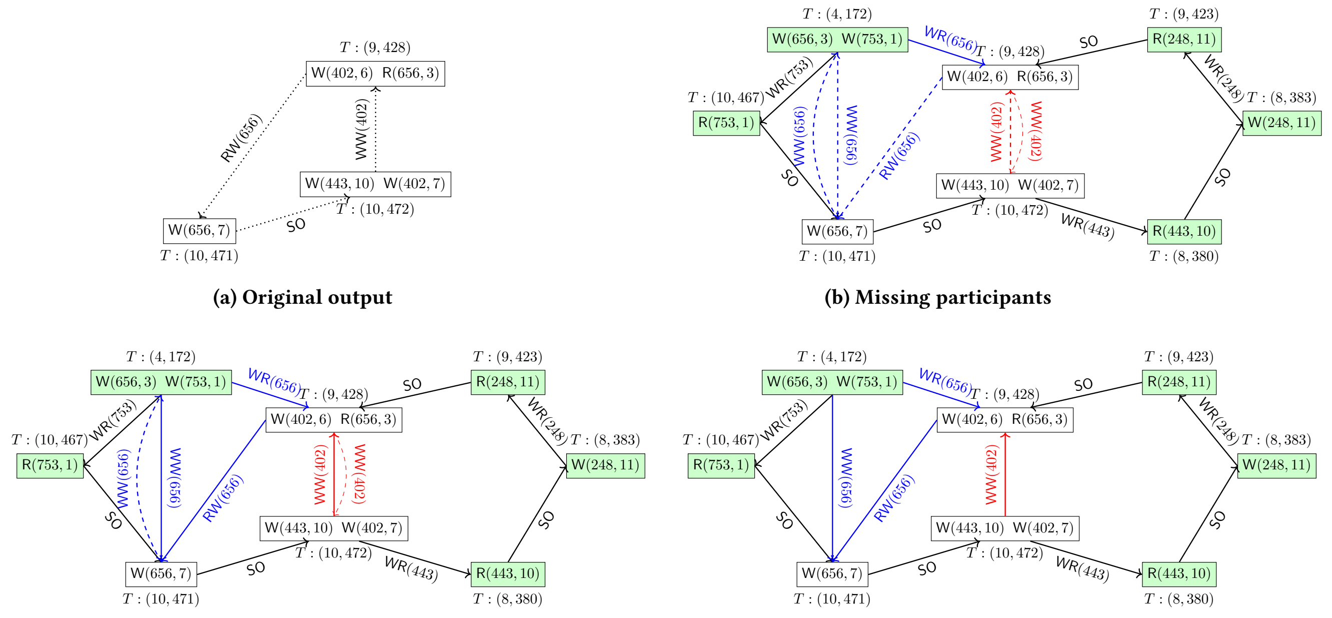
5.4.2 Decomposition Analysis of PolySI. We measure PolySI's checking time in terms of stages: *constructing*, which builds up a generalized polygraph from a given history; *pruning*, which prunes constraints in the generalized polygraph; *encoding*, which encodes

[6]In a read-modify-write transaction each read is followed by a write on the same key.



(a) Original output **(b) Missing participants** **(c) Recovered scenario** **(d) Finalized scenario**

Figure 5: Lost update: the SI violation found in MariaDB-Galera. The original output dependencies are indicated by dotted black arrows. The recovered dependencies are colored in red/blue with dashed and solid arrows representing uncertain and certain dependencies, respectively. The (only) missing transaction is colored in green. We omit key 0, associated with all dependencies.



(a) Original output **(b) Missing participants**

(c) Recovered scenario **(d) Finalized scenario**

Figure 6: Causality violation: the SI anomaly found in Dgraph. Dashed and solid arrows represent uncertain and certain dependencies, respectively. Recovered transactions are colored in green. The core dependencies involved in the two sub-scenarios are colored in red and blue, respectively.

the graph and the remaining constraints; and *solving*, which runs the MonoSAT solver.

Figure 9 depicts the results on six different datasets. Constructing a generalized polygraph is relatively inexpensive. The overhead of pruning is fairly constant, regardless of the workloads; PolySI can effectively prune (resp. resolve) a huge number of constraints (resp. unknown dependencies) in this phase, e.g., for TPC-C, all constraints are pruned and all uncertainties are resolved. See Table 3 for details. The encoding effort is moderate; TPC-C incurs more overhead as the number of operations in total is 5x more than the others. The solving time depends on the remaining constraints and unknown dependencies after pruning, e.g., the left four datasets incur negligible overhead (see Table 3).

5.4.3 Differential Analysis of PolySI. To investigate the contributions of PolySI's two major optimizations, we experiment with three variants: (i) PolySI itself; (ii) PolySI without pruning (P) constraints;

Table 3: Number of constraints and unknown dependencies before and after pruning (P) in the six benchmarks.

