TRANSACTIONS OF THE
AMERICAN MATHEMATICAL SOCIETY, SERIES B
Volume 8, Pages 1–38 (February 2, 2021)
https://doi.org/10.1090/btran/56

PERSISTENT OBSTRUCTION THEORY FOR A MODEL CATEGORY OF MEASURES WITH APPLICATIONS TO DATA MERGING

ABRAHAM D. SMITH, PAUL BENDICH, AND JOHN HARER

Abstract. Collections of measures on compact metric spaces form a model category ("data complexes"), whose morphisms are marginalization integrals. The fibrant objects in this category represent collections of measures in which there is a measure on a product space that marginalizes to any measures on pairs of its factors. The homotopy and homology for this category allow measurement of obstructions to finding measures on larger and larger product spaces. The obstruction theory is compatible with a fibrant filtration built from the Wasserstein distance on measures.

Despite the abstract tools, this is motivated by a widespread problem in data science. Data complexes provide a mathematical foundation for semi-automated data-alignment tools that are common in commercial database software. Practically speaking, the theory shows that database JOIN operations are subject to genuine topological obstructions. Those obstructions can be detected by an obstruction cocycle and can be resolved by moving through a filtration. Thus, any collection of databases has a persistence level, which measures the difficulty of JOINing those databases. Because of its general formulation, this persistent obstruction theory also encompasses multi-modal data fusion problems, some forms of Bayesian inference, and probability couplings.

1. Introduction

We begin this paper with an abstraction of a problem familiar to any large enterprise. Imagine that the branch offices within the enterprise have access to many data sources. The data sources exposed to each office are related and overlapping but non-identical. Each office attempts to merge its own data sources into a cohesive whole, and reports its findings to the head office. This is done by humans, often aided by ad-hoc data-merging software solutions. Presumably, each office does a good job of merging the data that they see. Now, the head office receives these cohesive reports, and must combine them into an overall understanding.

This paper provides a mathematical foundation combining methods from measure theory, simplicial homotopy, obstruction theory, and persistent cohomology (Section 1(a) gives an overview) for semi-automated data-table-alignment tools (e.g, HumMer [14]) that are common in commercial database software. Data tables are

Received by the editors December 4, 2019, and, in revised form, August 7, 2020.
2020 *Mathematics Subject Classification.* Primary 55U10; Secondary 55S35.
Work by all three authors was partially supported by the DARPA Simplex Program, under contract # HR001118C0070. The last two authors were also partially supported by the Air Force Office of Scientific Research under grant AFOSR FA9550-18-1-0266.

abstracted as measures over value spaces. The problem of merging tables, or indeed merging previously-merged tables, is recast as the search for a measure that marginalizes correctly.

For example, one data table might record the ages and heights and weights of patients in a hospital, abstracted as a formal sum of point-mass atoms. Another data table might be an actuarial table giving ages and heights and weights for an entire population, abstracted as a smooth distribution where the heights and weights form 2-dimensional elliptical Gaussian distributions for each age and height, the means and singular values varying with age. Both examples would be data tables on the same age-by-height-by-weight space. A third data table might be a simulated probabilistic model of injury severity during vehicle collisions based on height and weight of the passenger. This data table on height-by-weight-by-severity space may or may not merge with each of the previous examples over height-by-weight, within some error. One can imagine many other data tables collected from myriad sources (motor-vehicle records, longitudinal studies, clinical trials) related to this example that may be of interest.

Our first fundamental result (Theorem 3.11) uses this measure-theoretic lens to draw a surprising correspondence between the process of JOIN in database engineering and the Kan extension property for simplicial sets.

This abstraction, and the model-theoretic tools that come with it, permits several advances over the current state of the art, which are collected in our second fundamental result (Theorem 4.13). First, inconsistencies in table collections are automatically detected as *obstructions* in the sense of Steenrod (i.e, a certain cocycle is not zero). Second, when inconsistencies are detected, the obstruction theoretic tools, combined with persistent cohomology, provide two potential remedies: a) if algebraically permitted (i.e, a certain co-class is zero), the user may retreat back one level of merging, repair, and proceed; b) else, the user may settle for a measure that only marginalizes approximately correctly, with the degree of possible correctness computed automatically by persistent obstruction theory.

More broadly, we are interested in the following three meta-problems:

Problem 1.1 (Testing problem)**.** Given several sources of partial information, how do we test that a hypothetical explanation is reasonably consistent with that partial information?

Problem 1.2 (Merging problem)**.** Given several sources of partial information, how do we merge that partial information into a cohesive whole?

Problem 1.3 (Meta-merging problem)**.** Given many sources of partial information, and several partial attempts to merge some combinations of them, is there a way to merge these partial merges into a cohesive whole?

By "sources of partial information" we mean, roughly, collected data (databases, spreadsheets, pictures), statistical models, established theories, simulations, and general population trends. In this article, we define a formal mathematical structure—a *Data Complex* (Section 2)—that can encapsulate a wide range of problems like 1.1, 1.2, and 1.3. A data complex is meant to encode each source of information as a finite measure on an appropriate value space. Whether these measures arise from collected data as in Problem 1.2 or some model/theory/simulation/trend/merger derived from previous work as in Problem 1.3, we call them

data tables. By using measures, we combine Problems 1.2 and 1.3 into a single problem.

1(a). **Overview of technical approach.** Often, formal mathematics looks very different than its intuitive purpose, so we want to make sure the reader understands our intent, absent its formal notation.

We want a mathematically rigorous way to encode and solve Problems 1.1–1.3. When translated to the language of data complexes, a physically reasonable process for "merge [data tables] into a cohesive whole" can be expressed in terms of four key mathematical ingredients: homological algebra for simplicial sets, simplicial homotopy [10, 18], obstruction theory [23], and persistent (co)homology across a filtration [4].

The first ingredient (homological algebra) is used because data tables may overlap partially, meaning that we need a formal simplicial complex to encode all possible intersections of all possible data tables. Moreover, simplicial sets allow data tables with repeated columns. The marginalization integral provides a face map and a boundary operator, and an analogue of the diagonal measure within a product provides the degeneracy map. The face and boundary operators tell us whether one "narrow" data table reproduces the data of another "wider" data table via marginalization integrals. Thus, the question of whether several "narrow" data tables can be merged into a few "wider" data tables becomes the question of whether a k-chain is the boundary of a $(k+1)$-chain. That is, the ability to merge overlapping partial information sources as in Problem 1.2 is encoded as the homology of a data complex.

The second ingredient (simplicial homotopy) arises because Problems 1.1, 1.2, 1.3 suggest that we want "simple" solutions. Even when partial merging is possible in the sense of homology, it may be that the result is too complicated to be merged further. In the study of geometric bundles, the fundamental group and higher homotopy groups of the fiber play a key role, and we use simplicial homotopy in a similar way here. A simple solution to Problem 1.2/1.3 or a simple hypothesis in Problem 1.1 corresponds to a single data table (as opposed to a complicated chain of many data tables), which is indicated by triviality in the simplicial homotopy group.

An important side effect of introducing simplicial homotopy (via model categories) is that we see that the Kan extension condition means "merging operations are possible." The process we call *merging* is similar to *JOIN'ing* in database software, to *fusion* in multi-modal data analysis, and to *coupling* in probability theory. This link reinforces the intuition that data complexes are a good way to encode Problems 1.1/1.2/1.3 for modern data mining when using spreadsheets, DataFrames, and SQL databases. Indeed, our first fundamental result (Theorem 3.11) explicitly formalizes this correspondence.

The reader may be wondering why we introduce something as abstract as simplicial homotopy into something so concrete and common as data merging. Consider the typical database operation

```
SELECT * FROM table1 INNER JOIN table2
ON table1.column1 = table2.column2
WHERE condition;
```

When issuing such a command, the administrator *must* designate two tables to be `JOIN`ed and choose specific columns from the two tables to be identified via the `ON` clause. The `SELECT * ...;` command returns a table, whose columns must appear in some order that is determined by the ordering of attributes in `table1` and `table2`, by their placement in the command, and by the columns in the `ON` clause. Thus, in the language of Section 2, the database software and the working administrator must agree on a total set of attributes, the attributes in each table, and an ordered attribute inclusion to be used for the `ON` clause.

This command also indicates why we formalize "data tables" as measures over products of attribute value spaces. Replacing `SELECT *` with a `SELECT columnList` corresponds to the ability to re-order the attribute list and to marginalize the output to a sublist of attributes; hence, arbitrary finite products are possible. The optional `WHERE condition` clause allows one to impose additional restrictions on the values to be considered by imposing logical conditions on the entries, such as `WHERE (age > 18 AND height > 200)`. These conditions allow one to restrict the data table to any[1] measurable subset of the value space. The entries of a `WHERE`-restricted data table constitute the mass of this measurable set, with respect to the data table. (Finally, for those fluent in SQL subtleties, note that the ability to perform `LEFT`, `RIGHT`, and `OUTER JOIN` instead of `INNER JOIN` will be encompassed by approximate join and face operations in Section 4.)

The third ingredient (Steenrod's obstruction theory as in [22]) provides guidance on how to combine homological algebra and homotopy theory to detect and describe any obstructions to an iterative merging process. In its original formulation, obstruction theory asks whether a section σ of a fiber bundle p defined over a topological space B can be extended to a section defined over a larger topological space $A \supset B$? The most famous example is the smooth category, where one computes characteristic cohomology classes to indicate whether sections of a bundle can be extended globally. Steenrod studied this problem in the case of fibrations over general topological spaces. Typically, assuming one has some sort of CW structure on A, one tries to extend σ first over the 0-skeleton of A and then the 1-skeleton of A, and so forth. Assuming one has already extended σ to a section over the $(n-1)$-skeleton of A, Steenrod's obstruction cocycle is an element ξ_σ of $C^n(A, B; \pi_{n-1,\sigma}(F_0))$, where $\pi_{n-1,\sigma}(F_0)$ is the homotopy group of the fiber F_0 of p, as twisted by σ; loosely, the co-chain ξ_σ is defined on each n-cell c of A by restricting σ to the boundary of c, but there is some nuance coming from the twisting needed to turn this into a homotopy class of the fiber. If this cocyle $\xi_\sigma : C_n \to \pi_{n-1}$ is trivial in homotopy, then the section σ can be extended, and otherwise it cannot. However, if this cocycle is a coboundary, $[\xi_\sigma] = 0$, then there is another section τ, agreeing with σ on the $(n-2)$-skeleton of A, with ξ_τ trivial in homotopy. Hence, obstructions are discovered dimension-by-dimension via homotopy-valued cohomology, and the obstruction computation can often permit the "correction" of initial extensions of the section to avoid higher-dimensional obstructions.

This concept of an obstruction cocycle was introduced by Steenrod in [22] and revisited many times, such as [15] and [11]. Its importance motivated early work in category theory. The entire *raison d'être* of defining fibrant objects and model categories in [10] and [18] was to establish the most general context in which these (co)homology and homotopy calculations remain sensible for more general notions

[1] Any measurable subset—in principle and given a sufficiently generous SQL implementation.

of "weak equivalence." In particular, one does not require actual topological spaces to perform obstruction theory, merely fibrant objects in a model category.

Here, we establish homology and homotopy theory for data tables by relying on these categorical foundations, giving us an obstruction theory directly analogous to Steenrod's. When sequential merging is impossible, the obstruction cochain can compute specific data tables that obstruct the process. That is, obstruction theory determines when Problem 1.3 is solvable locally but not globally.

The fourth ingredient, persistent (co)homology, provides a mathematically robust way to measure how much the underlying original data tables would have to be altered, in order to overcome an obstruction. This is a key feature of the theory, because from a practical perspective, multiple information sources are *never* perfectly consistent. Typos and transpositions and omissions and error bars always exist, and must be accounted for. We use a filtration built from the Wasserstein distance on measures to ensure that the desired simplicial homotopy is possible throughout all levels of the filtration. This allows for a well-defined notion of *persistent obstruction theory.* Our second fundamental result (Theorem 4.13) formalizes the idea that when inconsistencies are identified, one of two remedies may be available[2]

- the head office should retreat back one merging level, repair (with repair suggested by algebra), then again seek consensus
- the head office should settle for only approximate consensus, where the desired measure only approximately marginalizes correctly, with the degree of approximation computed via persistence.

The ultimate result of this article is a mathematically robust framework for data merging that is reasonably applicable to real-world data. In this framework, Problems 1.1 and 1.2/1.3 become Problems 2.38 and 2.39, which are answered by Theorem 4.13 and Definition 4.14.

1(b). **Related work.** To the best of our knowledge, this is the first work to combine all of the tools above to build a robust obstruction theory for databases. Other authors have used different aspects of these tools to address databases. For example, recent work by Fong and Spivak uses database schemas and type/value relationships as a motivational example to introduce functors and (co)limits [6, Chapter 3]. Specifically Example 3.99 and the chapter's final remark are somewhat in the same spirit as the approach taken here. Our category of data complexes in Section 2 is similar to the categorical presentation in [20, 21], but our data tables are built from measures (not sets) in order to flexibly address the errors that are inevitable in applications. Finally, other recent work by Abramsky, Morton, and collaborators uses obstruction theory (in the sheaf-theoretic context) to detect non-contextuality in quantum theory, with an application to the non-existence of a universal data table that contains a set of given tables [1, 2, 13]. We expect that further interweaving of these measure-theoretic, sheaf-theoretic, and simplicial/categorical perspectives will be fruitful in the future.

1(c). **Outline.** The rest of this paper is organized as follows. Section 2 defines the basic object of study, a *data complex*, and draws a mapping between its simplicial set structure and the choices that must be made by any database administrator. Categorical language is alluded to in this section, but a full categorical treatment of data complexes is confined to the Appendix. Section 3 connects simplicial homotopy

[2]In fact, the second is always available, but may be less desirable!

to the notion of JOIN, and shows how obstruction theory detects the impossibility of merging. Section 4 describes our notion of persistent obstruction theory and its application to the idea of fuzziness of consensus. The paper concludes with discussion of practical considerations for applications in Section 5.

2. Attributes and data tables

This section provides a practical developmental discussion of a Data Subcomplex that should be accessible to a fairly wide mathematical audience, with full categorical language found in the Appendix. The basic definitions appear in Sections 2(a) and 2(b), culminating in Theorem 2.14 which shows that we have indeed defined a simplicial set. Operations that are specifically useful to standard database operations (inclusion/merge/join) are defined in Section 2(c). Then Section 2(d) makes plain the analogue of "section of a bundle," which permits the rephrasing of our fundamental problems in mathematical language, and Section 2(e) defines the (co)homology of data complexes needed for obstruction theory.

2(a). **Data subcomplex as a simplicial set.** Our definitions are aimed at making precise the following real-life scenario in data administration.

(1) The administrator chooses a set A of all attributes (column names and variable types) of interest.
(2) For each attribute a in the list A, the administrator chooses a space of possible values, and a "reasonable" metric ρ_a that can provide the distance between any two values in that space. Our notion of "reasonable" includes compactness, which is typically guaranteed by boundedness of realistic integer or vector-valued entries.
(3) The administrator acquires "data" for some lists of attributes, and attempts to reconcile these into a joint view across all attributes in A. The reconciliation process involves "join" operations that could be represented by SQL commands such as

```
SELECT * FROM table1 INNER JOIN table2
ON table1.column1 = table2.column2
WHERE condition;
```

(4) When reconciling, the administrator may choose to alter the data, as long as the alterations are "small" with respect to both the individual values via ρ_a and with respect to the overall information-theoretic content of the data.

The former two items are choices that must be made. The latter two items are a process to be accomplished. The mathematical structure developed here is informed deeply by the example SQL command, as discussed in Section 1(a).

Let us define our objects. It is convenient to use language of category theory; see Appendix A for our conventions.

Consider a finite set A. The elements are called *attributes*. For each attribute $a \in A$, there is a compact metric space $(\mathbb{V}(a), \rho_a)$, called the *value space*.[3] These assumed objects (the finite set of attributes and a compact metric space assigned

[3]These assumptions imply that $\mathbb{V}(a)$ is complete, separable and is a Radon space. In many applications, the space $\mathbb{V}(a)$ is finite or a closed interval in $\mathbb{R}$, so one needn't imagine esoteric spaces to grasp the theory.

to each attribute) are user-supplied by a data administrator; after these choices are made, everything else proceeds as defined.

Each $\mathbb{V}(a)$ is a Radon space (in particular, a measurable space) using the usual Borel algebra from the metric ρ_a. These metrics will be used in Section 4 to quantify levels of acceptable imprecision when marginalizing measures.

An *attribute list* $T = [a_0, a_1, \ldots, a_n]$ is a finite sequence of attributes; that is, an attribute list is a function $T : \{0, \ldots, n\} \to A$. The *length* of an attribute list is $\mathrm{len}(T) \coloneqq n + 1$. An attribute list T is called *nondegenerate* if it contains no repetitions; that is if the function T is one-to-one. The longest nondegenerate attribute lists are permutations of A.

For any attribute list T, the product space $\mathbb{V}(T) \coloneqq \prod_{i=0}^{n} \mathbb{V}(a_i)$ is well-defined. The product space $\mathbb{V}(T)$ admits the L^∞ metric $\rho_T = \max_{a \in T} \rho_a$ and is measurable via the corresponding tensor-product algebra.[4] For any list $\tilde{A}$ representing a permutation of the set A, then $\mathbb{V}(\tilde{A})$ is the correspondingly ordered total product of all the measurable spaces of all attributes. At the other extreme, we equip the empty attribute list $[]$, of length 0, with the *trivial value space* as $\mathbb{V}([]) = \{*\}$, a singleton set.

Definition 2.1 (Set of attribute lists)**.** Let $\mathcal{A}$ denote the set of all attribute lists in A. For each $n \geq -1$, let $\mathcal{A}_n \subset \mathcal{A}$ denote the set of all attribute lists of length $n+1$. $\mathcal{A}$ is a small category. Using the notation from Appendix A, an object in $\mathcal{A}$ is a function $T : \mathbf{n} \to A$. The case $n = -1$, giving the empty list $T = []$, is allowed. A morphism of attribute lists $T \to T'$ is given by $\ell : \mathbf{n}' \to \mathbf{n}$ (an order-preserving function, which is a morphism of $\mathbf{\Delta_a}$ as in Appendix A) such that $T' = T \circ \ell$, which is natural for the commutative diagram (2.1).

$$\begin{array}{ccc} \mathbf{n} & \xleftarrow{\ \ell\ } & \mathbf{n}' \\ {\scriptstyle T}\downarrow & & \downarrow{\scriptstyle T'} \\ A & \xleftarrow{\ =\ } & A \end{array} \tag{2.1}$$

In Section 2(b) it is shown that for $n \geq 0$, each $\mathcal{A}_n$ is equipped with face maps $d_i : \mathcal{A}_n \to \mathcal{A}_{n-1}$ (by omission of the ith element as in Definition 2.7) and degeneracy maps $s_i : \mathcal{A}_n \to \mathcal{A}_{n+1}$ (by repetition of the ith element as in Definition 2.11). When omitting the trivial -1-level, $\mathcal{A}$ is the simplicial set whose elements are generated by the permutations of A via the face and degeneracy maps. Including the trivial -1-level, $\mathcal{A}$ is the augmented simplicial set generated this way. See Appendix A for a summary of the standard definition of (augmented) simplicial sets.

For any attribute list T, let $\mathbf{M}(T)$ denote the space of finite measures on $\mathbb{V}(T)$. A *data table* is a pair (T, τ) for $\tau \in \mathbf{M}(T)$ for any $T \in \mathcal{A}$. Note that $\mathbf{M}([]) \cong \mathbb{R}_{\geq 0}$, as a measure on the singleton set $\mathbb{V}([])$ is determined by the mass $M \geq 0$ of $\{*\}$. A *trivial data table* is any data table of the form (T, τ) where $T = []$ and $\tau = M \geq 0$ is a measure on the singleton set $\mathbb{V}([]) = \{*\}$. We sometimes abbreviate our notation for data tables from (T, τ) to τ, because any $\tau \in \mathbf{M}(T)$ is equipped with a domain (the measurable sets in $\mathbb{V}(T)$), so T is understood in context.

[4] We use the ∞-metric for ease of proof when studying filtrations. Other p-metrics or more general product metrics might carry the whole theory, too, but we have not yet verified this.

In the first example alluded to in the introduction, we could have $T = [\text{age}, \text{height}, \text{weight}]$ with $\mathbb{V}(\text{age}) = \{0, 1, \ldots, 150\}$ in integer years, $\mathbb{V}(\text{height}) = [0, 500] \subset \mathbb{R}$ in centimeters, and $\mathbb{V}(\text{weight}) = [0, 1000] \subset \mathbb{R}$ in kilograms, each with the standard metric. The space $\mathbf{M}(T)$ would be the set of measures on the compact set $\mathbb{V}(T) \subset \mathbb{R}^3$ given by the product. An attribute list $[\text{height}, \text{height}]$ is also permissible, and might arise for example if heights were compared from two different sources (driver's license versus medical chart).

For practical purposes, because $\mathbb{V}(T)$ is a compact metric space, one might use the Radon–Nikodym theorem to write any $\tau \in \mathbf{M}(T)$ using a density function with respect to the uniform[5] probability measure on the compact set; however, for simplicity we use the language and notation of measures instead of the language of functions and integrals.

Definition 2.2 (Ambient data complex)**.** Given A, the *ambient data complex* over A is the set of all data tables,

$$\mathcal{X} = \{(T, \tau) \ : \ T \in \mathcal{A}, \ \tau \in \mathbf{M}(T)\}.$$

For $-1 \leq n$, let $\mathcal{X}_n = \{(T, \tau) \in \mathcal{X} \ : \ T \in \mathcal{A}_n, \ \tau \in \mathbf{M}(T)\}$. Let $p : \mathcal{X} \to \mathcal{A}$ denote the forgetful map $p : (T, \tau) \mapsto T$.

Theorem 2.14 shows that the ambient data complex is a simplicial set (augmented when including $\mathcal{X}_{-1}$) with faces given by the marginalization integrals (Definition 2.9) and degeneracies given by Dirac diagonalizations or intersections (Definition 2.13). The ambient data complex $\mathcal{X}$ is a small category, whose morphisms are generated by faces and degeneracies. The forgetful functor p is a simplicial map between the small categories $\mathcal{X}$ and $\mathcal{A}$.

Definition 2.3 (Data subcomplex)**.** Given an ambient data complex $p : \mathcal{X} \to \mathcal{A}$, a *Data Subcomplex* is a subset/subcategory $\mathcal{X}' \subseteq \mathcal{X}$ that is closed under the face and degeneracy maps defined in 2.9 and 2.13. Because p is a simplicial map, the attribute base

$$\mathcal{A}' = p(\mathcal{X}') = \{T \in \mathcal{A}_n \ : \ \exists n \geq -1, \ \exists (T, \tau) \in \mathcal{X}'_n\}$$

is a simplicial subset of $\mathcal{A}$.

Definition 2.4 (Finitely generated)**.** A data subcomplex $p : \mathcal{S} \to \mathcal{B}$ of an ambient $p : \mathcal{X} \to \mathcal{A}$ is said to be *finitely generated* iff there is a finite set $\{(T_1, \tau_1), \ldots, (T_K, \tau_K)\} \subset \mathcal{S}$ such that every data table in $\mathcal{S}$ is obtained from this finite set via face and degeneracy maps. We write $\mathcal{S} = \langle (T_1, \tau_1), \ldots, (T_K, \tau_K) \rangle$ or just $\mathcal{S} = \langle \tau_1, \ldots, \tau_K \rangle$.

Definition 2.5 (Closed under permutation)**.** A subset $\mathcal{B}$ of $\mathcal{A}$ is said to be *closed under permutation* iff for any $T \in \mathcal{S}$ with $\text{len}(T) = n{+}1$ and for any permutation (that is, bijection)[6] $\varsigma : \{0, \ldots, n\} \to \{0, \ldots, n\}$, there exists $\tilde{T} = T \circ \varsigma \in \mathcal{B}$. A data subcomplex $p : \mathcal{S} \to \mathcal{B}$ of an ambient $p : \mathcal{X} \to \mathcal{A}$ is said to be *closed under permutation* iff for any $(T, \tau) \in \mathcal{S}$ with $\text{len}(T) = n{+}1$ and for any permutation $\varsigma : \{0, \ldots, n\} \to \{0, \ldots, n\}$, there exists $(\tilde{T}, \tilde{\tau}) \in \mathcal{S}$ and such that the measure $\tilde{\tau}$ is evaluated on the basis sets $U_{\varsigma(0)} \times \cdots \times U_{\varsigma(n)}$ of the Borel algebra of $\mathbb{V}(\tilde{T})$ by

$$\tilde{\tau}(U_{\varsigma(0)} \times \cdots \times U_{\varsigma(n)}) = \tau(U_0 \times \cdots \times U_n).$$

[5]That is, the measure depends only on r, for metric balls $B_r(x)$ of sufficiently small radius.

[6]Note that a nontrivial permutation is *not* a morphism in $\mathbf{\Delta_a}$.

Remark 2.6. Actual database merging problems encountered in real-life situations such as Problems 1.1–1.3 always present themselves as Finitely Generated Data Subcomplexes, because there is some finite set of database tables or spreadsheets under consideration. The face and degeneracy maps provide the logical relations between these tables that allow or prevent joining. Real-life situations are also closed under permutation; because, the "`SELECT * FROM ...`" clause in SQL allows the database engineer to re-order the columns of any table. In our earlier example, a data table given by listing patients' age-by-height-by-weight might be permuted to height-by-weight-by-age simply by reordering the columns of the spreadsheet.

Notational note. We always use $p : \mathcal{X} \to \mathcal{A}$ to refer to an ambient data complex. We use either $p : \mathcal{X}' \to \mathcal{A}'$ or $p : \mathcal{S} \to \mathcal{B}$ to refer to a data subcomplex of $p : \mathcal{X} \to \mathcal{A}$. We tend to use $p : \mathcal{S} \to \mathcal{B}$ when we imagine that this data subcomplex came from an actual data merging problem (so it is likely to be finitely generated and closed under permutation); however, we state explicitly these conditions when they are required for a result. When the projection p and the attribute simplicial sets $\mathcal{A}, \mathcal{B}$ are not used in a statement, we omit them and write "a data subcomplex $\mathcal{S}$ of an ambient $\mathcal{X}$."

2(b). **Morphisms of data tables.** This section establishes notation for common operations and proves that $\mathcal{A}$ and $\mathcal{X}$ are simplicial sets, establishing that they are small categories with morphisms that are well-understood in language of measures.

Definition 2.7 (Face of attribute list)**.** The face map on attribute lists, $d_i : \mathcal{A}_n \to \mathcal{A}_{n-1}$, is defined as omission of the ith entry a_i in $T = [a_0, \ldots, a_i, \ldots, a_n]$, so

$$d_i[a_0, \ldots, a_{i-1}, a_i, a_{i+1}, \ldots, a_n] = [a_0, \ldots, a_{i-1}, a_{i+1}, \ldots, a_n].$$

Remark 2.8 (Categorical interpretation). In Definition 2.7, $d_i T = T \circ d^i = (d^i)^* T$, where $d^i : \mathbf{n-1} \to \mathbf{n}$ is the co-face monomorphism in $\mathbf{\Delta_a}$, as in Appendix A.

Definition 2.9 (Face of data table)**.** For a data table $(T, \tau) \in \mathcal{X}_n$ with $T = [a_0, \ldots, a_i, \ldots, a_n]$, let $d_i(\tau) \in \mathbf{M}(d_i(T))$ be the measure evaluated on the basis sets $U_0 \times \cdots \times U_{i-1} \times U_{i+1} \times \cdots \times U_n$ of the Borel algebra on $\mathbb{V}([a_0, \ldots, a_{i-1}, a_{i+1}, \ldots, a_n]) = \mathbb{V}(d_i T)$ as

$$d_i(\tau)(U_0 \times \cdots U_{i-1} \times U_{i+1} \times \cdots \times U_n) \coloneqq \tau(U_0 \times \cdots \times U_{i-1} \times \mathbb{V}(a_i) \times U_{i+1} \times \cdots \times U_n).$$

This is the measure obtained by marginalization to omit the ith factor, which could also be written as $d_i \tau \coloneqq \int_{\mathbb{V}(a_i)} \tau$. Let $d_i(T, \tau) \coloneqq (d_i T, d_i \tau)$, which is well-defined in $\mathcal{X}_{n-1}$.

In our earlier example with individual patients as atomic point-masses, the face map d_0 from age-by-height-by-weight to height-by-weight represents deleting the age column of the spreadsheet, and allowing new duplicate entries to add (that is, integrate) measure.

Face maps can be applied multiple times, and the following lemma provides the desired re-ordered "commutation" property. For attribute lists the proof is immediate; for data tables it is the Fubini–Tonelli Theorem applied to the measures.

Lemma 2.10 (Fubini–Tonelli Theorem)**.** *For any $i < j$, $d_i \circ d_j = d_{j-1} \circ d_i$.*

Definition 2.11 (Degeneracy of attribute list)**.** The degeneracy map on attribute lists, $s_i : \mathcal{A}_n \to \mathcal{A}_{n+1}$, is defined as repetition of the ith entry a_i in $T = [a_0, \ldots, a_i, \ldots, a_n]$, so $s_i T = [a_0, \ldots, a_i, a_i, \ldots, a_n]$.

Remark 2.12 (Categorical interpretation). In Definition 2.11, $s_i T = T \circ s^i = (s^i)^* T$, where $d^i : \mathbf{n} + \mathbf{1} \to \mathbf{n}$ is the co-degeneracy epimorphism in $\mathbf{\Delta_a}$, as in Appendix A.

Definition 2.13 (Degeneracy of data table)**.** For a data table $(T, \tau) \in \mathcal{X}_n$, let $s_i(\tau) \in \mathbf{M}(s_i(T))$ be the measure evaluated on the basis sets $U_0 \times \cdots U_i \times U_i' \times \cdots U_n$ of the Borel algebra on $\mathbb{V}([a_0, \ldots, a_i, a_i, \ldots, a_n]) = \mathbb{V}(s_i T)$ as

$$s_i(\tau)(U_0 \times \cdots U_i \times U_i' \times \cdots \times U_n) \coloneqq \tau(U_0 \times \cdots \times (U_i \cap U_i') \times \cdots \times U_n).$$

Then, $s_i(T, \tau) = (s_i T, s_i \tau)$ is well-defined in $\mathcal{X}_{n+1}$.

If the measure is expressed as a density function via the Radon–Nikodym theorem, then this is the Dirac-delta

$$s_i(\tau)(x_0, \ldots, x_i, x_i', \ldots, x_n) \coloneqq \tau(x_0, \ldots, x_i, \ldots, x_n)\delta(x_i, x_i'),$$

Theorem 2.14 (Simplicial sets)**.** *Let $\mathcal{X}$ be the ambient data complex over an attribute set A. For any $(T, \tau) \in \mathcal{X}_n$, consider the face maps $d_i(T, \tau)$ and degeneracy maps $s_i(T, \tau)$ as in the definitions above. Then*

1. $d_i \circ d_j = d_{j-1} \circ d_i$, *if* $i < j$;
2. $d_i \circ s_j = s_{j-1} \circ d_i$, *if* $i < j$;
3. $d_j \circ s_j = d_{j+1} \circ s_j = \mathrm{id}$;
4. $d_i \circ s_j = s_j \circ d_{i-1}$, *if* $i > j + 1$; *and*
5. $s_i \circ s_j = s_{j+1} \circ s_i$, *if* $i \leq j$.

Moreover, the forgetful map $p : \mathcal{X} \to \mathcal{A}$ commutes with d_i and s_i. That is, $\mathcal{X} = (\mathcal{X}_n, d, s)$ and $\mathcal{A} = (\mathcal{A}_n, d, s)$ are augmented simplicial sets as in Lemma A.4. *They are simplicial sets when omitting the trivial $\mathcal{X}_{-1}$ and $\mathcal{A}_{-1}$.* [7, Definition 3.2], [9, Equation (1.3)], [12, Definition 1.1]

Proof. This is direct with no surprises, by working on the Borel basis sets $U_0 \times \cdots \times U_n$ for $\mathbb{V}([a_0, \ldots, a_n])$. The $d_i d_j$ condition was already seen as Fubini–Tonelli. □

2(c). **Inclusions, merges, and joins.** We now establish[7] additional operations (inclusion, sum, merge, join) that are special to $\mathcal{A}$ and $\mathcal{X}$ and do not apply to general simplicial sets.

Definition 2.15 (Attribute inclusions)**.** An *attribute inclusion*

$$[a_0, a_1, \ldots, a_{n'}] \hookrightarrow [b_0, b_1, \ldots, b_n]$$

is given by a map $\iota : \{0, \ldots, n'\} \to \{0, \ldots, n\}$ such that

1. $\iota(i) \leq \iota(j)$ if and only if $i \leq j$ (order-preserving),
2. $a_i = b_{\iota(i)}$ (compatible),
3. ι is one-to-one (implying $n' \leq n$),

Although ι itself is a map of index sets, we use the compatibility property to overload notation and write $\iota : [a_0, \ldots, a_{n'}] \hookrightarrow [b_0, \ldots, b_n]$.

Remark 2.16 (Categorical interpretation). An attribute inclusion is a morphism $T \to T'$ in the category $\mathcal{A}$ such that $T' = T \circ \iota = \iota^* T$ where $\iota : \mathbf{n}' \to \mathbf{n}$ is a monomorphism in $\mathbf{\Delta_a}$. We overload notation (that is, omit the upper-star) and write $\iota : T' \hookrightarrow T$. The functor $\mathbf{\Delta_a} \to \mathcal{A}$ is contravariant, so attribute inclusions are actually epimorphisms in $\mathcal{A}$; however, it is reasonable to call them "inclusions"

[7]We are particularly indebted to Tony Falcone for technical discussions that motivated the formalism in this subsection.

because the $\mathbf{n}'$-ordered multiset $T'(\mathbf{n}')$ is an ordered subset of the $\mathbf{n}$-ordered multiset $T(\mathbf{n})$. One could avoid this overloaded notation by working in the opposite category, but we decline to add another layer of notation since the meaning is always clear in context.

Example 2.17. Consider $A = \{a, b, c, d\}$. Example attribute lists[8] are $T' = [a, a, a, c, d]$ and $T = [a, a, a, a, a, b, c, c, d]$. There are 20 possible inclusions $T' \hookrightarrow T$, which are obtained by choosing the ordered image of the a's and c's. One possible inclusion is

$$\iota = \{(0 \mapsto 0), (1 \mapsto 1), (2 \mapsto 3), (3 \mapsto 7), (4 \mapsto 8)\}, \tag{2.2}$$

which can be summarized as $\iota = [0, 1, 3, 7, 8]$. We can abbreviate this by decorating the entries in T that are included from T',

$$\iota : [a, a, a, c, d] \hookrightarrow [\underline{a}, \underline{a}, a, \underline{a}, a, b, c, \underline{c}, \underline{d}]. \tag{2.3}$$

Lemma 2.18 (Quotient inclusion)**.** *For any attribute inclusion $\iota : T' \hookrightarrow T$, there is an attribute list T/ι (called the* quotient*) that enumerates the entries of T that are not in the image of ι. This enumeration equips the quotient with an attribute inclusion $\iota^c : (T/\iota) \hookrightarrow T$, and ι^c corresponds to the complimentary monomorphism from Lemma* A.1*.*

Example 2.19. Consider the earlier example of an attribute inclusion. The quotient list is $T/\iota = [a, a, b, c]$. The quotient inclusion is

$$\iota^c : [a, a, b, c] \hookrightarrow [a, a, \underline{a}, a, \underline{a}, \underline{b}, \underline{c}, c, d]. \tag{2.4}$$

The next lemma and corollary make clear that face maps and attribute inclusions are related tightly.