Benchmark	#cons. before P	#cons. after P	#unk. dep. before P	#unk. dep. after P
TPC-C	386k	0	3628k	0
GeneralRH	4k	29	39k	77
RUBiS	14k	149	171k	839
C-Twitter	59k	277	307k	776
GeneralRW	90k	2565	401k	5435
GeneralWH	167k	6962	468k	14376

and (iii) PolySI without both compacting (C) and pruning the constraints. Figure 10 demonstrates the acceleration produced by each optimization. Note that the two variants without optimization exhibit (16GB) memory-exhausted runs on TPC-C, which contain

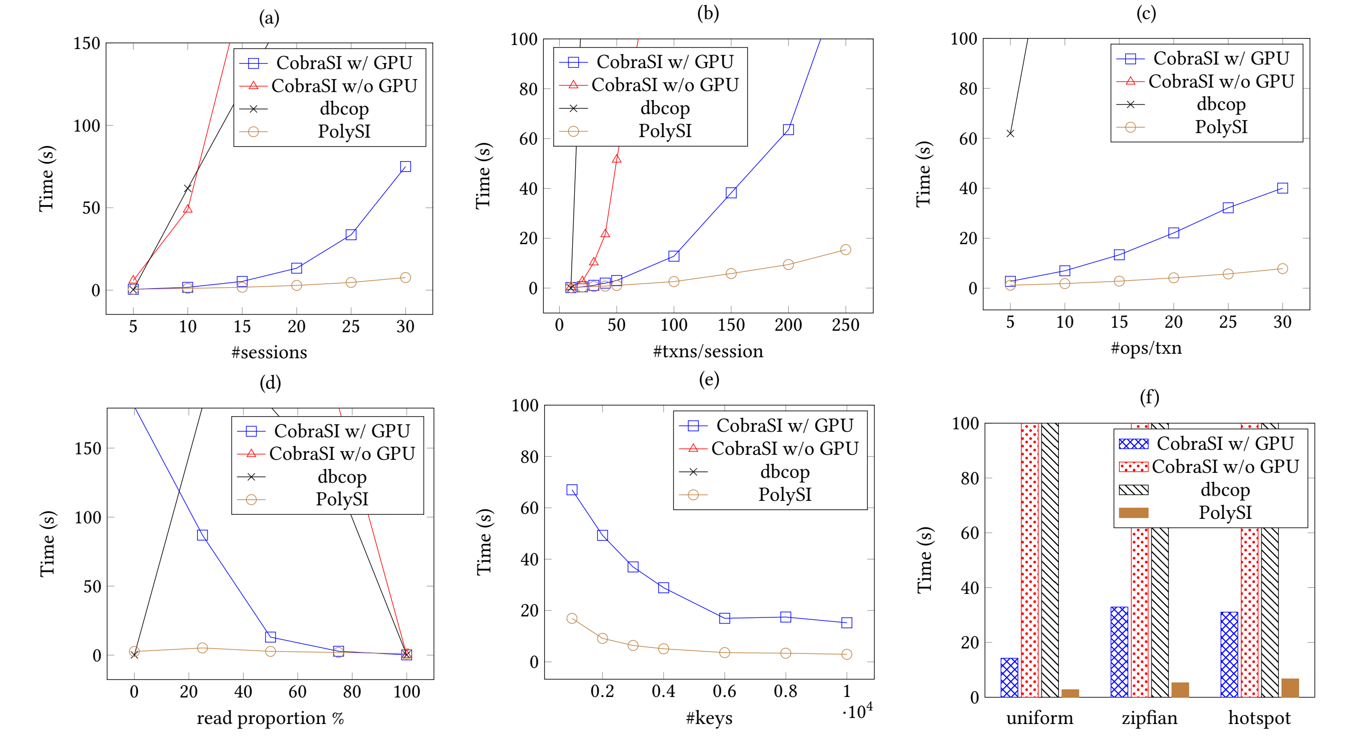


Figure 7: Performance comparison with the competing SI checkers under various workloads. Experiments time out at 180s; data points are not plotted for timed-out experiments.

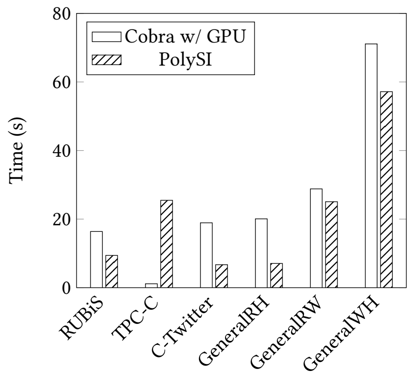


Figure 8: PolySI vs. Cobra w/ GPU on serializable histories.

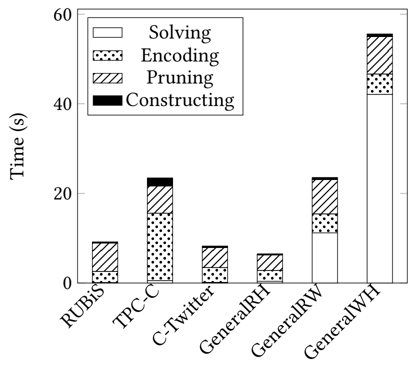


Figure 9: Decomposing PolySI's checking time into stages.