Lemma 2.20. *Any face map $d_i : T \to d_i T$ in $\mathcal{A}$ is equipped with an attribute inclusion $d_i T \hookrightarrow T$ defined by the index function that skips i, namely the co-face monomorphism d^i in $\mathbf{\Delta}_\mathbf{a}$. Its quotient inclusion is the index function $\{0\} \to \{0, \dots, n\}$ by $0 \mapsto i$ that gives $[a_i] \hookrightarrow T$.*

Corollary 2.21. *For any attribute inclusion $\iota : T' \hookrightarrow T$, there is a sequence*[9] *of face maps $d_{j_0}, \dots, d_{j_k}$ such that $d_{j_0} \cdots d_{j_k} T = T'$ and such that the attribute inclusion induced by the sequence of face maps is ι. Moreover, any permutation of this sequence obtained by re-indexing the face-maps according to Lemma* 2.10 *is equivalent. If $j_0 \leq \cdots \leq j_k$, then j is the index function for ι^c, the quotient inclusion.*

Remark 2.22. In light of Theorem 2.14, Corollary 2.21 is a partial version of Lemma A.4, which says face maps and degeneracy maps generate all the morphisms in a simplicial set. This is because the co-face and co-degeneracy maps in $\mathbf{\Delta}_\mathbf{a}$ generate all order-preserving maps.

Attribute inclusions provide surjections on value spaces and measures, according to the following "contravariant" definition.

[8]The fact that these lists are in alphabetical order is merely aesthetic, and is not required in the definition of an attribute list.

[9]The backwards ordering here is intentional. Because of the indexing situation and Lemma 2.10, it is simpler to remove attributes from the end. To remove an entire list, one could write $d_0 d_1 \cdots d_n T$ or $d_0 d_0 \cdots d_0 T$, because d_0 is like "pop" on the front of the list.

Definition 2.23 (Reduction)**.** Consider an attribute inclusion $\iota : T' \hookrightarrow T$, where $T' = [a_0, \ldots, a_{n'}]$ and $T = [b_0, \ldots, b_n]$. Write an element of $\mathbb{V}(T)$ as $(x_0, \ldots, x_n)$ where $x_i \in \mathbb{V}(b_i)$ and write an element of $\mathbb{V}(T')$ as $(y_0, \ldots, y_{n'})$ where $y_j \in \mathbb{V}(a_j)$. Define the surjective function $\downarrow_\iota \colon \mathbb{V}(T) \twoheadrightarrow \mathbb{V}(T')$ by

$$(x_0, \ldots, x_n) \mapsto (y_0, \ldots, y_{n'}) = (x_{\iota(0)}, \ldots, x_{\iota(n')}).$$

Similarly, define the surjective function $\downarrow_\iota \colon \mathbf{M}(T) \twoheadrightarrow \mathbf{M}(T')$ by sequential application of face maps according to the previous corollary: For any $\tau \in \mathbf{M}(T)$, let

$$\downarrow_\iota (\tau)(U_0 \times \cdots \times U_{n'}) \coloneqq \tau(W_0 \times \cdots \times W_n), \text{ for } W_i = \begin{cases} U_j, & \text{if } i = \iota(j) \\ \mathbb{V}(a_i), & \text{otherwise.} \end{cases}$$

That is, $\downarrow_\iota \tau$ is the measure on $\mathbb{V}(T')$ obtained by marginalizing τ to remove the factors specified by ι^c.

When the attribute inclusion $\iota : T' \to T$ is understood from context, we abuse notation and write $\downarrow_{T'} \tau$ instead of $\downarrow_\iota \tau$. Note that $\downarrow_{[]} \tau = (d_0 \cdots d_n)\tau = \tau(\mathbb{V}(T)) = \int_{\mathbb{V}(T)} \tau$, so we use this notation as shorthand for "the total integral of a measure."

Definition 2.24 (Sum of attribute lists)**.** Given attribute lists T_1 and T_2 in $\mathcal{A}$, define $T_1 \oplus T_2$ as the attribute list obtained by concatenating T_1 and T_2.

Note that $T_1 \oplus T_2$ and $T_2 \oplus T_1$ are related by a permutation, which (excepting the identity permutation) does *not* correspond to a morphism in the categories $\mathcal{A}$ or $\mathbf{\Delta_a}$. The concatenation process provides specific attribute inclusions $T_1 \hookrightarrow T_1 \oplus T_2$ and $T_2 \hookrightarrow T_1 \oplus T_2$. More generally, for attribute inclusion $\iota : T' \to T$ as in Lemma 2.18, it is true that T and $T' \oplus (T/\iota)$ are related by a permutation; because, the concatenation provides inclusions $T' \hookrightarrow T$ and $(T/\iota) \hookrightarrow T$ that may *not* be the original ι and ι^c. On the other hand, for any sum $T = T_1 \oplus T_2$, it is true that T_2 is the quotient of T by the concatenation-induced inclusion of T_1, and vice-versa.

Definition 2.24 implies $\mathbb{V}(T_1 \oplus T_2) = \mathbb{V}(T_1) \times \mathbb{V}(T_2)$ and $\mathbf{M}(T_1 \oplus T_2)$ are well-defined. But, beware of multivariable calculus: $\mathbf{M}(T_1 \oplus T_2) \supsetneq \mathbf{M}(T_1) \times \mathbf{M}(T_2)$, as not every measure on a product space is an elementary product of measures!

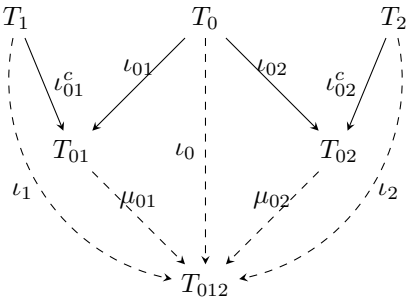
Example 2.25. Consider $T_1 = [a, a, a, c, d]$ and $T_2 = [a, a, b, c]$. Then their sum is $T_1 \oplus T_2 = [a, a, a, c, d, a, a, b, c]$. The concatenation is equipped with inclusions

$$\begin{aligned} \iota_1 = \iota_2^c : [a, a, a, c, d] &\hookrightarrow [\underline{a}, \underline{a}, \underline{a}, \underline{c}, \underline{d}, a, a, b, c] \\ \iota_2 = \iota_1^c : [a, a, b, c] &\hookrightarrow [a, a, a, c, d, \underline{a}, \underline{a}, \underline{b}, \underline{c}]. \end{aligned} \tag{2.5}$$

Definition 2.26 (Permutation notation)**.** Suppose that T_{12}, T_1, and T_2 are attribute lists such that $\varsigma(T_1 \oplus T_2) = T_{12}$ for a permutation ς. Then $\iota = \varsigma|_{T_1} : T_1 \to T_{12}$ and $\iota^c = \varsigma|_{T_2} : T_2 \to T_{12}$ are complimentary attribute inclusions. If the permutation or attribute inclusions are well-known in context, then for any subsets $U_1 \subseteq \mathbb{V}(T_1)$ and $U_2 \subseteq \mathbb{V}(T_2)$, let $U_1 \tilde{\times} U_2 \in \mathbb{V}(T_{12})$ denote the subset for which elements $x_1 \times x_2 \in U_1 \times U_2 \subseteq \mathbb{V}(T_1 \oplus T_2)$ and $x_1 \tilde{\times} x_2 \in U_1 \tilde{\times} U_2 \subseteq \mathbb{V}(T_{12})$ correspond with respect to the ς-permuted indices.

Because Lemma A.2 provides an ordered form of the inclusion–exclusion principle, we can define an indexed form of the inclusion–exclusion principle.

Definition 2.27 (Merge of attribute lists)**.** Suppose $T_0, T_{01}, T_{02} \in \mathcal{X}$, and that $\iota_{01} : T_0 \hookrightarrow T_{01}$ and $\iota_{02} : T_0 \hookrightarrow T_{02}$ are attribute inclusions. Define Merge$(T_{01}, T_{02}, \iota_{01} \sim \iota_{02})$ as the attribute list obtained by performing the index merge specified by Figure1 as in Lemma A.2; this merge concatenates sublists spliced between the entries aligned by $\iota_{01} \sim \iota_{02}$. Writing T_{012} for Merge$(T_{01}, T_{02}, \iota_{01} \sim \iota_{02})$, Diagram A.1 becomes a diagram of attribute inclusions.



(2.6)

Note that the choice of ordering in Definition 2.27 and Figure 1 is partially arbitrary. In particular, one may draw an equivalent diagram with any choice of interleaving pattern, as long as the T_0 entries remain fixed. However, this choice is irrelevant, as the theory developed in Section 3 will encompass all allowable permutations. Regarding the permutation notation introduced earlier, for any Borel sets $U_0 \subseteq \mathbb{V}(T_0)$, $U_1 \subseteq \mathbb{V}(T_1)$, and $U_2 \subseteq \mathbb{V}(T_2)$, we write $U_0 \tilde{\times} U_1 \tilde{\times} U_2 \subseteq \mathbb{V}(T_{012})$ for the appropriately permuted copy of the set $U_0 \times U_1 \times U_2$ in $\mathbb{V}(T_0 \oplus T_1 \oplus T_2)$, since the definition and algorithm give a well-defined permutation. This $\tilde{\times}$ notation is required in Theorem 3.11 and elsewhere.

Example 2.28. Compare this to Example A.3. Consider $A = \{a, b, c, d\}$. Consider attribute lists $T_{01} = [a, a, a, a, b, c]$ and $T_{02} = [a, a, b, b, d]$, and $T_0 = [a, b]$ with the attribute inclusions $\iota_{01} = [1, 4]$ and $\iota_{02} = [1, 3]$. Visually, the merged indexing means

$$\iota_{01} : [a, b] \mapsto [a, \underline{a}, a, a, \underline{b}, c]$$
$$\iota_{02} : [a, b] \mapsto [a, \underline{a}, b, \underline{b}, d]$$
yields
$$\iota_0 : [a, b] \mapsto [a, a, \underline{a}, a, a, b, \underline{b}, c, d]$$
$$\mu_{01} : [a, a, a, a, b, c] \mapsto [\underline{a}, a, \underline{a}, \underline{a}, \underline{a}, b, \underline{b}, \underline{c}, d]$$
$$\mu_{02} : [a, a, b, b, d] \mapsto [a, \underline{a}, \underline{a}, a, a, \underline{b}, \underline{b}, c, \underline{d}]$$

Our choice of ordering in Merge() provides that the trivial merge

$$\text{Merge}(T_{01}, T_{02}, []) = T_{01} \oplus T_{02}$$

is the sum from Definition 2.24.

As with Definition 2.24, the attribute list Merge$(T_{01}, T_{02}, \iota_{01} \sim \iota_{02})$ is well-defined regardless of the preference of T_{01} versus T_{02} and regardless of the indices specified by ι_{01} and ι_{02}. This list is identical to the list obtained by constructing the sum $T_{01} \oplus T_{02}$ then applying face maps to remove the image of $d_{\iota_{02}(i)}$ for each i indexing T_0. But, again beware that the partitioned merge-sort construction equips

T_{012} with specific attribute inclusions $T_{01} \hookrightarrow T_{012}$ and $T_{02} \hookrightarrow T_{012}$ such that the composed attribute inclusion $T_0 \hookrightarrow T_{012}$ is well-defined through both compositions. In general, these inclusions are not the same as the inclusions obtained through the "sum and face" construction.

Lemma 2.29 (Decomposition of merged lists)**.** *Suppose $T_0, T_{01}, T_{02} \in \mathcal{X}$, and that $\iota_{01} : T_0 \hookrightarrow T_{01}$ and $\iota_{02} : T_0 \hookrightarrow T_{02}$ are attribute inclusions. Let T_{012} denote* $\mathrm{Merge}(T_{01}, T_{02}, \iota_{01} \sim \iota_{02})$. *Let $\iota_{01}^c : T_1 \hookrightarrow T_{01}$ and $\iota_{02}^c : T_2 \hookrightarrow T_{02}$ denote the complements of these inclusions, so $T_1 \coloneqq T_{01}/\iota_{01}$ and $T_2 \coloneqq T_{02}/\iota_{02}$.*

Then T_{012} is partitioned by the inclusions $\iota_0 : T_0 \hookrightarrow T_{012}$, $\iota_1 : T_1 \hookrightarrow T_{012}$, and $\iota_2 : T_2 \hookrightarrow T_{012}$. That is, $\mathbb{V}(T_{012}) = \mathbb{V}(T_0) \tilde{\times} \mathbb{V}(T_1) \tilde{\times} \mathbb{V}(T_2)$ is a permutation of $\mathbb{V}(T_0) \times \mathbb{V}(T_1) \times \mathbb{V}(T_2)$, with ordering of projections determined by the inclusion and quotient maps, $\downarrow_{\iota_0}$, $\downarrow_{\iota_1}$, and $\downarrow_{\iota_2}$.

2(d). **Data sections.** Because the forgetful map $p : \mathcal{X} \to \mathcal{A}$ acts like a projection, it allows a notion of section.

Definition 2.30 (Data section)**.** Consider a data subcomplex $p : \mathcal{X}' \to \mathcal{A}'$ of an ambient $p : \mathcal{X} \to \mathcal{A}$. A *data section* is a natural[10] map $\sigma : \mathcal{A}' \to \mathcal{X}'$ such that $p \circ \sigma = 1_{\mathcal{A}'}$.

Remark 2.31. In Section 4, data sections will be specified as $\sigma : \mathcal{A}'_n \to \mathcal{X}'_n$, on a single level of the simplicial-set grading, where the other levels are inferred by the face and degeneracy maps. This omits all nondegenerate elements of level $n+1$, so is interpreted as a section on the n-skeleton.

The following definition captures a condition describing data subcomplexes that are "as compatible as possible."

Definition 2.32 (Well-aligned)**.** A data subcomplex $\mathcal{X}'$ of an ambient $\mathcal{X}$ is called *well-aligned* if: for all $(T_{01}, \tau_{01}), (T_{02}, \tau_{02}) \in \mathcal{X}'$ and all T_0 with attribute inclusions $\iota_{01} : T_0 \hookrightarrow T_{01}$ and $\iota_{01} : T_0 \hookrightarrow T_{02}$, there exists $(T_0, \tau_0) \in \mathcal{X}'$ with

$$\downarrow_{\iota_{01}} \tau_{01} = \tau_0 = \downarrow_{\iota_{02}} \tau_{02}.$$

The next lemma shows that well-aligned data subcomplexes in this theory play the role of "submanifolds transverse to the fiber" from classical bundle theory and of "holonomic submanifolds" in geometric PDE theory. That is, they represent local sections.

Lemma 2.33. *Suppose that $p : \mathcal{X}' \to \mathcal{A}'$ is a data subcomplex of an ambient data complex $p : \mathcal{X} \to \mathcal{A}$ such that $\mathcal{X}'$ contains a nontrivial data table. The following are equivalent:*

1. *$\mathcal{X}'$ is well-aligned.*
2. *There is a data section $\sigma : \mathcal{A}' \to \mathcal{X}'$ such that $\sigma(\mathcal{A}') = \mathcal{X}'$.*
3. *$p : \mathcal{X}' \to \mathcal{A}'$ is a simple cover via the isomorphism p.*

Proof. (2) implies (1): Note that well-alignedness is implied by the commutation of σ with the face maps.

(1) implies (2): The case of $T_0 = []$ implies that all data tables in $\mathcal{X}'$ have the same mass, M, which is non-zero since $\mathcal{X}'$ contains at least one non-trivial table. The case of $T_{01} = T_{02} = T_0 = T$ implies that each $T \in \mathcal{A}'$ admits exactly one $(T, \tau) \in \mathcal{X}'$.

[10]Natural means that it respects the face and degeneracy maps, as in (A.2) and Lemma A.4.

It is immediate that (2) and (3) are equivalent. □

Remark 2.34. A database engineer would appreciate a database system that could be described as a well-aligned data subcomplex, because for each list of columns present within any combination of the given tables, there is only one possible table; that is, for each T there is exactly one (T, τ). Compare the well-aligned condition to the space of joins, Definition 3.5. Note too that well-aligned implies finitely generated. (Not every finitely-generated data subcomplex is well-aligned, as it could have multiple data tables over the same attribute lists.) Moreover, if $\mathcal{X}'$ is well-aligned (and contains a nontrivial data table), then all data tables can be re-scaled by their shared mass M to yield probability measures.

In our definitions above, the set A of attributes is finite, and each level $\mathcal{A}_n$ of the simplicial set $\mathcal{A}$ is finite. Therefore, for any simplicial subset $\mathcal{A}' \subseteq \mathcal{A}$, we can consider the finite graph whose vertices are the 0-cells of $\mathcal{A}'_0$ (singleton attribute lists) and whose edges are the 1-cells $\mathcal{A}'_1$ (including loops, as degenerate 1-cells like $[a, a]$).

Definition 2.35 (Connected)**.** Suppose that $p : \mathcal{X}' \to \mathcal{A}'$ is a data subcomplex of an ambient data complex $p : \mathcal{X} \to \mathcal{A}$ such that $\mathcal{X}'$ contains a nontrivial data table. The simplicial set $\mathcal{A}'$ of attributes is called *connected* if the finite graph with vertices $\mathcal{A}'_0$ and edges $\mathcal{A}'_1$ is connected.

Definition 2.36 (Path-connected)**.** A data subcomplex $\mathcal{X}'$ of an ambient $\mathcal{X}$ is called *path-connected* if for any attributes a, b in $\mathcal{A}'$, there is a sequence (T_{01}, τ_{01}), $(T_{12}, \tau_{12}), \ldots, (T_{k-1,k}, \tau_{k-1,k})$ in $\mathcal{X}'$ such that $a \in T_{0,1}$ and $b \in T_{k-1,k}$ and for all $i = 1 \ldots, k$ there is an attribute list $T_i \neq []$ equipped with inclusions $T_i \hookrightarrow T_{i-1,i}$ and $T_i \hookrightarrow T_{i,i+1}$ such that

$$\downarrow_{T_i} \tau_{i-1,i} = \downarrow_{T_i} \tau_{i,i+1}.$$

Lemma 2.37. *Suppose $\mathcal{A}'$ is connected as a simplicial set. If $\mathcal{X}'$ is well-aligned, then $\mathcal{X}'$ is path-connected.*

With the language of simplicial sets, we can now re-state our original motivating questions. The remaining sections of this document construct a precise way to answer these questions, and ensure that the notion of "distance" is well-defined. An appropriate notion of distance appears in Definition 4.1. When all the definitions and lemmas are in place, these problems are answered by the Obstruction Cocycle in Definition 4.8.

Problem 2.38 (Testing problem, bis)**.** Consider a data subcomplex $p : \mathcal{S} \to \mathcal{B}$ of an ambient $p : \mathcal{X} \to \mathcal{A}$. Given a data section $\sigma^+ : \mathcal{A}_{n+1} \to \mathcal{X}_{n+1}$ of the form $\sigma^+ : T^+ \mapsto (T^+, \sigma^+)$ for $T \in \mathcal{A}_{n+1}$, is it true that $\partial\sigma^+$ lies entirely within $\mathcal{S}_n$? If not, what is the distance from $\partial\sigma^+$ to $\mathcal{S}_n$ in $\mathcal{X}_n$?

Problem 2.39 (Merging and meta-merging problems, bis)**.** Consider a data subcomplex $p : \mathcal{S} \to \mathcal{B}$ of an ambient $p : \mathcal{X} \to \mathcal{A}$. Suppose that there is a simplicial map $\sigma : \mathcal{B}_n \to \mathcal{S}_n$ of the form $\sigma : T \mapsto (T, \sigma)$. Does there exist an extension $\sigma^+ : \mathcal{A}_{n+1} \to \mathcal{X}_{n+1}$ of σ, meaning $\partial\sigma^+(T^+) = \sigma(\partial T^+)$ for all $T^+ \in \mathcal{A}_{n+1}$ such that $\partial T^+ \in \mathcal{B}_n$? If not, what is the minimal distance that would allow an approximate extension?

2(e). **Homology.** We use the traditional definition of chains, summarized here to fix notation. Fix[11] a ring R. For $k \geq 0$, a k-chain $Y \in C_k(\mathcal{A}, R) \coloneqq C_k(\mathcal{A}, \mathbb{Z}) \otimes R$ is a formal linear combination

$$Y = \sum_j r_j T_j, \quad r_j \in R,\ T_j \in \mathcal{A}_k, \tag{2.7}$$

where negative coefficients indicate formally reversed orientation. We define (-1)-graded chains as elements of the 1-dimensional R-module, $C_{-1}(\mathcal{A}) = R \cdot \{[]\} \cong R$. For $k \geq 0$, define the usual simplicial boundary operator $C_k(\mathcal{A}, R) \to C_{k-1}(\mathcal{A}, R)$, as

$$\begin{aligned} \partial &\coloneqq \sum_{i=0}^{k} (-1)^i d_i \\ \partial &: [a_0, \ldots, a_k] \mapsto \sum_{i=0}^{k} (-1)^i [a_0, \ldots, \hat{a_i}, \ldots, a_k]. \end{aligned} \tag{2.8}$$

It is immediate that $\partial : C_k(\mathcal{A}, R) \to C_{k-1}(\mathcal{A}, R)$ satisfies $\partial^2 = 0$, so homology $H_\bullet(\mathcal{A}, R)$ is well-defined.

For $k \geq 0$, a k-chain $(Y, \psi) \in C_k(\mathcal{X}, R) \coloneqq C_k(\mathcal{X}, \mathbb{Z}) \otimes R$ is a formal linear combination

$$(Y, \psi) = \sum_j r_j (T_j, \tau_j), \quad r_j \in R,\ T_j \in \mathcal{A}_k,\ \tau_j \in \mathbf{M}(T_j), \tag{2.9}$$

where negative coefficients indicate formally reversed orientation. We can also define (-1)-graded chains as $C_{-1}(\mathcal{X}, R)$, the R-module generated by $\mathbf{M}([]) = R_{\geq 0}$. Moreover, for any $(T, \tau) \in \mathcal{X}_k$, note that $r(T, \tau)$ and $(T, r\tau)$ are formally distinct unless $r = 1_R$; hence, the graded module $C_\bullet(\mathcal{S}, R)$ is *very large*, especially if $\mathbb{V}(a)$ is infinite for any $a \in A$. For $k \geq 0$, define the usual simplicial boundary operator $C_k(\mathcal{X}, R) \to C_{k-1}(\mathcal{X}, R)$, as

$$\begin{aligned} \partial &\coloneqq \sum_{i=0}^{k} (-1)^i d_i \\ \partial &: ([a_0, \ldots, a_k], \tau) \mapsto \sum_{i=0}^{k} (-1)^i ([a_0, \ldots, \hat{a_i}, \ldots, a_k], \downarrow_{[a_0, \ldots, \hat{a_i}, \ldots, a_k]} \tau). \end{aligned} \tag{2.10}$$

The next lemma is easy, but important; it means the usual notions of *cycle/closed* and *boundary/exact* apply to chains in $\mathcal{X}$.

Lemma 2.40. *$\partial : C_k(\mathcal{X}, R) \to C_{k-1}(\mathcal{X}, R)$ satisfies $\partial^2 = 0$. In particular, the homology $H_\bullet(\mathcal{X}, R)$ is well-defined.*

Proof. Suppose that $T = [a_0, a_1, \ldots, a_k]$ and that $\tau \in \mathbf{M}(T)$. Recall that $\mathbb{V}(T)$ is a product of the attributes' measurable spaces, and by our definitions, the measure $(\mathbb{V}(T), \tau)$ is finite, therefore σ-finite, so the Fubini–Tonelli theorem holds. In particular, Corollary 2.21 shows that the reduction $\downarrow_{[a_0, \ldots, \hat{a_i}, \ldots, \hat{a_j}, \ldots, a_k]}$ is symmetric, so because the double-sum is alternating, all terms will cancel.

[11] For practical reasons we use $R = \mathbb{Z}/2\mathbb{Z} = \mathbb{F}_2$ in applications; however, chains are sensible for any ring.

For example, suppose $\tau = \tau_{0123}$ on $T = [a_0, a_1, a_2, a_3]$, where the index shows which variables are still free. Then

(2.11)
$$\begin{aligned}\partial^2\tau_{0123} &= \partial(\tau_{123} - \tau_{023} + \tau_{013} - \tau_{012}) \\ &= (\tau_{23} - \tau_{13} + \tau_{12}) - (\tau_{23} - \tau_{03} + \tau_{02}) + (\tau_{13} - \tau_{03} + \tau_{01}) - (\tau_{12} - \tau_{02} + \tau_{01}) \\ &= 0 \in C_{k-2}(\mathcal{X}).\end{aligned}$$

Note that it is irrelevant in this proof whether T is a degenerate attribute list, as repeated attribute value spaces are treated as distinct factors for the sake of integration. □

Define the projection $p : C_k(\mathcal{X}, R) \to C_k(\mathcal{A}, R)$ by $p(T, \tau) \coloneqq T$ and extending by linearity.

Lemma 2.41. $\partial \circ p = p \circ \partial$

Proof. Suppose that $T = [a_0, a_1, \ldots, a_k]$ and that $\tau \in \mathbf{M}(T)$. Then
(2.12)
$$p\partial(T, \tau) = p\sum(-1)^i d_i(T, \tau) = \sum(-1)^i p(d_i T, d_i \tau) = \sum(-1)^i d_i T = \partial T = \partial p(T, \tau).$$

□

Corollary 2.42. *The induced homomorphism* $[p] : H_\bullet(\mathcal{X}, R) \to H_\bullet(\mathcal{A}, R)$ *is well-defined.*

Similarly, the chain modules and homology are well-defined for any data subcomplex $p : \mathcal{X}' \to \mathcal{A}'$ of an ambient $p : \mathcal{X} \to \mathcal{A}$.

Corollary 2.43. *If* $p : \mathcal{X}' \to \mathcal{A}'$ *is well-aligned, then there are canonical isomorphisms* $C_\bullet(\mathcal{X}', R) \cong C_\bullet(\mathcal{A}', R)$ *and* $H_\bullet(\mathcal{X}', R) \cong H_\bullet(\mathcal{A}', R)$ *induced by* p.

We are particularly interested in the case $R = \mathbb{Z}/2\mathbb{Z}$, so that a chain $C_\bullet(\mathcal{A}, \mathbb{Z}/2\mathbb{Z})$ (respectively $C_\bullet(\mathcal{X}, \mathbb{Z}/2\mathbb{Z})$) is interpreted as a set of attribute lists (respectively, data tables), without any consideration for multiplicity or orientation. It is therefore sensible to apply the condition *well-aligned* to a chain $(Y, \psi) \in C_n(\mathcal{X}', \mathbb{Z}/2\mathbb{Z})$, so that a well-aligned chain in $(Y, \psi) \in C_n(\mathcal{X}, \mathbb{Z}/2\mathbb{Z})$ can be interpreted equivalently to a section $\sigma : p(\mathcal{X}') \to \mathcal{X}'$, where $\mathcal{X}'$ is the data subcomplex generated by the elements of ψ.

3. Homotopy as joins

In the previous section, we established that a data complex is equipped with simplicial homology, and framed data complexes as simplicial sets. This section contains several payoffs for that effort. First, Section 3(a) builds to Theorem 3.11, our first key result, which shows that the simplicial set language enables a connection between our framework and standard database engineering; later results show that the framework enables further insights into data merging problems that transcend standard database engineering. Then, Section 3(b) explores the simplicial homotopy of data complexes and reframes Problems 2.38 and 2.39 in the language of obstruction theory for simplicial sets, as in [7, 9, 10, 12, 18].

3(a). **Database joins and the kan condition.** Recall these three standard definitions from the well-established theory of simplicial sets, as in Appendix A and [7, 9, 10, 12, 18].

Definition 3.1 (Simplex)**.** The standard n-simplex Δ^n is the simplicial set generated (via face and degeneracy maps) by the ordered set $\mathbf{n} = \{0, \ldots, n\}$ in the simplex category $\mathbf{\Delta}$.

Definition 3.2 (Horn)**.** The kth *horn* Λ^n_k of the n-simplex Δ^n is the simplicial subset generated by the union of all the faces of Δ^n except the kth face.

As is standard in the literature, we abuse notation slightly by referring to both $\Delta^n \to X$ (which is an infinite collection of sets) and $\mathbf{n} \mapsto x \in X_n$ (which is the generator of that collection) as "an n-simplex in X."

Note! The (categorical) n-simplex Δ^n is *not* the same as the (topological) n-simplex $|\Delta^n|$. The former is an infinite set of formal objects in the simplex category; it has no notion of "interior" or "continuity." The latter is a compact topological space obtained defined via convex linear combinations in $\mathbb{R}^{n+1}$. There is a relationship between their respective categories called *realization*, as discussed in [18, §3] and [9, Chap I.2].

Remark 3.3 (Data tables as simplices). A data subcomplex $\mathcal{X}'$ in an ambient $\mathcal{X}$ is an (augmented) simplicial set by Theorem 2.14. Thus, a data table (T, τ) with $T = [a_0, \ldots, a_n]$ can be seen as (the generator of) an n-simplex, which includes its faces $d_0(T, \tau), \ldots, d_n(T, \tau)$ and degeneracies $s_0(T, \tau), \ldots, s_n(T, \tau)$, and so-on. The $n{+}1$ "vertices" are (generated by) the single-attribute data tables $(T_0, \tau_0), \ldots, (T_n, \tau_n)$ obtained by applying sequences of n face maps. For $m \leq n$, the m-simplices within (T, τ) are (generated by) the data tables obtained by applying sequences of face maps and degeneracy maps until the result has $m{+}1$ attributes. The picture of "two simplices that share a boundary component" is realized in $\mathcal{X}'$ as a pair of data tables (T_{01}, τ_{01}) and (T_{02}, τ_{02}) and attribute inclusions $\iota_{01} : T_0 \to T_{01}$ and $\iota_{02} : T_0 \to T_{02}$ such that there is a data table (T_0, τ_0) with $\downarrow_{\iota_{01}} \tau_{01} = \tau_0 = \downarrow_{\iota_{02}} \tau_{02}$. If $\text{len}(T_{01}) = \text{len}(T_{02}) = 2$ and $\text{len}(T_0) = 1$, then this information generates a 2-horn. A completion of the 2-horn to a 2-simplex would be (generated by) a data table (T_{012}, τ_{012}) that has (T_{01}, τ_{01}) and (T_{02}, τ_{02}) as two of its three faces. Depending on the available simplices in $\mathcal{X}'$, it may or may not be possible to find such (T_{012}, τ_{012}).

Definition 3.4 (Kan condition)**.** A simplicial set X is said to satisfy the *Kan condition* iff any map from a horn $\Lambda^n_k \to X$ extends to a compatible map from the simplex, $\Lambda^n_k \hookrightarrow \Delta^n \to X$.

The Kan condition means that the simplicial set is closed under simplicial deformation, so it has a well-defined homotopy group. The Kan condition is not specific to data complexes; it is a definition for general simplicial sets, and gives the appropriate notion of *fibrant* for many model categories. For our purposes, we require a slight variation on the Kan condition to provide an adequate notion of *fibrant* data contexts, which we now develop as Definition 3.13.

We define the space of joins for two data tables with designated attribute inclusions.

Definition 3.5 (Space of joins)**.** Suppose attribute lists $T_0, T_{01}, T_{02} \in \mathcal{A}$ are equipped with attribute inclusions $\iota_{01} : T_0 \to T_{01}$ and $\iota_{02} : T_0 \to T_{02}$, and let

T_{012} denote the attribute list $\mathrm{Merge}(T_{01}, T_{02}, \iota_{01} \sim \iota_{02})$ as in Definition 2.27, which is equipped with inclusions $\iota'_{01} : T_{01} \hookrightarrow T_{012}$ and $\iota'_{02} : T_{02} \hookrightarrow T_{012}$. For any data tables $(T_{01}, \tau_{01}), (T_{02}, \tau_{02}) \in \mathcal{X}$, let

$$\mathrm{Joins}(\tau_{01}, \tau_{02}, \iota_{01} \sim \iota_{02}) \coloneqq \{(T_{012}, \tau_{012}) \in \mathcal{X} : \downarrow_{\iota'_{01}} \tau_{012} = \tau_{01},\ \downarrow_{\iota'_{02}} \tau_{012} = \tau_{02}\}. \tag{3.1}$$

Note that "$\mathrm{Joins}(\tau_{01}, \tau_{02}, \iota_{01} \sim \iota_{02}) \neq \emptyset$" requires $\downarrow_{\iota'_{01} \circ \iota_{01}} \tau_{01} = \downarrow_{\iota'_{01} \circ \iota_{01}} \tau_{02}$, as $\downarrow_{T_0} \tau_{012}$ must be well-defined.

Similarly to Definition 2.27, we write this set as $\mathrm{Joins}(\tau_{01}, \tau_{02}, T_0)$ for notational convenience when the attribute inclusions are understood from context.

Note! This is *not* the same notion of "join" that one sees in traditional topology, or in categorical references such as [5, 19] and `https://ncatlab.org/nlab/show/join+of+simplicial+sets`. It is not yet clear whether there is a useful relationship to joins in ergodic theory [8]. We choose the term "join" to mimic the terminology in database engineering discussed in Section 1. Definition 3.5 reminds one of couplings from statistics, as in [3]; however, the generality here allows repetition and distinct measures and overlaps.

Definition 3.6 (Join conditions)**.** A data subcomplex $p : \mathcal{X}' \to \mathcal{A}'$ of an ambient $p : \mathcal{X} \to \mathcal{A}$ is said to satisfy the *weak join condition* iff, for any (T_{01}, τ_{01}) and $(T_{02}, \tau_{02}) \in \mathcal{X}'$ with attribute inclusions $T_0 \hookrightarrow T_{01}$ and $T_0 \hookrightarrow T_{02}$ and $\downarrow_{T_0} \tau_{01} = \downarrow_{T_0} \tau_{02}$, then $\mathrm{Joins}(\tau_{01}, \tau_{02}, T_0) \cap \mathcal{X}'$ is nonempty. It satisfies the *strong join condition* iff $\mathrm{Joins}(\tau_{01}, \tau_{02}, T_0) \subset \mathcal{X}'$.

The weak join condition means that the simplicial set admits some database JOIN operation between any well-aligned pair of data tables. The strong join condition requires that *all possible* joins exist in $\mathcal{X}'$.

The trivial join (when $T_0 = []$) is of particular interest as it provides a generalization of independent products of measures.

Definition 3.7 (Admission of trivial joins)**.** A data subcomplex $p : \mathcal{X}' \to \mathcal{A}'$ of an ambient $p : \mathcal{X} \to \mathcal{A}$ is said to *admit trivial joins* iff, for every $(T_1, \tau_1), (T_2, \tau_2) \in \mathcal{X}'$ with $\downarrow_{[]} \tau_1 = \downarrow_{[]} \tau_2 = M$, there exists some $\tau_{12} \in T_1 \oplus T_2$ such that $\downarrow_{T_1} \tau_{12} = \tau_1$ and $\downarrow_{T_2} \tau_{12} = \tau_2$.