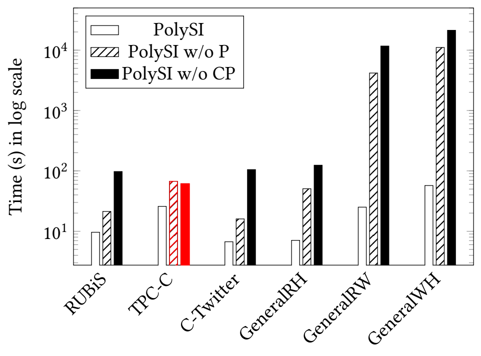


Figure 10: Differential analysis of PolySI. Memory-exhausted runs are colored in red.

considerably more uncertain dependencies (3628k) and constraints (386k) without pruning than the other datasets (see Table 3).

5.5 Discussion

Fault Injection. We have found SI violations in three production databases without injecting faults, such as network partition and clock drift. Since PolySI is an off-the-shelf checker, it is straightforward to integrate it into existing testing frameworks with fault injection such as Jepsen [27] and CoFI [14]; both have been demonstrated to effectively trigger bugs in distributed systems.

Database Schema. In our testing of production databases, we used a simple, yet effective, database schema adopted by most of black-box checkers [8, 42, 50, 51]: a two-column table storing key-value pairs. Extending it to multi-columns or even the column-family data model could be done by: (i) representing each cell in a table as a compound key, i.e., "TableName:PrimaryKey:ColumnName", and a single value, i.e., the content of the cell; and (ii) utilizing the compiler in [9] to rewrite (more complex) SQL queries to key-value read/write operations.

Predicates. To the best of our knowledge, none of the state-of-the-art black-box checkers [8, 9, 30, 42, 50, 51] considers predicates nor can they detect predicate-specific anomalies. Given the non-predicate violations found by PolySI (as well as dbcop [8] and Elle [30]), we conjecture that more anomalies would arise with predicates. It is therefore interesting future work to extend our SI characterization to represent predicates and to explore optimizations with respect to encoding and pruning.

6 RELATED WORK

6.1 Characterizing Snapshot Isolation

Many frameworks and formalisms have been developed to characterize SI and its variants. Berenson et al. [6] considers SI as a multi-version concurrency control mechanism (described also in Section 2.1). Adya [1] presents the first formal definition of SI using dependency graphs, which, as pointed out by [13], still relies on low-level implementation choices such as how to order start and commit events in transactions. Cerone et al. [12] proposes an axiomatic framework to declaratively define SI with the dual notions of visibility (what transactions can observe) and arbitration (the order of installed versions/values). The follow-up work [13] characterizes SI solely in terms of Adya's dependency graphs requiring no additional information about transactions. Crooks et al. [20] introduces an alternative implementation-agnostic formalization of SI and its variants based on client observations (of key-value pairs read/written). Xiong et al. [48] provides an operational definition of SI, together with an operational semantics for client-observable behaviours of distributed transactions over key-value stores.

Driven by black-box testing of SI, we base our GP-based characterization on Cerone and Gotsman's formal specification [13]. In particular, our new characterization:

1. targets the prevalent *strong session* variant of SI [13, 23], where *sessions*, advocated by Terry et al. [43], have been adopted by many production databases in practice (e.g., DGraph [24], Galera [17], and CockroachDB [18]);
2. does not rely on implementation details such as concurrency control mechanism as in [6] and timestamps as in [1], and the operational semantics of the underlying database as in [48], which are usually invisible to the outsiders; and
3. naturally models uncertain dependencies inherent to black-box testing using generalized constraints (Section 3) and enables the acceleration in SMT solving by compacting constraints (Section 5.4).

Regarding the comparison with [20], despite its promising characterization of SI suitable for black-box testing, we are unaware of any checking algorithm based on it. A straightforward (suboptimal) implementation would require enumerating all permutations of the transactions in a history, e.g., 10k transactions in our experiment indicate 10k-factorial permutations.

6.2 Dynamic Checking of SI

Dynamic checking of SI determines whether a collected history from dynamically executing a database satisfies SI. We are unaware of any black-box SI checker that satisfies SIEGE+.

dbcop [8] is the most efficient black-box SI checker to date. The underlying checking algorithm runs in $O(n^c)$, with n and c the number of transactions and clients involved in a single history, respectively. The authors devise both a polynomial-time algorithm for checking serializabilty (also with a fixed number of client sessions) and a polynomial-time algorithm for reducing checking SI to checking serializabilty. However, as demonstrated in Section 5.4, dbcop is practically not as efficient as our PolySI tool under various workloads. Moreover, dbcop is incomplete as it does not check non-cycle anomalies such as *aborted reads* and *intermediate reads* (Section 4.5). Finally, dbcop provides no details upon a violation; only a "false" answer is returned.

Elle [30], which is part of the Jepsen [27] testing framework, is a checker for a variety of isolation levels including SI and has identified bugs in real-world database systems. Elle requires specific data models, e.g., lists, in workloads to infer the version order (or the WW dependencies in our case) and specific APIs to perform, e.g., list-specific operations such as "append". Moreover, as Elle builds upon Adya's formalization, it also relies on the start and commit timestamps of transactions, which may not always be exposed to clients. In contrast, PolySI targets both general and production workloads and uses standard key-value and SQL APIs; the underlying SI characterization does not rely on any implementation details. Both Elle and PolySI are sound and complete (modulo determinate transactions), and return informative counterexamples.