Definition 3.8 (Closure under independent products)**.** A data subcomplex $p : \mathcal{X}' \to \mathcal{A}'$ of an ambient $p : \mathcal{X} \to \mathcal{A}$ is said to be *closed under independent products* iff, for every $(T_1, \tau_1), (T_2, \tau_2) \in \mathcal{X}'$ and $\downarrow_{[]} \tau_1 = \downarrow_{[]} \tau_2 = M$, we have $(T_1 \oplus T_2, \frac{\tau_1 \tau_2}{M}) \in \mathcal{X}'$.

Remark 3.9. The independent product is an example of a trivial join. If a data subcomplex is closed under independent products, then it also includes all IID measures built from its various data tables; this property is important for applications to statistics.

Lemma 3.10. *If a data subcomplex $p : \mathcal{X}' \to \mathcal{A}'$ of an ambient $p : \mathcal{X} \to \mathcal{A}$ satisfies the strong join condition, then $p : \mathcal{X}' \to \mathcal{A}'$ is closed under permutations.*

Proof. Because A is finite, it suffices to prove that $\mathcal{X}'$ is closed under permutations that are swaps (that is, transpositions or 2-cycles). Moreover, it suffices to consider only swaps of adjacent entries, as any swap $i \leftrightarrow j$ can be written by migration of j past i, then i to the original position of j.

Suppose that $(T, \tau) \in \mathcal{X}'$ with $T = [a_0, \ldots, a_i, a_j, \ldots, a_n]$ with $j = i + 1$. Then, consider $(T_{01}, \tau_{01}) = d_i(T, \tau)$ and $(T_{02}, \tau_{02}) = d_i(T, \tau)$, and $(T_0, \tau_0) = d_j d_i(T, \tau)$. By construction, there are well-defined attribute inclusions $T_0 \hookrightarrow T_{01}$ and $T_0 \hookrightarrow T_{01}$ that satisfy $\downarrow_{T_0} \tau_{01} = d_j d_i \tau = d_{j-1} d_i \tau = \downarrow_{T_0} \tau_{02}$. Note that $T = \text{Merge}(T_{01}, T_{02}, T_0)$, and $\tau \in \text{Joins}(\tau_{01}, \tau_{02}, \tau)$. Consider the attribute list $\tilde{T} = [a_0, \ldots, a_j, a_i, \ldots, a_n]$ obtained by swapping the adjacent entries a_i and a_j, and let $\tilde{\tau}$ denote the correspondingly permuted measure obtained from τ. Note that $\tilde{T} = \text{Merge}(T_{02}, T_{01}, T_0)$, and $\tilde{\tau} \in \text{Joins}(\tau_{02}, \tau_{01}, T_0)$. By the strong join condition, $\tilde{\tau} \in \mathcal{X}'$. □

Theorem 3.11 (Fundamental theorem of data complexes)**.** *For any data subcomplex $p : \mathcal{X}' \to \mathcal{A}'$ of an ambient $p : \mathcal{X} \to \mathcal{A}$.*

1. *If $\mathcal{X}'$ satisfies the strong join condition, then $\mathcal{X}'$ admits trivial joins and $\mathcal{X}'$ satisfies the Kan condition as a simplicial set.*
2. *If $\mathcal{X}'$ admits trivial joins and $\mathcal{X}'$ satisfies the Kan condition, then $\mathcal{X}'$ satisfies the weak join condition.*

(We would not be surprised if the Kan condition and the strong join condition are equivalent, under some reasonable assumptions, but we have not pursued that claim.)

Proof of (2). Suppose that $\mathcal{X}'$ satisfies the Kan condition and admits trivial joins. Admission of trivial joins provides the weak join condition in the case $T_0 = []$. Suppose that (T_{01}, τ_{01}) and (T_{02}, τ_{02}) are elements of $\mathcal{X}'$. Suppose that there are inclusions $\iota_{01} : T_0 \to T_{01}$ and $\iota_{02} : T_0 \to T_{02}$ for some T_0, and suppose that $\downarrow_{T_0} \tau_{01} = \downarrow_{T_0} \tau_{02}$. Each of (T_{01}, τ_{01}) and (T_{02}, τ_{02}) and (T_0, τ_0) provides all faces of all lower dimensions. We prove the existence of $(T_{012}, \tau_{012}) \in \text{Joins}(\tau_{01}, \tau_{02}, T_0)$ by induction on the dimension. For simplicity, we use the language of simplicial sets, instead of the language of measures. Recall that a "vertex" is a data table obtained by marginalizing to a single attribute, and an n-simplex is a data table obtained by marginalizing to $n+1$ attributes, as in Remark 3.3. Fix a preferred vertex k in (T_0, τ_0). For any vertex i in (T_{01}, τ_{01}) and any vertex j in (T_{02}, τ_{02}), the 1-simplices $[i, k]$ and $[k, j]$ exist *a priori* (up to notational ordering). This is an example of a horn Λ^2_k. Therefore, by the Kan condition, the 2-simplex $[i, k, j]$ exists in $\mathcal{X}'$. Hence, every 2-face including k and vertices in (T_{01}, τ_{01}) or (T_{02}, τ_{02}) exists in $\mathcal{X}'$. Assume for induction that every n-face containing vertex k exists. Any n of those n-faces form a horn Λ^{n+1}_k, so their $(n+1)$-face exists in $\mathcal{X}'$. So, every $(n+1)$-face containing vertex k exists in $\mathcal{X}'$. Therefore, there is a Data Table (T_{012}, τ_{012}) that involves all vertices in (T_{01}, τ_{01}) and (T_{02}, τ_{02}). □

Proof of (1). We prove part (1) under the notable assumption that $\mathbb{V}(a)$ is a compact metric space for all $a \in A$. Hence, for any attribute list T, the space of measures $\mathbf{M}(T)$ includes a uniform[12] probability measure κ_T.

Suppose that $\mathcal{X}'$ satisfies the strong join condition. The case $T_0 = []$ implies admission of trivial joins.

In this proof, we assume that $k = 0$ is the common vertex in a horn Λ^n_k, but that is only for notational simplicity; the proof certainly applies for any other specified vertex k, by appropriate re-ordering. Consider data tables giving a horn

[12]That is, $\kappa_T(B_r(x))$ depends only on r, for metric balls $B_r(x)$ of sufficiently small radius.

Λ_0^n. These data tables are of the form $(T_{\widehat{m}}, \tau_{\widehat{m}})$ for $1 \leq m \leq n$, where $T_{\widehat{m}} = [a_0, \ldots, a_{m-1}, a_{m+1}, \ldots, a_n]$. Let $T = [a_0, \ldots, a_n]$. As a horn, these data tables are well-aligned; that is, they match on all corresponding faces according to $d_i\tau_{\widehat{j}} = d_{j-1}\tau_{\widehat{i}}$ for $i < j$ as noted after Definition A.7. In particular, all these data tables share the same total mass, M. To establish the Kan condition, we construct a compatible n-simplex; that is, a data table (T, τ) such that $d_m(T, \tau) = (T_{\widehat{m}}, \tau_{\widehat{m}})$ for $1 \leq m \leq n$.

A brief outline of the argument: First, we construct a data table $(T, \tau^{(n-1)})$ such that $d_{n-1}\tau^{(n-1)} = \tau_{\widehat{n-1}}$ and $d_n\tau^{(n-1)} = \tau_{\widehat{n}}$. The measure $\tau^{(n-1)}$ is built from a parameterized family of partial measures on $\mathbb{V}(T_{n-1} \oplus T_n)$ by recursively bifurcating the parameter set $\mathbb{V}(T_{0\cdots(n-2)})$ into dyadic sets, which allows $\tau^{(n-1)}$ to be defined via countable disjoint unions. This data table $\tau^{(n-1)}$ serves as the base case for an inductive argument for a sequence of partial solutions $\tau^{(n-1)}, \ldots, \tau^{(m)}, \ldots, \tau^{(1)}$ such that $\tau^{(m)}$ has the desired faces d_m through d_n. Finally, $(T, \tau^{(1)})$ is the desired n-simplex. Let us proceed.

For $K = \mathbb{V}(T_{0\cdots(n-2)})$, a compact set, consider the measures

$$\begin{cases} \mu_{K,n-1} : & U_{n-1} \mapsto \tau_{\widehat{n}}(K \tilde{\times} U_{n-1}) \text{ in } \mathbf{M}(T_{n-1}), \\ \mu_{K,n} : & U_n \mapsto \tau_{\widehat{n-1}}(K \tilde{\times} U_n) \text{ in } \mathbf{M}(T_n). \end{cases} \tag{3.2}$$

Consider any trivial join

$$t_K \in \text{Joins}(\mu_{K,n-1}, \mu_{K,n}, []) \subset \mathbf{M}(T_{n-1} \oplus T_n) = \mathbf{M}(\text{Merge}(T_{n-1}, T_n, []));$$

that is, $\downarrow_{T_{n-1}} t_K(U_{n-1}) = \tau_{\widehat{n-1}}(K \tilde{\times} U_{n-1})$ and $\downarrow_{T_n} t_K(U_n) = \tau_{\widehat{n-1}}(K \tilde{\times} U_n)$. Note that

$$\downarrow_{[]} t_K = d_n\tau_{\widehat{n-1}}(K) = d_n\tau_{\widehat{n}}(K) = \tau_{0\cdots(n-2)}(K) = M. \tag{3.3}$$

Fix any open $W \subset K$ such[13] that $\kappa_{T_{0\cdots(n-2)}}(W) = \frac{1}{2}\kappa_{T_{0\cdots(n-2)}}(K)$. Note that the measures

$$\begin{cases} \mu_{W,n-1} : & U_{n-1} \mapsto \tau_{\widehat{n}}(W \tilde{\times} U_{n-1}) \text{ in } \mathbf{M}(T_{n-1}), \\ \mu_{W,n} : & U_n \mapsto \tau_{\widehat{n-1}}(W \tilde{\times} U_n) \text{ in } \mathbf{M}(T_n) \end{cases} \tag{3.4}$$

can be joined to provide

$$t_W \in \text{Joins}(\mu_{W,n-1}, \mu_{W,n}, []) \subset \mathbf{M}(T_{n-1} \oplus T_n) = \mathbf{M}(\text{Merge}(T_{n-1}, T_n, [])).$$

That is, $\downarrow_{T_{n-1}} t_W(U_{n-1}) = \tau_{\widehat{n-1}}(W \tilde{\times} U_{n-1})$ and $\downarrow_{T_n} t_W(U_n) = \tau_{\widehat{n-1}}(W \tilde{\times} U_n)$. Note that

$$\downarrow_{[]} t_W = d_{n-1}\tau_{\widehat{n-1}}(W) = d_{n-1}\tau_{\widehat{n}}(W) = \tau_{0\cdots(n-2)}(\mathbb{V}(T_{0\cdots(n-2)})). \tag{3.5}$$

Further, by their definitions via trivial joins from $W \subset K$, one can choose t_W to guarantee that $t_W(U_{n-1} \times U_n) \leq t_K(U_{n-1} \times U_n)$ for all Borel sets $U_{n-1} \times U_n \subset \mathbb{V}(T_{n-1} \oplus T_n)$. In particular, $\downarrow_{[]} t_W \leq \downarrow_{[]} t_K$. Likewise, for the closed set $K - W$, define $t_{K-W} \coloneqq t_K - t_W$, which is also a measure in $\mathbf{M}(T_{n-1} \oplus T_n)$ by construction. Note that both the closure $\bar{W}$ and the complement $K - W$ are closed in K, therefore both are compact. Replacing K with $\bar{W}$ or $K - W$ in (3.2) means that we can establish measures $\{t_{W_\lambda}\}_{\lambda \in \Lambda}$ for a countable bifurcating collection $\{W_\lambda\}_{\lambda \in \Lambda}$ of open sets; any measurable set in $\mathbb{V}(T_{0\cdots(n-2)})$ can be κ-almost covered by disjoint sets in the collection. By analogy, we refer to the $\{W_\lambda\}_{\lambda \in \Lambda}$ as a dyadic collection.

[13]Of course the value of $\frac{1}{2}$ is not special, but aesthetic. Any $0 < \kappa(W) < \kappa(K)$ will do.

Given such a countable collection $\{t_{W_\lambda}\}_{\lambda\in\Lambda} \subset \mathbf{M}(T_{n-1} \oplus T_n)$, define a measure $\tau^{(n-1)} \in \mathbf{M}(T)$ on Borel sets $U_0 \tilde{\times} \cdots \tilde{\times} U_n$ by disjoint σ-additivity,

$$\tau^{(n-1)}(U_0 \tilde{\times} \cdots \tilde{\times} U_n) \coloneqq \sum_{\substack{\text{disjoint} \\ W_\lambda \subset U_0 \tilde{\times} \cdots \tilde{\times} U_{n-2}}} t_{W_\lambda}(U_{n-1} \times U_n). \tag{3.6}$$

By construction, $d_n \tau^{(n-1)} = \tau_{\widehat{n}} \in \mathcal{X}'$ and $d_{n-1}\tau^{(n-1)} = \tau_{\widehat{n-1}} \in \mathcal{X}'$. Therefore, by the strong join condition, $(T, \tau^{(n-1)}) \in \mathcal{X}'$. The data table $(T, \tau^{(n-1)})$ provides the base case for induction on faces.

Assume for induction that for some m satisfying $1 < m \leq n-1$ there exists a data table $(T, \tau^{(m)}) \in \mathcal{X}'$ such that $d_k \tau^{(m)} = \downarrow_{T_{\hat{k}}} \tau^{(m)} = \tau_{\hat{k}} \in \mathcal{X}'$ for all $m \leq k \leq n$. Denote the "error" of the d_{m-1} face as

$$\varepsilon_{m-1} \coloneqq \left(d_{m-1}\tau^{(m)} - \tau_{\widehat{m-1}} \right). \tag{3.7}$$

The error ε_{m-1} is a signed measure—not a measure—on $\mathbb{V}(T_{\widehat{m-1}})$, but the face operation of marginalization is still sensible. Then for $m \leq k \leq n-1$,

$$\begin{aligned} d_k \varepsilon_{m-1} &= d_k(d_{m-1}\tau^{(m)} - \tau_{\widehat{m-1}}) \\ &= d_{m-1}d_{k+1}\tau^{(m)} - d_k \tau_{\widehat{m-1}} = d_{m-1}\tau_{\widehat{k+1}} - d_{m-1}\tau_{\widehat{k+1}} = 0. \end{aligned} \tag{3.8}$$

Also, for application below, consider the pre-measure f on Borel sets $U_{m-1} \subset \mathbb{V}(T_{m-1})$ as defined by

$$f(U_{m-1}) = \inf \left\{ \frac{\tau^{(m)}(W \tilde{\times} U_{m-1} \tilde{\times} Z)}{\varepsilon_{m-1}(W \tilde{\times} Z)} \text{ for Borel } W \tilde{\times} Z \subset \mathbb{V}(T_{\widehat{n-1}}) \text{st} \varepsilon_{m-1}(W \tilde{\times} Z) > 0 \right\}. \tag{3.9}$$

Observe the inequality

$$f(\mathbb{V}(T_{m-1})) \geq 1, \tag{3.10}$$

which follows because for all Borel sets $W \tilde{\times} Z \subset \mathbb{V}(T_{\widehat{n-1}})$ satisfying $\varepsilon_{m-1}(W \tilde{\times} Z) > 0$, we have

$$\frac{\tau^{(m)}(W \tilde{\times} \mathbb{V}(T_{m-1}) \tilde{\times} Z)}{\varepsilon_{m-1}(W \tilde{\times} Z)} = \frac{\tau_{\widehat{m-1}}(W \tilde{\times} Z) + \varepsilon_{m-1}(W \tilde{\times} Z)}{\varepsilon_{m-1}(W \tilde{\times} Z)} = 1 + \frac{\tau_{\widehat{m-1}}}{\varepsilon_{m-1}}(W \tilde{\times} Z) \geq 1. \tag{3.11}$$

Let $\rho_{m-1} \in \mathbf{M}(T_{m-1})$ be a probability measure satisfying the condition

$$\rho_{m-1}(U_{m-1}) \leq f(U_{m-1}) \tag{3.12}$$

for all Borel $U_{m-1} \subseteq \mathbb{V}(T_{m-1})$. Such probability measures are guaranteed to exist by (3.10).

Then, define for any[14] Borel $W \tilde{\times} U_{m-1} \tilde{\times} Z \subseteq \mathbb{V}(T)$,

$$\tau^{(m-1)}(W \tilde{\times} U_{m-1} \tilde{\times} Z) \coloneqq \tau^{(m)}(W \tilde{\times} U_{m-1} \tilde{\times} Z) - \rho_{m-1}(U_{m-1}) \cdot \varepsilon_{m-1}(W \tilde{\times} Z), \tag{3.13}$$

and extend by additivity. By construction, $\tau^{(m-1)}$ is additive and zero-null. Non-negativity follows from (3.12) and the definition of f in (3.9); therefore, $\tau^{(m-1)}$ is

[14]Not necessarily meeting the $\varepsilon_{m-1} > 0$ condition above.

a measure on $\mathbb{V}(T)$. Moreover, $\tau^{(m-1)}$ satisfies the desired marginalizations, shown here:

$$\begin{aligned} d_{m-1}\tau^{(m-1)}(W\tilde{\times}Z) &= \tau^{(m)}(W\tilde{\times}Z) - \rho_{m-1}(\mathbb{V}(T_{m-1}))\cdot\varepsilon_{m-1}(W\tilde{\times}Z) \\ &= \tau^{(m)}(W\tilde{\times}Z) - 1\cdot\varepsilon_{m-1}(W\tilde{\times}Z) \\ &= \tau_{\widehat{m-1}}(W\tilde{\times}Z). \end{aligned} \tag{3.14}$$

And, for $m \le k \le n$, the properties (3.8) apply to give

$$d_k\tau^{(m-1)} = d_k\left(\tau^{(m)} - \rho_{m-1}\cdot\varepsilon_{m-1}\right) = d_k\tau^{(m)} - \rho_{m-1}\cdot 0 = d_k\tau^{(m)} = \tau_{\widehat{k}}. \tag{3.15}$$

Therefore, $(T, \tau^{(m-1)})$ is a data table that has the desired faces d_{m-1} through d_n. The inductive step is established. The ultimate data table $(T, \tau^{(1)})$ provides the n-simplex Δ^n completing Λ^n_0. □

Remark 3.12 (Freedom). The flexibility in choosing (T, τ) arises from an initial parametric choice of joined measures $\{t_x\}_{x\in\mathbb{V}(T_{0\cdots(n-2)})} \subset \mathbf{M}(T_{n-1} \oplus T_n)$ and a finite set of probability measures ρ_{n-2}, ρ_{n-3}, $\ldots$, ρ_1.