ConsAD [50] is a checker tailored to application servers as opposed to black-box databases in our setting. Its SI checking algorithm is also based on dependency graphs. To determine the WW dependencies, ConsAD enforces the commit order of update transactions using, e.g., artificial SQL queries, to acquire exclusive locks on the database records, resulting in additional overhead [41, 50]. Moreover, ConsAD is incapable of detecting non-cycle anomalies.

CAT [32] is a dynamic white-box checker for SI (and several other isolation levels). The current release is restricted to distributed databases implemented in the Maude formal language [15]. CAT must capture the internal transaction information, e.g., start/commit times, during a system run.

7 CONCLUSION

We have presented the design of PolySI, along with a novel characterization of SI using generalized polygraphs. We have established the soundness and completeness of our new characterization and PolySI's checking algorithm. Moreover, we have demonstrated PolySI's effectiveness by reproducing all of 2477 known SI anomalies and by finding new violations in three popular production cloud databases, its efficiency by experimentally showing that it outperforms the state-of-the-art tools, and its generality by experimenting with a wide range of workloads and databases of different kinds. Finally, we have leveraged PolySI's interpretation algorithm to locate the causes of the violations found.

PolySI is the first black-box SI checker that satisfies the SIEGE+ principle. The obvious next step is to apply SMT solving to build SIEGE+ black-box checkers for other data consistency properties such as transactional causal consistency [25, 33] and the recently proposed regular sequential consistency [26]. Moreover, we will pursue the three research directions discussed in Section 5.5.

REFERENCES

[1] Atul Adya. 1999. *Weak Consistency: A Generalized Theory and Optimistic Implementations for Distributed Transactions.* Ph.D. Dissertation. USA.

[2] Anonymous Author(s). 2022. *Efficient Black-box Checking of Snapshot Isolation in Databases.* Technical Report. https://github.com/anonymous-hippo/PolySI/polysi-tr.pdf.

[3] Peter Bailis, Aaron Davidson, Alan Fekete, Ali Ghodsi, Joseph M. Hellerstein, and Ion Stoica. 2013. Highly Available Transactions: Virtues and Limitations. *Proc. VLDB Endow.* 7, 3 (nov 2013), 181–192. https://doi.org/10.14778/2732232.2732237

[4] Peter Bailis, Alan Fekete, Ali Ghodsi, Joseph M. Hellerstein, and Ion Stoica. 2016. Scalable Atomic Visibility with RAMP Transactions. *ACM Trans. Database Syst.* 41, 3, Article 15 (jul 2016), 45 pages. https://doi.org/10.1145/2909870

[5] Sam Bayless, Noah Bayless, Holger H. Hoos, and Alan J. Hu. 2015. SAT modulo Monotonic Theories. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence (AAAI'15).* AAAI Press, 3702–3709.

[6] Hal Berenson, Phil Bernstein, Jim Gray, Jim Melton, Elizabeth O'Neil, and Patrick O'Neil. 1995. A Critique of ANSI SQL Isolation Levels. In *SIGMOD '95*. ACM, 1–10. https://doi.org/10.1145/223784.223785

[7] Philip A Bernstein, Vassos Hadzilacos, and Nathan Goodman. 1986. *Concurrency Control and Recovery in Database Systems.* Addison-Wesley Longman Publishing Co., Inc., USA.

[8] Ranadeep Biswas and Constantin Enea. 2019. On the Complexity of Checking Transactional Consistency. *Proc. ACM Program. Lang.* 3, OOPSLA, Article 165 (Oct. 2019), 28 pages. https://doi.org/10.1145/3360591

[9] Ranadeep Biswas, Diptanshu Kakwani, Jyothi Vedurada, Constantin Enea, and Akash Lal. 2021. MonkeyDB: Effectively Testing Correctness under Weak Isolation Levels. *Proc. ACM Program. Lang.* 5, OOPSLA, Article 132 (oct 2021), 27 pages. https://doi.org/10.1145/3485546

[10] Ahmed Bouajjani, Constantin Enea, Rachid Guerraoui, and Jad Hamza. 2017. On verifying causal consistency. In *POPL'17*. ACM, 626–638.

[11] Apache Cassandra. Accessed August, 2022. https://cassandra.apache.org/.

[12] Andrea Cerone, Giovanni Bernardi, and Alexey Gotsman. 2015. A Framework for Transactional Consistency Models with Atomic Visibility. In *CONCUR'15 (LIPIcs)*, Vol. 42. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, 58–71.

[13] Andrea Cerone and Alexey Gotsman. 2018. Analysing Snapshot Isolation. *J. ACM* 65, 2, Article 11 (Jan. 2018), 41 pages. https://doi.org/10.1145/3152396

[14] Haicheng Chen, Wensheng Dou, Dong Wang, and Feng Qin. 2020. CoFI: Consistency-Guided Fault Injection for Cloud Systems. In *ASE 2020*. IEEE. https://doi.org/10.1145/3324884.3416548

[15] Manuel Clavel, Francisco Durán, Steven Eker, Patrick Lincoln, Narciso Martí-Oliet, José Meseguer, and Carolyn Talcott. 2007. *All about Maude - a High-Performance Logical Framework: How to Specify, Program and Verify Systems in Rewriting Logic.* Springer-Verlag, Berlin, Heidelberg.

[16] Amazon Elastic Compute Cloud. Accessed August, 2022. https://aws.amazon.com/cn/ec2/.

[17] MariaDB Galera Cluster. Accessed August, 2022. https://mariadb.com/kb/en/what-is-mariadb-galera-cluster/.

[18] CockroachDB. Accessed August, 2022. https://www.cockroachlabs.com/.

[19] Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest, and Clifford Stein. 2009. *Introduction to Algorithms, Third Edition* (3rd ed.). The MIT Press.

[20] Natacha Crooks, Youer Pu, Lorenzo Alvisi, and Allen Clement. 2017. Seeing is Believing: A Client-Centric Specification of Database Isolation. In *PODC '17*. ACM, 73–82. https://doi.org/10.1145/3087801.3087802

[21] Ben Darnell. Accessed August, 2022. Lessons Learned from 2+ Years of Nightly Jepsen Tests. https://www.cockroachlabs.com/blog/jepsen-tests-lessons/.

[22] Oracle Database. Accessed August, 2022. https://www.oracle.com/database/.

[23] Khuzaima Daudjee and Kenneth Salem. 2006. Lazy Database Replication with Snapshot Isolation. In *VLDB'06*. VLDB Endowment, 715–726.

[24] Dgraph. Accessed August, 2022. https://dgraph.io/.

[25] Diego Didona, Rachid Guerraoui, Jingjing Wang, and Willy Zwaenepoel. 2018. Causal Consistency and Latency Optimality: Friend or Foe? *Proc. VLDB Endow.* 11, 11 (2018), 1618–1632.

[26] Jeffrey Helt, Matthew Burke, Amit Levy, and Wyatt Lloyd. 2021. Regular Sequential Serializability and Regular Sequential Consistency. In *SOSP'21*. ACM, 163–179.

[27] Jepsen. Accessed August, 2022. https://jepsen.io.

[28] Jepsen. Accessed August, 2022. Issue #824. https://github.com/YugaByte/yugabyte-db/issues/824.

[29] Nick Kallen. Accessed August, 2022. Big Data in Real Time at Twitter. https://www.infoq.com/presentations/Big-Data-in-Real-Time-at-Twitter/.

[30] Kyle Kingsbury and Peter Alvaro. 2020. Elle: Inferring Isolation Anomalies from Experimental Observations. *Proc. VLDB Endow.* 14, 3 (Nov. 2020), 268–280.

[31] Leslie Lamport. 1978. Time, Clocks, and the Ordering of Events in a Distributed System. *Commun. ACM* 21, 7 (1978), 558–565.

[32] Si Liu, Peter Csaba Ölveczky, Min Zhang, Qi Wang, and José Meseguer. 2019. Automatic Analysis of Consistency Properties of Distributed Transaction Systems in Maude. In *TACAS 2019 (LNCS)*, Vol. 11428. Springer, 40–57.

[33] Wyatt Lloyd, Michael J. Freedman, Michael Kaminsky, and David G. Andersen. 2013. Stronger semantics for low-latency geo-replicated storage. In *NSDI' 13*. USENIX Association, 313–328.

[34] Haonan Lu, Siddhartha Sen, and Wyatt Lloyd. 2020. Performance-Optimal Read-Only Transactions. In *OSDI 2020*. USENIX Association, 333–349.

[35] MongoDB. Accessed August, 2022. https://www.mongodb.com/.

[36] Christos H. Papadimitriou. 1979. The Serializability of Concurrent Database Updates. *J. ACM* 26, 4 (oct 1979), 631–653. https://doi.org/10.1145/322154.322158

[37] Daniel Peng and Frank Dabek. 2010. Large-Scale Incremental Processing Using Distributed Transactions and Notifications. In *OSDI'10*. USENIX Association, USA, 251–264.

[38] PostgreSQL. Accessed August, 2022. Transaction Isolation. https://www.postgresql.org/docs/current/transaction-iso.html.

[39] RUBiS. Accessed August, 2022. Auction Site for e-Commerce Technologies Benchmarking. https://projects.ow2.org/view/rubis/.

[40] Microsoft SQL Server. Accessed August, 2022. https://www.microsoft.com/en-us/sql-server/.

[41] Zechao Shang, Jeffrey Xu Yu, and Aaron J. Elmore. 2018. RushMon: Real-Time Isolation Anomalies Monitoring. In *SIGMOD '18*. ACM, 647–662. https://doi.org/10.1145/3183713.3196932

[42] Cheng Tan, Changgeng Zhao, Shuai Mu, and Michael Walfish. 2020. COBRA: Making Transactional Key-Value Stores Verifiably Serializable. In *OSDI'20*. Article 4, 18 pages.