3(b). **Simplicial homotopy of data complexes.**

Definition 3.13 (Fibrant)**.** A data complex $p : \mathcal{X}' \to \mathcal{A}'$ satisfying the strong join condition is called fibrant.

See Appendix A for a categorical version of this definition. The entire *raison d'être* of fibrant objects is that they admit homotopy, as proven by [10] and [18], which allows obstruction theory to be studied in direct analogy to Steenrod. In the category of simplicial sets, the term *fibrant* refers only to the Kan extension condition. Our practical desire to use joins as a weak-equivalence compels us to require the strong join condition. By Theorem 3.11(1) the traditional definition and all of its consequences are implied.

Corollary 3.14. *Suppose a data subcomplex $\mathcal{X}'$ of an ambient $\mathcal{X}$ is fibrant, and fix a basepoint data table (T_0, τ_0). The homotopy group $\pi_n(\mathcal{X}', \tau_0)$ is well-defined for all n, and satisfies the typical properties of homotopy categories over model categories.*

Theorem 3.15. *For any attribute set A and value spaces $\mathbb{V}(\ )$, the ambient data complex $p : \mathcal{X} \to \mathcal{A}$ is fibrant.*

Proof. The very definition of an ambient $\mathcal{X}$ is that it includes *all* finite measures over the relevant metric spaces, so it includes the set Joins() in particular. □

We now want to explore how a data subcomplex $p : \mathcal{S} \to \mathcal{B}$ of an ambient $p : \mathcal{X} \to \mathcal{A}$ interacts with any other attribute list $T \in \mathcal{A}$. The following sets are of interest.

Definition 3.16. Let $\mathcal{S}|_T \coloneqq \{(S, \sigma) \in \mathcal{S} : \exists\, \iota : S \hookrightarrow T\}$, the set of data tables in the data subcomplex that are detected by T. Let $\mathcal{B}|_T \coloneqq \{S \in \mathcal{B} : \exists\, \iota : S \hookrightarrow T\} = p(\mathcal{S}|_T)$, the set of attribute lists in the subcomplex that are detected by T.

A data subcomplex $p : \mathcal{S} \to \mathcal{B}$ may not be fibrant, so we define a convenient fibrant space that contains it. The notation $\mathcal{F}^0$ is meant to be suggestive; in Section 4, a larger filtration of simplicial sets is created (Definition 4.5) by turning the equality in the definition below into an inequality involving Wasserstein distance.

Definition 3.17 (Complex of perfect joins)**.** For any data subcomplex $\mathcal{X}'$ of an ambient $\mathcal{X}$, let $\mathcal{F}^0$ denote the subset of $\mathcal{X}$ defined by

$$(T,\tau) \in \mathcal{F}^0 \text{ if and only if } \forall\, a \in T,\ \exists\, (S,\sigma) \in \mathcal{S}|_T \text{ such that } a \in S \text{ and } \downarrow_S \tau = \sigma.$$

Note: the quantifier "$\forall a \in T$" refers to each entry in the attribute list, which means repeated entries must have corresponding measures.

The definition of $\mathcal{F}^0$ is a convenient way to say "consider everything that can be generated from $\mathcal{S}$ using Joins()," as justified by the following lemma. Similarly, the upcoming Definition 4.5 of $\mathcal{F}^t$ gives a convenient way of saying "consider everything that can be approximated to an acceptable level of uncertainty from $\mathcal{S}$ using Joins()."

Lemma 3.18. *$(T,\tau) \in \mathcal{F}^0$ if and only if there is a sequence $(T_0,\tau_0),(T_1,\tau_1),\ldots,(T_k,\tau_k)$ such that*

- *$(T_0,\tau_0) = (S_0,\sigma_0) \in \mathcal{S}$, and*
- *$\forall i = 1,\ldots,k,\ (T_i,\tau_i) \in \text{Joins}(\tau_{i-1},\sigma_i,T_{i-1}\cap S_i)$ for some $(S_i,\sigma_i) \in \mathcal{S}$, and*
- *$(T_k,\tau_k) = (T,\tau)$.*

Proof. Suppose $(T,\tau) \in \mathcal{F}^0$. Let $a_0 \in T$ denote the first attribute of T. By the definition of $\mathcal{F}^0$, there exists $(S_0,\sigma_0) \in \mathcal{S}$ with an attribute inclusion $\iota_0 : S_0 \hookrightarrow T$ such that $\downarrow_{\iota_0} \tau = \sigma_0$ and such that a_0 is in the image of ι_0. Let $(T_0,\tau_0) = (S_0,\sigma_0)$. By reducing T_0 if necessary, we may ensure that $\iota_0(T_0)$ is contiguous within T. If $T_0 = T$, then the sequence is complete.

Otherwise, there exists some first attribute a_1 in T/ι_0. By the definition of $\mathcal{F}^0$, there exists $(S_1,\sigma_1) \in \mathcal{S}$ with an attribute inclusion $\iota_1 : S_1 \hookrightarrow T$ such that $\downarrow_{\iota_1} \tau = \sigma_1$ and such that a_1 is in the image of ι_1. By reducing S_1 if necessary, we may ensure that $\iota_1(S_1)$ is contiguous within T, and that $T_0 \cap S_1$ is also contiguous. With these reductions, the orderings are consistent such that $T_1 \coloneqq \text{Merge}(T_0,S_1,T_0\cap S_1)$ is equipped with a list inclusion $T_1 \hookrightarrow T$. Because τ is given, let $\tau_1 = \downarrow_{T_1} \tau$, which by construction is an element of $\text{Joins}(\tau_0,\sigma_1,T_0\cap S_1)$. Repeat this process until all elements a_i of T are in the image of some inclusion $T_i \hookrightarrow T$.

For the converse, note that each $a \in T$ is included in some S_i, which is sufficient. □

Corollary 3.19. *$\mathcal{F}^0$ includes all independent products formed from data tables in $\mathcal{S}$.*

Lemma 3.20. *For any data subcomplex $\mathcal{S}$ of an ambient $\mathcal{X}$, the complex of perfect joins $\mathcal{F}^0$ is fibrant.*

Proof. The data subcomplex $\mathcal{S}$ is closed under face maps and degeneracy maps, so application of those maps to all (S,σ) in the definition shows that $\mathcal{F}^0$ is closed under the face maps and degeneracy maps as well. To verify that $\mathcal{F}^0$ is fibrant, suppose that $(T_{012},\tau_{012}) \in \mathcal{X}$ is any join of (T_{01},τ_{01}) and (T_{02},τ_{02}) in $\mathcal{F}^0$. Because every $a \in T_{012}$ appears in T_{01} or T_{02}, the existence of $(S,\sigma) \in \mathcal{S}$ is inherited from (T_{01},τ_{01}) and (T_{02},τ_{02}). □

We conclude this section by tying simplicial homotopy theory to Problem 2.39.

Lemma 3.21. *Suppose $\mathcal{X}'$ is a fibrant data subcomplex of an ambient $\mathcal{X}$. A basepoint-preserving simplicial map $f : \partial\Delta^n \to \mathcal{X}'$ defines a class in $\alpha(f) \in \pi_{n-1}(\mathcal{X}')$. Moreover, $\alpha(f) = e$ if and only if f admits an extension $f^+ : \Delta^n \to \mathcal{X}'$.*

Proof. The first claim reduces to Lemma 9.6 in [7]. The second claim reduces to Lemma 7.4 in [9]. Our definition of fibrant implies path-connectedness, so a spanning tree can be used for locality such as in [10]. □

Corollary 3.22. *Suppose that $p : \mathcal{S} \to \mathcal{B}$ is a data subcomplex of an ambient $p : \mathcal{X} \to \mathcal{A}$ such that $\mathcal{B}_{n-1} = \mathcal{A}_{n-1}$ for some $n \geq 1$. Fix a simplicial section $\sigma : \mathcal{B}_{n-1} \to \mathcal{S}_{n-1}$. The following are equivalent (omitting basepoints for brevity).*

(1) *For every composition*

$$\partial\Delta^n \xrightarrow{c} \mathcal{B}_{n-1} \xrightarrow{\sigma} \mathcal{S}_{n-1} \xrightarrow{\iota} \mathcal{F}^0,$$

we have $\alpha(\iota \circ \sigma \circ c) = e \in \pi_{n-1}(\mathcal{F}^0)$.

(2) *σ admits an extension of the form $\sigma^+ : \mathcal{A}_n \to \mathcal{F}^0_n$.*

Proof. Because $\mathcal{A}_{n-1} = \mathcal{B}_{n-1}$, the boundary of every n-simplex in $\mathcal{A}$ appears in $\mathcal{B}$. Apply the previous lemma for each $f = \iota \circ \sigma \circ c$ as a map $f : \partial\Delta^n \to \mathcal{X}'$ for $\mathcal{X}' = \mathcal{F}^0$. □

This corollary is revisited as Lemma 4.9. The corollary fails when no such extension can be found. Then, the question remains: how to measure the failure of this corollary? That measurement is the purpose of filtered obstruction theory.

4. Filtrations and obstructions

This section concludes the theoretical framework outlined in Section 1(a). Together, obstructions and filtrations allow us to detect when merging is possible; if merging appears obstructed, we can determine whether merging can be achieved by reverting a previous merge or by altering some of the data tables. Section 4(a) introduces a filtration from a data subcomplex $\mathcal{S}$ to its ambient $\mathcal{X}$ using the Wasserstein distance. Each level of the filtration is fibrant, which allows one to define an obstruction cocycle (Section 4(b)) at each level of the filtration. Eventually, for a high enough level in the filtration, the obstruction cocycle becomes trivial, so the importance of the obstruction cocycle can be measured using topological persistence. This statement is formalized in Theorem 4.13, which can be seen as the main payoff of our theoretical development in terms of database engineering. As promised in the introduction, the theory of data complexes does not just mathematize the notion of table merging; rather, it provides further powerful operations when traditional merging is impossible.

4(a). **Filtrations from data subcomplexes.** A general notion of persistence on simplicial sets appears in [16]. In summary, a fibrant filtration of simplicial sets is a bi-graded collection of sets $\{\mathcal{F}^t_n\}$ for $0 \leq t \leq \infty$ and $n \in \mathbb{N}$ equipped with maps d_i and s_i such that

(1) $(\mathcal{F}^t, d_i, s_i)$ is a simplicial set for each t,
(2) $\mathcal{F}^s_i \subseteq \mathcal{F}^t_i$ for all $s \leq t$, and
(3) $(\mathcal{F}^t, d_i, s_i)$ is fibrant for each t.

The fibrant condition implies that $\pi_n(\mathcal{F}^t)$ is well-defined for all t, and the inclusion maps $\mathcal{F}^s \hookrightarrow \mathcal{F}^t$ induce maps on homotopy, $\pi_n(\mathcal{F}^s) \to \pi_n(\mathcal{F}^t)$.

We now define a specific filtration for a data subcomplex that is designed to meet our application regarding joining data tables. Recall that $(\mathbb{V}(a), \rho_a)$ is a Radon space for each attribute a.

Definition 4.1 (Wasserstein distance)**.** For any $a \in A$ with $(\mathbb{V}(a), \rho_a)$, and $\tau_1, \tau_2 \in \mathbf{M}(a)$, let

$$w_a(\tau_1, \tau_2) \coloneqq \inf \left\{ \int_{\mathbb{V}([a,a])} \rho_a(x_1, x_2)\, \mathrm{d}\mu(x_1, x_2) \;:\; \mu \in \mathbf{M}([a,a]), \downarrow_1 \mu = \tau_1,\ \downarrow_2 \mu = \tau_2 \right\}. \tag{4.1}$$

The reductions $\downarrow_1$ and $\downarrow_2$ refer to the two copies of the attribute a.

For any $T \in \mathcal{A}$ and $\tau_1, \tau_2 \in \mathbf{M}(T)$, let

$$w_T(\tau_1, \tau_2) \coloneqq \inf \left\{ \int_{\mathbb{V}(T \oplus T)} \rho_T(x_1, x_2) \mathrm{d}\mu(x_1, x_2) \;:\; \mu \in \mathbf{M}(T \oplus T), \downarrow_1 \mu = \tau_1,\ \downarrow_2 \mu = \tau_2 \right\}. \tag{4.2}$$

The reductions $\downarrow_1$ and $\downarrow_2$ refer to the two interwoven copies of the attribute list T.

Remark 4.2. Recall that $\rho_T(x_1, x_2) = \max_{a \in T} \rho_a(x_{1,a}, x_{2,a})$, the L^∞-metric obtained from the individual attribute metrics. Also, in the special case that $\downarrow_{[]} \tau_1 = \downarrow_{[]} \tau_2$, the infimum argument μ lies in the space of trivial joins, $\mathrm{Joins}(\tau_1, \tau_2, [])$, so the Wasserstein distance is tied to our notion of fibrant data complexes.

Lemma 4.3. *Suppose that $\tau_1, \tau_2 \in T$. If $w_T(\tau_1, \tau_2) = t$, then $w_{d_i T}(d_i \tau_1, d_i \tau_2) \leq t$.*

Proof. Let a indicate the ith attribute of T, and write $T' = d_i T$ with inclusion $T' \hookrightarrow T$ and quotient inclusion $[a] \hookrightarrow T$. Then write $(T', \tau_1') \coloneqq d_i(T, \tau_1)$ and $(T', \tau_2') \coloneqq d_i(T, \tau_2)$. For any $\mu \in \mathbf{M}(T \oplus T)$ such that $\downarrow_1 \mu = \tau_1$ and $\downarrow_2 \mu = \tau_2$, let $\mu' \in \mathbf{M}(T' \oplus T')$ be the reduction of μ obtained by applying both copies of $\downarrow_{T'} = d_i$. Also, we use the notational convention $x = (y, z)$ for $x \in \mathbb{V}(T)$, $y \in \mathbb{V}(T')$, $z \in \mathbb{V}([a])$, so the L^∞ metric gives $\rho_T(x_1, x_2) \geq \rho_{T'}(y_1, y_2)$.

$$\begin{aligned} &\int_{(x_1,x_2) \in \mathbb{V}(T \oplus T)} \rho_T(x_1, x_2) \mathrm{d}\mu(x_1, x_2) \\ &\geq \int_{((y_1,z_1),(y_2,z_2)) \in \mathbb{V}(T \oplus T)} \rho_{T'}(y_1, y_2) \mathrm{d}\mu((y_1, z_1), (y_2, z_2)) \\ &= \int_{(y_1,y_2) \in \mathbb{V}(T' \oplus T')} \int_{(z_1,z_2) \in \mathbb{V}([a,a])} \rho_{T'}(y_1, y_2) \mathrm{d}\mu((y_1, z_1), (y_2, z_2)) \\ &= \int_{(y_1,y_2) \in \mathbb{V}(T \oplus T)} \rho_{T'}(y_1, y_2) \mathrm{d}\mu'(y_1, y_2). \end{aligned} \tag{4.3}$$

Therefore, the infimum defining $w_{T'}(\tau_1', \tau_2')$ cannot be greater than the infimum defining $w_T(\tau_1, \tau_2)$. □

Lemma 4.4. *Suppose that $\tau_1, \tau_2 \in T$. If $w_T(\tau_1, \tau_2) = t$, then $w_{s_i T}(s_i \tau_1, s_i \tau_2) \leq t$.*

Proof. Let a indicate the ith attribute of T. Let $T^+ \coloneqq s_i T$, equipped with the degeneracy inclusion $T \hookrightarrow T^+$ and its quotient $[a] \hookrightarrow T^+$. Then write $(T^+, \tau_1^+) \coloneqq s_i(T, \tau_1)$ and $(T^+, \tau_2^+) \coloneqq s_i(T, \tau_2)$. Now, for any $\mu \in \mathbf{M}(T \oplus T)$ such that $\downarrow_1 \mu = \tau_1$ and $\downarrow_2 \mu = \tau_2$, let $\mu^+ \in \mathbf{M}(T^+ \oplus T^+)$ be the degeneracy of μ obtained by applying both copies of s_i. Consider the integral $\int_{(x_1,x_2) \in \mathbb{V}(T^+ \oplus T^+)} \rho_{T^+} \mathrm{d}\mu^+$. Note that the

distributional form of the degeneracy is a delta,

(4.4)
$$\mathrm{d}\mu^+(x_1,x_2) = \begin{cases} \mathrm{d}\mu(y_1,y_2), & \text{if } x_1 = s_i y_1,\ x_2 = s_i y_2 \text{ for some } (y_1,y_2) \in \mathbb{V}(T \oplus T), \\ 0, & \text{otherwise.} \end{cases}$$

Moreover, if $x_1 = s_i y_1$, $x_2 = s_i y_2$ for some $(y_1, y_2) \in \mathbb{V}(T \oplus T)$, then $\rho_{T^+}(x_1, x_2) = \rho_T(y_1, y_2)$. Together, these give $\int_{(x_1,x_2)\in\mathbb{V}(T^+\oplus T^+)} \rho_{T^+}\mu^+ = \int_{(y_1,y_2)\in\mathbb{V}(T\oplus T)} \rho_T \mathrm{d}\mu$. Therefore, the infimum defining $w_{s_iT}(s_i\tau_1, s_i\tau_2)$ cannot be greater than the infimum defining $w_T(\tau_1, \tau_2)$. □

Now we produce a particular fibrant filtration for a data subcomplex.

Definition 4.5 (The complex of approximate joins)**.** Let $p : \mathcal{S} \to \mathcal{B}$ be a data subcomplex of an ambient $p : \mathcal{X} \to \mathcal{A}$. For any $0 \le t \le \infty$, let

$$\mathcal{F}^t \coloneqq \{(T,\tau) \in \mathcal{X} \ :\ \forall a \in T, \exists (S,\sigma) \in \mathcal{S},\ [a] \hookrightarrow S \hookrightarrow T,\ w_S(\downarrow_S \tau, \sigma) \le t\}.$$

Note that the case $t = 0$ reproduces the complex of perfect joins, $\mathcal{F}^0$. Note also that $\mathcal{F}^\infty = \mathcal{X}$.

Theorem 4.6. *For each $t \in [0,\infty]$, $\mathcal{F}^t$ is a fibrant data subcomplex of $\mathcal{X}$.*

The proof is identical to the proof of Lemma 3.20, replacing the equality with an inequality.

Proof. Recall that the data complex $\mathcal{S}$ is closed under face maps and degeneracy maps. Note the face and degeneracy bounds for the Wasserstein distance given above. Application of those maps to the (S,σ) and (T,τ) in the definition shows that $\mathcal{F}^t$ is closed under the face maps and degeneracy maps as well. Therefore, $\mathcal{F}^t$ is a data subcomplex.

To verify that $\mathcal{F}^t$ is fibrant, apply Theorem 3.15 to obtain all joins $(T_{012}, \tau_{012}) \in \mathcal{X}$ from any (T_{01}, τ_{01}) and (T_{02}, τ_{02}) in $\mathcal{F}^t$. We must show such τ_{012} lies in $\mathcal{F}^t$. Fix $a \in T_{012}$. Because every $a \in T_{012}$, it appears in T_{01} or T_{02}. For concreteness, assume $a \in T_{01}$. There is some $(S,\sigma) \in \mathcal{S}$ such that $w_S(\downarrow_S \tau_{01}, \sigma) \le t$. By the construction of τ_{012}, we have $\downarrow_{T_{01}} \tau_{012} = \tau_{01}$, so $\downarrow_S \tau_{012} = \downarrow_S \tau_{01}$. Hence, $w_S(\downarrow_S \tau_{012}, \sigma) \le t$. □

Because $\mathcal{F}^t$ is fibrant, all of the usual consequences apply in homotopical algebra, such as

Corollary 4.7. *Fix a data subcomplex $\mathcal{S}$ of an ambient $\mathcal{X}$. For each $t \in [0,\infty]$, and for each $n \ge 0$, the pointed homotopy group $\pi_n(\mathcal{F}^t, *)$ is well-defined. Moreover, for $t_1 \le t_2$, the inclusion of data subcomplexes $\mathcal{F}^{t_1} \subset \mathcal{F}^{t_2}$ induces a homomorphism of pointed homotopy groups $\pi_n(\mathcal{F}^{t_1}, *) \to \pi_n(\mathcal{F}^{t_2}, *)$.*

4(b). **Persistent obstruction theory for data subcomplexes.** Because we have established fibrant objects with resulting homotopy and homology, we are equipped to extend obstruction theory to our application. Although our category is not classical, the next several results are modeled on the classical work summarized in Section 6 of [24], Section 34 of [22], Section 4 of [11], and [15]. The discussion culminates in Definition 4.8 and Theorem 4.13.

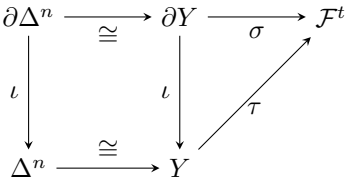
Definition 4.8 (Obstruction cocycle)**.** Let $\mathcal{S} \subseteq \mathcal{F}^0 \subset \cdots \mathcal{F}^t \subset \cdots \subset \mathcal{F}^\infty = \mathcal{X}$ be the filtration of a path-connected data complex. Fix a dimension n such that

$d_i Y \in \mathcal{B}_{n-1}$ for all faces d_i of all $Y \in \mathcal{A}_n$. Let $\sigma : \mathcal{B}_{n-1} \to \mathcal{S}_{n-1}$ be a data section. For a fixed basepoint $(T_0, \tau_0) \in \mathcal{S} \subset \mathcal{F}^0$, define

$$\xi_\sigma^t \in C^n(\mathcal{A}, R; \pi_{n-1}(\mathcal{F}^t, (T_0, \tau_0))) \tag{4.5}$$

to be the element of $\pi_{n-1}(\mathcal{F}^t, (T_0, \tau_0))$ that is represented by the loop corresponding[15] to the cycle $\sigma(\partial Y) \in C_{n-1}(\mathcal{S}) \subset C_{n-1}(\mathcal{F}^t)$ for any $Y \in \mathcal{A}_n$. Extend by linearity for $Y \in C_n(\mathcal{A}, R)$. We typically omit the basepoint and ring for brevity, so $\xi_\sigma^t \in C^n(\mathcal{A}; \pi_{n-1}(\mathcal{F}^t))$.

Lemma 4.9. *Fix $Y \in \mathcal{A}_n$. If $\xi_\sigma^t(Y) = e \in \pi_{n-1}(\mathcal{F}^t)$, then there exists $(Y, \tau) \in \mathcal{F}_n^t$ such that the diagram commutes*



Lemma 4.10. *The cochain ξ_σ^t is a cocycle. So, it defines a cohomology class $[\xi_\sigma^t] \in H^n(\mathcal{A}, R; \pi_{n-1}(\mathcal{F}^t, (T_0, \tau_0)))$.*

Proof. For any $X \in \mathcal{A}_{n+1}$, we have $\delta\xi_\sigma^t(X) = \xi_\sigma^t(\partial X)$, but then the trivial cycle $0 = \partial(\partial X) \in C_n(\mathcal{A})$ represents the trivial class $e \in \pi_{n-1}(\mathcal{F}^0)$. □

Remark 4.11. Obstructions in dimension $n-1 = 1$ detect loops in $\mathcal{F}^t$, which will prevent some $n+1 = 3$ data tables from being mutually joinable.

Obstructions in dimension $n-1 = 2$ detect spheres in $\mathcal{F}^t$, which will prevent some $n+1 = 4$ data tables from being mutually joinable.

Obstructions in dimension $n-1 = 0$ detect non-path-connectedness of $\mathcal{F}^t$, which would prevent some $n+1 = 2$ data tables from being joinable (but this is impossible with our definitions including trivial joins).

The next theorem is an adaptation of Theorem 34.6 and Corollary 34.7 in [22], which is summarized in Theorem 4.5 of [11]. It relies on defining a *difference cochain* that compares a homology class of sections.

Theorem 4.12. *Fix a data section $\sigma : \mathcal{B}_{n-1} \to \mathcal{S}_{n-1}$. Suppose $\xi_\sigma^t = \delta\eta$ for some $\eta \in C^{n-1}(\mathcal{A}; \pi_{n-1}(\mathcal{F}^t))$. Then there exists a data section $\tau : \mathcal{A}_n \to \mathcal{F}_n^t$ such that $\tau|_{n-2} = \sigma|_{n-2}$. The converse holds as well.*

Theorem 4.13 restates Theorems 4.9 and 4.12 in practical language.

Theorem 4.13 (Steenrod's trichotomy)**.** *Fix a data subcomplex $\mathcal{S}$ of an ambient $\mathcal{X}$, with Wasserstein filtration $(\mathcal{F}^t)$. Exactly one of the following is true.*

1. *$\xi_\sigma^t = e$ as a cocycle. Every $n-1$-cycle of $n+1$ data tables in $\mathcal{S}$ over a total of $n+1$ attributes can be approximately joined to a single data table over those $n+1$ attributes, allowing error at-most t in any reduction to the original data.*

[15]The well-definedness of this loop is implied by our assumption $R = \mathbb{Z}/2\mathbb{Z}$.

(2) $\xi^t_\sigma \neq e$ *as a cocycle, but* $[\xi^t_\sigma] = e$ *as a cohomology class. There is some* $(n-1)$*-cycle of* $n+1$ *data tables* $(T_{\hat{0}}, \tau_{\hat{0}}), \ldots, (T_{\hat{n}}, \tau_{\hat{n}})$ *in* $\mathcal{S}$ *such that the combined attribute list* $T = [a_0, \ldots, a_n]$ *does* not *admit an approximate join* (T, τ) *with error at-most* t*. However, if one considers all of the faces of these data tables, then there is an approximate join to* (T, τ) *of error at-most* t*.*

(3) $[\xi^t_\sigma] \neq e$ *as a cohomology class. There is some* $(n-1)$*-cycle of* $n+1$ *data tables* $(T_{\hat{0}}, \tau_{\hat{0}}), \ldots, (T_{\hat{n}}, \tau_{\hat{n}})$ *in* $\mathcal{S}$ *such that the combined attribute list* $T = [a_0, \ldots, a_n]$ *does* not *admit an approximate join* (T, τ) *with error at-most* t*, even when omitting attributes from the original data tables. The only way to produce a single joined table is to increase the error threshold* t*.*

Definition 4.14 (Persistence of obstruction)**.** Let $\mathcal{S} \subseteq \mathcal{F}^0 \subset \cdots \mathcal{F}^t \subset \cdots \subset \mathcal{F}^\infty = \mathcal{X}$ be the filtration of a path-connected data complex. Fix a dimension n such that $d_i Y \in \mathcal{B}_{n-1}$ for all faces d_i of all $Y \in \mathcal{A}_n$. Let $\sigma : \mathcal{B}_{n-1} \to \mathcal{S}_{n-1}$ be a data section. Fix a basepoint $(T_0, \tau_0) \in \mathcal{S} \subset \mathcal{F}^0$. Define

$$t_n(\sigma) \coloneqq \inf\{t \,:\, \xi^t_\sigma = e \in C^n(\mathcal{A}; \pi_{n-1}(\mathcal{F}^t))\}$$

and

$$t'_n(\sigma) \coloneqq \inf\{t \,:\, [\xi^t_\sigma] = e \in H^n(\mathcal{A}; \pi_{n-1}(\mathcal{F}^t))\}.$$

Note that $t'_n(\sigma) \leq t_n(\sigma)$.

Remark 4.15. Consider a data section $\sigma : \mathcal{B} \to \mathcal{S}$. A specific value $t_n(\sigma) = t$ means that σ admits an extension into $\mathcal{F}^t$, but not for any level of the filtration less than t. In other words, there is no obstruction to extension beyond the mere existence of the data section $\sigma : \mathcal{B}_{n-1} \to \mathcal{F}^t_{n-1}$. Similarly, by Theorem 4.13, a specific value $t'_n(\sigma) = t$ means that there is no obstruction to extension beyond the mere existence of the data section $\sigma|_{n-2} : \mathcal{B}_{n-2} \to \mathcal{F}^t_{n-2}$.

Remark 4.16. When obstructions are resolved, there are typically many solutions to Problems 1.2/1.3. That is, if any hypothesis is consistent in 1.1, then there are typically many other hypotheses that are consistent as well. Typical methods for choosing among them often involve posing and then solving some optimization problem. We might propose enriching those optimization problems via inclusion of a measure of global inconsistency. More precisely, the cost of a proposed data section σ might be some combination of a local cost and some decreasing function of $t_n(\sigma)$ or $t'_n(\sigma)$; in other words, one might penalize proposed local mergers based on the degree of difficulty they cause in forming global consensus with other local mergers.

5. Discussion

This paper provides a mathematical foundation for semi-automated data-table-alignment tools that are common in commercial database software. Data tables are abstracted as measures over value spaces, and the problem of merging tables, or indeed merging previously-merged tables, is recast as the search for a measure that marginalizes correctly. This abstraction, and the simplicial set structure built with it, permits several advances over the current state of the art in database engineering. Ongoing and future work will focus on developing clear algorithms for application of persistent obstruction theory to real-world database engineering and related problems in data science.

We conclude this paper with several brief remarks about further work and also some practicalities for future use of this theory:

- A data sample X in any metric space V provides a measure, by counting. The measure is $\mu(U) = \#(U \cap X)$ or normalized as $\mu(U) = \frac{\#(U \cap X)}{\#X}$ for any $U \in 2^V$.
- For computational purposes, most infinite metric spaces can be considered as compact or finite spaces, using bounds or bins or kernel methods or distributional coordinates that are appropriate to the problem at hand.
- On the compact metric spaces $\mathbb{V}(T)$, measures of interest can be described as density functions via a Radon–Nikodym comparison to the uniform probability measure κ_T.
- One attribute can represent models on other attributes, providing an interpretation of Bayesian inference and an opportunity to apply persistent obstruction theory to compact parameterized model spaces. In machine learning, one could use this framework to describe the compatibility of solutions in ensemble methods.
- Any list of attributes can be considered as a single attribute, because it still provides measures over some metric space. There is no requirement that attribute value spaces are "minimal" or "1-dimensional" in any sense.
- Filtrations other than L^∞-Wasserstein might work, too, but someone has to prove that all levels of the filtration are fibrant.
- To study a complex of approximate joins, $\mathcal{F}^t$, one must compute Wasserstein distances as in Definition 4.1. This can be done efficiently using the tools of optimal transport as in [17].
- To apply Theorem 4.13, one must compute ξ^t_σ in the simplicial homotopy group $\pi_{n-1}(\mathcal{F}^t)$. This is definitely the greatest challenge for realizing these mathematical advances as actual software, because homotopy groups are notoriously difficult to compute in general. The task is simplified in our case by several factors. First, we do not necessarily need to know the group structure of $\pi_{n-1}(\mathcal{F}^t)$ to know whether a particular element ξ^t_σ is trivial in that group. Second, a data subcomplex $p : \mathcal{S} \to \mathcal{B}$ is always finitely generated with $\mathcal{B}$ finite, and that finite number is small (several, not several trillion) in most use-cases. Third, because any list of attributes can be considered as a single attribute, problems that are *a priori* high-dimensional can be studied with a smaller list of formal attributes. Fourth, we expect the homotopy $\pi_{n-1}(\mathcal{F}^t)$ to simplify as t increases, so for practical purposes it may be easy to bound $t_n(\sigma)$ even if each $\pi_{n-1}(\mathcal{F}^t)$ is difficult to compute. We expect (or hope) that π_1 and π_2 are often sufficient for practical problems.
- The most important conclusions of this work are: *Any manual or automatic data-merging system must analyze homotopy in order to guarantee success;* and *Obstructions can only be resolved two ways—backing up one step, or allowing additional leeway in the data comparison.*

Appendix A. Categorical definitions

This appendix provides a rapid summary of a categorical interpretation of the development in Section 2. For more on these topics, and for the notion of homotopy for fibrant objects in model categories, see [7,9,10,12,18]. The reader is warned that

each of these references uses a slightly different convention for ordering, opposite categories, and co-/contra-variant functors.

A(a). **Simplex.** Let **Set** denote the set category, whose objects are sets and whose morphisms are functions.

Let $\mathbf{\Delta}$ denote the simplex category, whose objects are the *nonempty* sets of natural numbers with the standard ordering $\leq$, written $\mathbf{n} \coloneqq \{0, 1, \cdots, n\}$, and whose morphisms are order-preserving functions. Let $\mathbf{\Delta_a}$ denote the augmented simplex category, whose objects are sets of natural numbers with the standard ordering, and whose morphisms are order-preserving functions. The augmented simplicial category includes the empty set, denoted $-\mathbf{1}$ or $\emptyset$, which is the initial object in the category. So, $\mathbf{\Delta_a} = \mathbf{\Delta} \cup \{\emptyset\}$. A monomorphism in $\mathbf{\Delta_a}$ is a one-to-one order-preserving function. The only bimorphisms/isomorphisms in $\mathbf{\Delta_a}$ are the identity maps. Among the morphisms in $\mathbf{\Delta}$ and $\mathbf{\Delta_a}$ are the co-faces d^i and co-degeneracies s^i, defined as follows.

$d^i : \mathbf{n} - \mathbf{1} \to \mathbf{n}$ by

$d^i : (0, \ldots, i-1, i, i+1, \ldots, n-1) \mapsto (0, \ldots, i-1, i+1, i+2, \ldots, n)$, respectively.

$s^i : \mathbf{n} + \mathbf{1} \to \mathbf{n}$ by

$s^i : (0, \ldots, i-1, i, i+1, \ldots, n+1) \mapsto (0, \ldots, i-1, i, i, \ldots, n)$, respectively.

These morphisms satisfy the conditions

(1) $d^j \circ d^i = d^i \circ d^{j-1}$, if $i < j$;
(2) $s^j \circ d^i = d^i \circ s^{j-1}$, if $i < j$;
(3) $s^j \circ d^j = d^{j+1} \circ s^j = \mathrm{id}$;
(4) $s^j \circ d^i = d^{i-1} \circ s^j$, if $i > j + 1$; and
(5) $s^j \circ s^i = s^i \circ s^{j+1}$, if $i \leq j$.

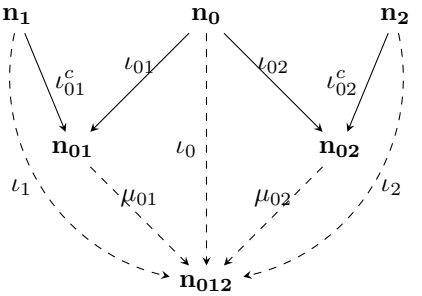
Every non-identity morphism in $\mathbf{\Delta}$ or $\mathbf{\Delta_a}$ can be written as a finite composition of co-face and co-degeneracy morphisms, so these five properties essentially characterize $\mathbf{\Delta}$ and $\mathbf{\Delta_a}$.