[43] Douglas B. Terry, Alan J. Demers, Karin Petersen, Mike Spreitzer, Marvin Theimer, and Brent B. Welch. 1994. Session Guarantees for Weakly Consistent Replicated Data. In *PDIS*. IEEE Computer Society, 140–149.

[44] Jepsen testing of MongoDB 4.2.6. Accessed August, 2022. http://jepsen.io/analyses/mongodb-4.2.6.

[45] Jepsen testing of TiDB 2.1.7. Accessed August, 2022. https://jepsen.io/analyses/tidb-2.1.7.

[46] TiDB. Accessed August, 2022. https://en.pingcap.com/tidb/.

[47] TPC. Accessed August, 2022. TPC-C: On-Line Transaction Processing Benchmark. https://www.tpc.org/tpcc/.

[48] Shale Xiong, Andrea Cerone, Azalea Raad, and Philippa Gardner. 2020. Data Consistency in Transactional Storage Systems: A Centralised Semantics. In *ECOOP'20*, Vol. 166. 21:1–21:31. https://doi.org/10.4230/LIPIcs.ECOOP.2020.21

[49] YugabyteDB. Accessed August, 2022. https://www.yugabyte.com/.

[50] Kamal Zellag and Bettina Kemme. 2014. Consistency anomalies in multi-tier architectures: automatic detection and prevention. *VLDB J.* 23, 1 (2014), 147–172. https://doi.org/10.1007/s00778-013-0318-x

[51] Rachid Zennou, Ranadeep Biswas, Ahmed Bouajjani, Constantin Enea, and Mohammed Erradi. 2019. Checking Causal Consistency of Distributed Databases. In *NETYS 2019 (LNCS)*, Vol. 11704. Springer, 35–51. https://doi.org/10.1007/978-3-030-31277-0_3

A PROOFS

A.1 Proof of Theorem 16

Proof. The proof proceeds in two directions.

("$\Longrightarrow$") Suppose that $\mathcal{H}$ satisfies SI. By Theorem 6, $\mathcal{H}$ satisfies Int and there exist WR, WW, and RW relations with which $\mathcal{H}$ can be extended to a dependency graph G such that $(\mathsf{SO}_G \cup \mathsf{WR}_G \cup \mathsf{WW}_G)\ ;\ \mathsf{RW}_G?$ is acyclic. We show that the generalized polygraph G' of $\mathcal{H}$ is SI-acyclic by constructing a compatible graph G'' with G' such that $G''|_{\mathcal{R}}$ is acyclic when the edge types are ignored: Consider a constraint $\langle either, or\rangle$ in G' and any WW edge $T \xrightarrow{\mathsf{WW}} S$ in G. If $(T, S, \mathsf{WW}) \in either$, add all the edges in *either* into G''. Otherwise, add all the edges in *or* into G''.

("$\Longleftarrow$") Suppose that $\mathcal{H} \models$ Int and the generalized polygraph G' of $\mathcal{H}$ is SI-acyclic. By Definition 15, there exists a compatible graph G'' with G' such that $G''|_{\mathcal{R}}$ is acyclic when the edge types are ignored. We show that $\mathcal{H}$ satisfies SI by constructing suitable WR, WW, and RW relations with which $\mathcal{H}$ can be extended to a dependency graph G such that $(\mathsf{SO}_G \cup \mathsf{WR}_G \cup \mathsf{WW}_G)\ ;\ \mathsf{RW}_G?$ is acyclic: We simply take G to be G'' by defining, e.g., $\mathsf{WW} = \{(a, b) \mid (a, b, \mathsf{WW}) \in E_{G''}\}$. □

A.2 Proof of Theorem 18

Proof. We prove (1) by induction on the number of iterations of pruning. Consider an arbitrary iteration of pruning $\mathcal{P}$ (lines 1:30–1:62) and denote the generalized polygraphs just before and after $\mathcal{P}$ by G_1 and G_2, respectively. We should show that G_1 is SI-acyclic if and only if $\mathcal{P}$ does not return False from line 1:51 or line 1:59 and G_2 is SI-acyclic. The following proof proceeds in two directions.

("$\Longrightarrow$") Suppose that G_1 is SI-acyclic. We then proceed by contradiction.

- Suppose that $\mathcal{P}$ returns False from line 1:51 or line 1:59. This happens when some constraint in G_1 constructed based on two write transactions, say T and S, on the same key cannot be resolved appropriately: every compatible graph with the induced SI graph of G_1 contains a cycle without adjacent RW edges, no matter whether $T \xrightarrow{\mathsf{WW}} S$ or $S \xrightarrow{\mathsf{WW}} T$ is in it. That is, G_1 is not SI-acyclic. Contradiction.
- Suppose that G_2 is not SI-acyclic. G_1 is not SI-acyclic because the set of constraints in G_2 is a subset of that in G_1 and $\mathcal{P}$ prunes a constraint only if one of its two possibilities cannot happen. Contradiction.

("$\Longleftarrow$") Suppose that $\mathcal{P}$ does not return False from line 1:51 or line 1:59 and G_2 is SI-acyclic. Therefore, there exists an acyclic compatible graph with the induced SI graph of G_2. This is also an acyclic compatible graph with the induced SI graph of G_1. Hence, G_1 is SI-acyclic.

The second part of the theorem holds because any compatible graph with the induced SI graph of G_p is also a compatible graph with the induced SI graph of G. □

B THE INTERPRETATION ALGORITHM

In this section, we describe the interpretation algorithm, called Interpret, that helps locate the causes of violations found by PolySI; see Algorithm 2. The algorithm takes as input the undesired cycles constructed from the log generated by MonoSAT, and outputs a dependency graph demonstrating the cause of the violation.

It is often difficult to locate the cause of a violation based solely on the original cycles because the cycles may miss some crucial information such as the core participating transactions and the dependencies between transactions. Therefore, Interpret first restores such information from the generalized polygraph of the history to reproduce the whole violating scenario (line 2:2). This may bring uncertain dependencies to the resulting polygraph, which are then resolved via pruning (line 2:3). Finally, Interpret finalizes the violating scenario by removing any remaining uncertain dependencies, as they are the "effect" of the violation instead of the "cause" (line 2:4).

Restore Transactions and Dependencies. The procedure Restore restores dependencies and the transactions involved in these dependencies in two steps (line 2:6). First, it checks each RW dependency in the input undesired cycles and restores its associated WW and WR dependencies if they are missing (lines 2:9 – 2:12). Then, for each WW dependency in an undesired cycle, including the ones just restored in the first step, it restores the other direction of this WW dependency (if missing) which may be involved in another undesired cycle (line 2:13 - line 2:24). Since the second step may introduce new RW dependencies, Restore repeats these two steps until no more dependencies (and involved transactions) are discovered (line 2:8).

Resolve Uncertain Dependencies. The restored violating scenario may contain WW and RW dependencies that are still uncertain at this moment. For an uncertain dependency to be part of the cause of a violation, it must be resolved. The procedure Resolve resolves as many uncertain dependencies as possible using the same idea of pruning (Section 4.3). Specifically, if an uncertain dependency from *either* (resp. *or*) in a constraint would create an undesired cycle with other certain dependencies, it, along with its associated dependencies in *either* (resp. *or*) is removed and the dependencies in *or* (resp. *either*) become certain (lines 2:33 - 2:39).

Finalize the Violating Scenario. The procedure Finalize finalizes the violating scenario by removing all the remaining uncertain dependencies, as they are the "effect" of the violation instead of the "cause" (line 2:40).

C A CAUSALITY VIOLATION FOUND IN YUGABYTEDB

Figure 11 shows how Algorithm 2 interprets a causality violation found in YugabyteDB. MonoSAT reports an undesired cycle $T : (0, 7) \xrightarrow{\mathsf{WW}(10)} T : (1, 15) \xrightarrow{\mathsf{WR}(13)} T : (0, 6) \xrightarrow{\mathsf{SO}} T : (0, 7)$; see Figure 11a. In this scenario, no transactions observe values that have been causally overwritten; consider the only transaction $T : (0, 6)$ in Figure 11a that reads. To locate the cause of the causality violation, PolySI first finds the (only) "missing" transaction $T : (0, 9)$ (colored in green) and the associated dependencies, as shown in Figure 11b. The WW and RW dependencies are uncertain at this moment (represented by dashed arrows), while the WR dependency is certain (represented by solid arrows, colored in blue). PolySI then restores the violating scenario by resolving such uncertainties. Specifically, as shown in Figure 11c, PolySI determines that $W(10, 3)$

Algorithm 2 The interpretation algorithm.

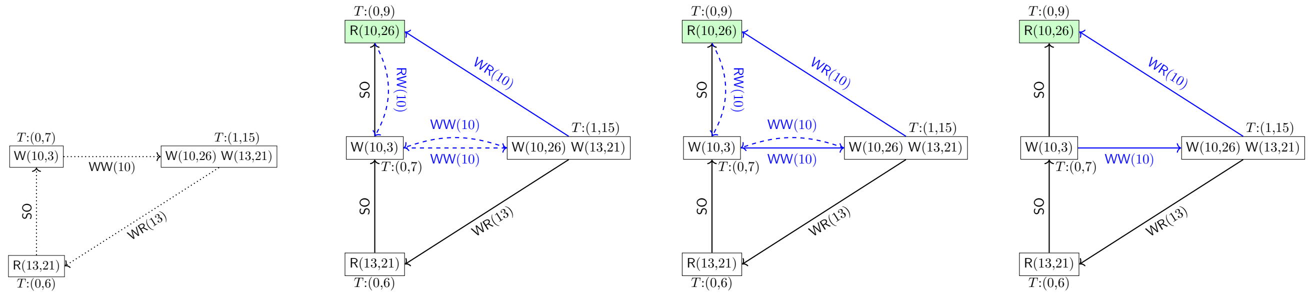
```
    $\mathcal{H} = (\mathcal{T}, \mathrm{SO})$: the original history
    $G = (V\_1, E\_1, C\_1)$: the polygraph
    $C = (V\_2, E\_2)$: The cycle found by PolySI

procedure INTERPRET($\mathcal{H}, G, C$)
    $Graph_{recovered} \leftarrow$ RESTORE($\mathcal{H}, G, C$)
    $Graph_{tagged} \leftarrow$ RESOLVE($Graph_{recovered}, \mathcal{H}, G, C$)
    $Graph_{finalized} \leftarrow$ FINALIZE($Graph_{tagged}$)
    return $Graph_{recovered}, Graph_{tagged}, Graph_{finalized}$

procedure RESTORE($\mathcal{H}, G, C$)
    $Graph_{recovered} \leftarrow C$
    while ( $Graph_{recovered}$ was changed in the last loop)
        for all $(u \xrightarrow{\mathrm{RW}} v) \in E\_2$
            for all $c \in C\_1$
                if $((u \xrightarrow{\mathrm{RW}} v) \in c) \land (\exists w, (w \xrightarrow{\mathrm{WW}} v) \in c)$
                    $Graph_{recovered} \leftarrow Graph_{recovered} \cup \{w \xrightarrow{\mathrm{WW}} v\} \cup \{w \xrightarrow{\mathrm{WR}} u\}$
        for all $dep \in (E\_2 \cap c.either)$, where $c \in C\_1$
            if $(\exists dep' \in c.or, dep' \in Graph_{recovered})$ continue
            for all $dep' \in c.or$
                if $(\exists$ a cycle $(V', E') \subseteq G) \land (dep' \in E')$
                    $Graph_{recovered} \leftarrow Graph_{recovered} \cup (V', E')$
                    break;
        for all $dep \in (E\_2 \cap c.or)$, where $c \in C\_1$
            if $(\exists dep' \in c.either, dep' \in Graph_{recovered})$ continue
            for all $dep' \in c.either$
                if $(\exists$ a cycle $(V', E') \subseteq G) \land (dep' \in E')$
                    $Graph_{recovered} \leftarrow Graph_{recovered} \cup (V', E')$
                    break;
    return $Graph_{recovered}$

procedure RESOLVE($Graph_{recovered}, \mathcal{H}, G, C$)
    $Graph_{tagged} \leftarrow Graph_{recovered}$
    for all $dep \in Graph_{tagged}$
        if $dep \in E\_1$
            $dep.tag \leftarrow$ 'certain'
        else
            $dep.tag \leftarrow$ 'uncertain'
    while $Graph_{tagged}$ was changed in the last loop
        for all $dep \in Graph_{tagged}$
            if $(dep.tag =$ 'uncertain'$) \land (dep$ in a cycle $(V', E'))$
                if $\forall dep' \in E' \land dep' \neq dep, dep'.tag =$ certain
                    $dep.tag \leftarrow$ 'uncertain'
                    for all $dep_{opposite}$, where $\{dep/dep_{opposite}\} \in C\_1$
                        $dep_{opposite}.tag \leftarrow$ 'certain'
        return $Graph_{tagged}$

procedure FINALIZE($Graph_{tagged}$)
    $Graph_{finalized} \leftarrow Graph_{tagged}$
    for all $dep \in Graph_{finalized}$
        if $dep.tag =$ 'uncertain'
            $Graph_{finalized} \leftarrow Graph_{finalized} \setminus \{dep\}$
    return $Graph_{finalized}$
```



(a) Original output **(b) Missing participants** **(c) Recovered scenario** **(d) Finalized scenario**

Figure 11: Causality violation: the SI anomaly found in YugabyteDB. The recovered dependencies are colored in blue with dashed and solid arrows indicating uncertain and certain dependencies, respectively. The (only) missing transaction is colored in green.

of transaction $T : (0, 7)$ was actually installed before $W(10, 26)$ of transaction $T : (1, 15)$, i.e., $T : (0, 7) \xrightarrow{\mathrm{WW}(10)} T : (1, 15)$. Otherwise, the other direction of dependency $T : (1, 15) \xrightarrow{\mathrm{WW}(10)} T : (0, 7)$ would enforce the dependency $T : (0, 9) \xrightarrow{\mathrm{RW}(10)} T : (0, 7)$, which would create an undesired cycle with the known dependency $T : (0, 7) \xrightarrow{\mathrm{SO}} T : (0, 9)$. Finally, PolySI finalizes the violating scenario by removing the remaining uncertainties from Figure 11c to obtain Figure 11d.

Thanks to the participation of transaction $T : (0, 9)$, the cause of the causality violation becomes clear: Transaction $T : (0, 7)$ causally depends on transaction $T : (1, 15)$, via $T : (1, 15) \xrightarrow{\mathrm{WR}(13)} T : (0, 6) \xrightarrow{\mathrm{SO}} T : (0, 7)$. However, transaction $T : (0, 9)$ following transaction $T : (0, 7)$ on the same session reads the value 26 of key 10 from transaction $T : (1, 15)$, which should have been overwritten by transaction $T : (0, 7)$.