For our applications, the following lemmas about monomorphisms in $\mathbf{\Delta_a}$ are very useful. They are elementary, but do not appear in the standard references in this form. Merged indexing is merely an ordered formulation of the inclusion–exclusion principle.

Lemma A.1 (Complimentary monomorphism)**.** *For any monomorphism $\iota : \mathbf{n}' \to \mathbf{n}$ in $\mathbf{\Delta_a}$, write $m = n - n' - 1$. There is a monomorphism $\iota^c : \mathbf{m} \to \mathbf{n}$ in $\mathbf{\Delta_a}$ that enumerates the entries of $\mathbf{n}$ that are not in the image of ι.*

Lemma A.2 (Merged indexing)**.** *In the category $\mathbf{\Delta_a}$, suppose $\mathbf{n_0}, \mathbf{n_{01}}, \mathbf{n_{02}}$ are equipped with monomorphisms $\iota_{01} : \mathbf{n_0} \hookrightarrow \mathbf{n_{01}}$ and $\iota_{02} : \mathbf{n_0} \hookrightarrow \mathbf{n_{02}}$. Then, for $n_{012} = n_{01} + n_{02} - n_0$, there are monomorphisms $\mu_{01} : \mathbf{n_{01}} \hookrightarrow \mathbf{n_{012}}$ and $\mu_{02} : \mathbf{n_{02}} \hookrightarrow \mathbf{n_{012}}$ such that $\iota_0 \coloneqq \mu_{01} \circ \iota_{01} = \mu_{02} \circ \iota_{02} : \mathbf{n_0} \to \mathbf{n}$ is well-defined. Moreover, the complementary monomorphisms ι_{01}^c and ι_{02}^c provide monomorphisms ι_1 and ι_2,*

as in Diagram (A.1). *The images of* $\iota_0, \iota_1, \iota_2$ *are disjoint.*



(A.1)

Proof. The sets $\mathbf{n_{01}}, \mathbf{n_{02}}, \mathbf{n_0}$ have sizes $n_0{+}1, n_{01}{+}1, n_{02}{+}1$ respectively. Then, $n_{012} \coloneqq n_{01} + n_{02} - n_0$ satisfies $n_{012} + 1 \coloneqq (n_{01} + 1) + (n_{02} + 1) - (n_0 + 1)$ and defines the object $\mathbf{n_{012}} = \{0, \ldots, n_{012}\}$ in $\mathbf{\Delta_a}$.

Monomorphisms μ_{01} and μ_{02} can be constructed via the algorithm in Figure 1, which is merely a sequence of concatenations spliced between aligned entries of ι_{01} and ι_{02}. The resulting maps are indeed morphisms, as they are guaranteed to be order-preserving. □

Example A.3. Consider $n_0 = 1$ and $n_{01} = 5$ and $n_{02} = 4$. Then $n_{012} = 8$. Let $\iota_{01} : \mathbf{1} \mapsto \mathbf{5}$ be the monomorphism that is written as the sequence $[1, 4]$. Let $\iota_{02} : \mathbf{1} \mapsto \mathbf{5}$ be the monomorphism that is written as the sequence $[1, 3]$. Visually, the merged indexing means

$$\begin{aligned}
\iota_{01} : \{0,1\} &\mapsto \{0, \quad \underline{1}, 2, 3, \quad \underline{4}, 5 \quad \} \\
\iota_{02} : \{0,1\} &\mapsto \{ \quad 0, \underline{1}, \qquad 2, \underline{3}, \quad 4\} \\
&\text{yields} \\
\mu_{01} : \{0,1,2,3,4,5\} &\mapsto \{\underline{0}, 1, \underline{2}, \underline{3}, \underline{4}, 5, \underline{6}, \underline{7}, 8\} \\
\mu_{02} : \{0,1,2,3,4\} &\mapsto \{0, \underline{1}, \underline{2}, 3, 4, \underline{5}, \underline{6}, 7, \underline{8}\} \\
&\text{so} \\
\iota_0 : \{0,1\} &\mapsto \{0, 1, \underline{2}, 3, 4, 5, \underline{6}, 7, 8\} \\
\iota_1 : \{0,1,2,3\} &\mapsto \{\underline{0}, 1, 2, \underline{3}, \underline{4}, 5, 6, \underline{7}, 8\} \\
\iota_2 : \{0,1,2\} &\mapsto \{0, \underline{1}, 2, 3, 4, \underline{5}, 6, 7, \underline{8}\}
\end{aligned}$$

For abbreviation and programming, the constructed monomorphisms can be written as lists.

$$\begin{aligned}
\mu_{01} &= [0, 2, 3, 4, 6, 7] \\
\mu_{02} &= [1, 2, 5, 6, 8] \\
\iota_0 &= [2, 6] \\
\iota_1 &= [0, 3, 4, 7] \\
\iota_2 &= [1, 5, 8]
\end{aligned}$$

```
def merge_idx(n01, n02, iota01, iota02):
    """
    Parameters:
        n01, a non-negative integer
        n02, a non-negative integer
        iota01, an increasing list within [0,..,n01]
        iota02, an increasing list within [0,..,n02]
        (iota01 and iota02 must be the same length)
    Returns:
        mu01, an increasing list of n01+1 integers
        mu02, an increasing list of n02+1 integers
        iota0, an increasing list, same length as iota01,iota02
    """
    n0 = len(iota01) - 1
    n012 = n01 + n02 - n0
    mu01 = []
    mu02 = []
    i0 = i01 = i02 = i012 = 0
    while i0 <= n0:
        while i01 < iota01[i0]:
            mu01.append(i012)
            i012 += 1
            i01 += 1
        while i02 < iota02[i0]:
            mu02.append(i012)
            i012 += 1
            i02 += 1
        # now, both i01 and i02 correspond to i0
        mu01.append(i012)
        mu02.append(i012)
        i0 += 1
        i01 += 1
        i02 += 1
        i012 += 1
    # mutual terms are extinguished. concatenate.
    while i01 <= n01:
        mu01.append(i012)
        i012 += 1
        i01 += 1
    while i02 <= n02:
        mu02.append(i012)
        i012 += 1
        i02 += 1

    iota0 = [ mu01[i] for i in iota01 ]
    #     = [ mu02[i] for i in iota02 ]
    return ( mu01, mu02, iota0 )
```

FIGURE 1. Merged indexing algorithm.

A(b). **Simplicial sets.** For any category $\mathbf{C}$, the "simplicial category over $\mathbf{C}$" is $\mathbf{sC}$. An object in $\mathbf{sC}$ is a contravariant functor $X : \mathbf{\Delta} \to \mathbf{C}$. That is, an object in $\mathbf{sC}$ is an assignment of:

- for each object $\mathbf{n}$ in $\mathbf{\Delta}$, an object X_n in $\mathbf{C}$;
- for each morphism (order-preserving function) $\mu : \mathbf{n}' \to \mathbf{n}$ in $\mathbf{\Delta}$, a morphism $X(\mu) : X_n \to X_{n'}$ in $\mathbf{C}$.

The augmented simplicial category, $\mathbf{asC}$, allows a terminal object in $\mathbf{C}$ to correspond to the initial object $-\mathbf{1} \in \mathbf{\Delta_a}$. That is, the trivial map $-\mathbf{1} \to \mathbf{n}$ yields a corresponding map $X_n \to X_{-1}$, if the category $\mathbf{C}$ happens to admit a terminal object.

The morphisms $X \to Y$ in $\mathbf{sC}$ or $\mathbf{asC}$ are the natural transformations as in (A.2).

$$\begin{array}{ccc} X & & X_{n'} \xleftarrow{X(\mu)} X_n \\ \downarrow & & \downarrow \quad\quad \downarrow \\ Y & & Y_{n'} \xleftarrow{Y(\mu)} Y_n \end{array} \qquad \mathbf{n}' \xrightarrow{\mu} \mathbf{n} \tag{A.2}$$

The most important case is $\mathbf{sSet}$, the category of simplicial sets, which is augmented to $\mathbf{asSet}$. The following lemma shows that augmented simplicial sets are given by face and degeneracy maps.

Lemma A.4. *Any object in* $\mathbf{asSet}$ *is a set X (called an* augmented simplicial set*) graded by $-1, 0, 1, 2, \ldots$ and equipped with morphisms $d_i : X_n \to X_{n-1}$ and $s_i : X_n \to X_{n+1}$ for $0 \le i \le n$ such that*

1. $d_i \circ d_j = d_{j-1} \circ d_i$, *if* $i < j$;
2. $d_i \circ s_j = s_{j-1} \circ d_i$, *if* $i < j$;
3. $d_j \circ s_j = d_{j+1} \circ s_j = \mathrm{id}$;
4. $d_i \circ s_j = s_j \circ d_{i-1}$, *if* $i > j+1$; *and*
5. $s_i \circ s_j = s_{j+1} \circ s_i$, *if* $i \le j$.

Proof. The objects are apparent. As for morphisms, each co-face $d^i : \mathbf{n} - \mathbf{1} \to \mathbf{n}$ and co-degeneracy $s^i : \mathbf{n} + \mathbf{1} \to \mathbf{n}$ morphism in $\mathbf{\Delta_a}$ must correspond to face $d_i : X_n \to X_{n-1}$ and boundary $s_i : X_n \to X_{n+1}$ morphisms in X. Because the co-face and co-degeneracy morphisms generate all non-identity morphisms in $\mathbf{\Delta_a}$, it is sufficient to specify the face and degeneracy maps. □

Corollary A.5 (Simplicial maps)**.** *The morphisms of* $\mathbf{sSet}$ *or* $\mathbf{asSet}$ *(called* simplicial maps*) from* (A.2) *are set functions $f : X \to Y$ such that $d_i \circ f = f \circ d_i$ and $s_i \circ f = f \circ s_i$.*

A particularly important example of a simplicial set is Δ^n, the n-simplex. (See 3(a).)

Definition A.6 (Simplex)**.** The standard n-simplex Δ^n is the simplicial set generated (via face and degeneracy maps) by the ordered set $\mathbf{n} = \{0, \ldots, n\}$ in the simplex category $\mathbf{\Delta}$.

By the Yoneda Lemma, a simplicial set X is characterized by the simplicial maps $\Delta^n \to X$; that is, a simplicial set is characterized by its simplices.

Definition A.7 (Horn)**.** The kth *horn* Λ^n_k of the n-simplex Δ^n is the simplicial subset generated by the union of all the faces of Δ^n except the kth face.

By Lemma A.4 and the Yoneda Lemma, if X is a simplicial set, then a horn in X is a collection of n $(n-1)$-simplices $f_0, \ldots, f_{k-1}, f_{k+1}, \ldots, f_n$ such that $d_i f_j = d_{j-1} f_i$ for $i < j$.

A simplicial map $f : X \to Y$ is called a *cofibration* iff it is a monomorphism. A simplicial map $f : X \to Y$ is called a *fibration* iff for any cofibration $i : \Lambda^n_k \hookrightarrow \Delta^n$, the commutative diagram (A.3) can be completed.

$$\begin{array}{ccc} \Lambda^n_k & \longrightarrow & X \\ {\scriptstyle i}\downarrow & \nearrow & \downarrow{\scriptstyle f} \\ \Delta^n & \longrightarrow & Y \end{array} \tag{A.3}$$

Weak-equivalences are defined to be compatible with fibrations and cofibrations according to [18]. See also [9]. These definitions of cofibration, fibration, and weak equivalence make **sSet** into a (closed) model category.

A simplicial set X is called *fibrant* or to satisfy the *Kan extension condition* if $f : X \to \{*\}$ is a fibration; that is, a simplicial set satisfies the Kan condition if and only if each horn Λ^n_k in X can be extended to a simplex Δ^n in X. Let $\mathbf{sSet}_f$ denote the subcategory of fibrant simplicial sets. Then there is a homotopy category $\Pi_n(\mathbf{sSet}_f)$, and any $X \in \mathbf{sSet}_f$ admits pointed homotopy groups $\pi_n(X, x)$ that characterize the weak equivalence. Moreover, the simplicial homotopy groups of $X \in \mathbf{sSet}_f$ are isomorphic to the continuous homotopy groups of its topological realization, $|X|$, as discussed in [18, §3] and [9, Chap I.2]. See also [10] and [12] for historical explanations that minimize categorical language.

A(c). **Data complexes.** Let **DataCplx** denote the category of data complexes. An object in **DataCplx** is a pair of augmented simplicial sets $(\mathcal{X}, \mathcal{A})$ with simplicial map $p : \mathcal{X} \to \mathcal{A}$ such that for each $\mathbf{n} \in \Delta_a$, the set $\mathcal{X}_n$ is a set of data tables over attribute lists $\mathcal{A}_n$ from some attribute set A, as in Section 2, with d_i and s_i by marginalization and Dirac-delta intersection, respectively.

A morphism in **DataCplx** is a simplicial map $f : (\mathcal{X}, \mathcal{A}) \to (\mathcal{Y}, \mathcal{B})$ as in (A.4) with some compatibility conditions.

$$\begin{array}{ccccc} \mathcal{X} & & \mathcal{X}_{n'} & \xleftarrow{\mathcal{X}(\mu)} & \mathcal{X}_n \\ {\scriptstyle f}\downarrow & & \downarrow & & \downarrow \\ \mathcal{Y} & & \mathcal{Y}_{n'} & \xleftarrow{\mathcal{Y}(\mu)} & \mathcal{Y}_n \end{array} \qquad \begin{array}{c} \\ \\ \mathbf{n}' \xrightarrow{\mu} \mathbf{n} \end{array} \tag{A.4}$$

The vertical maps are tuples $(\varphi_n, \{\psi_a\}_{a\in A}, f_n)$ satisfying the following compatibility conditions.

(1) $\varphi_n : \mathcal{A}_n \to \mathcal{B}_n$ is a level of a simplicial map $\varphi : \mathcal{A} \to \mathcal{B}$ on sets of attribute lists.
(2) $f_n : \mathcal{X}_n \to \mathcal{Y}_n$ is a level of a simplicial map $f : \mathcal{X} \to \mathcal{Y}$ on sets of measures, with $\varphi_n = p \circ f_n$.
(3) $\psi_a : \mathbb{V}([a]) \to \mathbb{V}([b])$ is a continuous function on metric spaces, where $[b] = \varphi_0([a]) \in \mathcal{B}_0$. This induces $\psi_T : \mathbb{V}(T) \to \mathbb{V}(\varphi_n(T))$ for all $T \in \mathcal{A}_n$.
(4) If $(T, \tau) \in \mathcal{X}_n$ with $\varphi_n(T) = S$ and $f_n(T, \tau) = (S, \sigma)$, then $\tau \circ \psi_T^{-1} = \sigma$ as measures. That is,

$$f_n : (T, \tau) \mapsto (\varphi_n(T), \tau \circ \psi_T^{-1}). \tag{A.5}$$

These conditions guarantee simply that the attribute lists T, the value spaces $\mathbb{V}(T)$, and the measure spaces $\mathbf{M}(T)$ remain compatible. As with **sSet**, in (A.4), the map $\mu : \mathbf{n}' \to \mathbf{n}$ can be taken to be $d^i : \mathbf{n} - \mathbf{1} \to \mathbf{n}$ or $s^i : \mathbf{n} + \mathbf{1} \to \mathbf{n}$ so that the diagram describes naturality with respect to face and degeneracy maps on $\mathcal{X}$ and $\mathcal{Y}$. These conditions are sensible for $n \geq 0$, so they apply to the trivial data table $([], M)$.

For each real number $M \geq 0$, there is a *singleton* data complex with $A = \{*\}$, $\mathbb{V}(*) = \{*\}$. For each $n \geq -1$, there is a single attribute list $[*, \ldots, *]$ with a singleton value space $\{*\}^n$ and one measure, M. For brevity, we refer to this singleton data complex as M.

Slightly more generally, there is a *terminal* data complex with $A = \{*\}$, $\mathbb{V}(*) = \{*\}$. For each $n \geq -1$, there is a single attribute list $[*, \ldots, *]$ with a singleton value space $\{*\}^n$ and measures M for each $M \geq 0$. The terminal data complex is the union of all the singleton data complexes. For brevity, we refer to the terminal data complex as $\mathbb{R}_{\geq 0}$.

Every data complex $\mathcal{X}$ admits a morphism to the terminal data complex $\mathbb{R}_{\geq 0}$. This terminal morphism f maps each data table $(T, \tau) \in \mathcal{X}$ to the singleton mass $([*, \ldots, *], \downarrow_{[]} \tau) \in \mathbb{R}_{\geq 0}$. If all data tables in $\mathcal{X}$ share the same mass (say, $M = 1$), then the image of the terminal morphism goes to some $M \subset \mathbb{R}_{\geq 0}$.

A morphism in **DataCplx** is called a cofibration iff it is a monomorphism. A morphism in **DataCplx** is called a fibration iff for any cofibration of from a well-aligned pair to a join $i : \langle \tau_{01}, \tau_{02} \rangle_{T_0} \to \langle \tau_{012} \rangle$, the commutative diagram (A.6) can be completed.

$$\begin{array}{ccc} \langle \tau_{01}, \tau_{02} \rangle_{T_0} & \longrightarrow & \mathcal{X} \\ \Big\downarrow{\scriptstyle i} & \nearrow & \Big\downarrow{\scriptstyle f} \\ \langle \tau_{012} \rangle & \longrightarrow & \mathcal{Y} \end{array} \tag{A.6}$$

A data complex $\mathcal{X}$ is called *fibrant* if the terminal morphism $\mathcal{X} \to \mathbb{R}_{\geq 0}$ is a fibration. By Theorem 3.11, if $\mathcal{X}$ is a fibrant data complex, then $\mathcal{X}$ is a fibrant simplicial set. Thus, the category **DataCplx** is a (closed) model category, and the fibrant subcategory $\mathbf{DataCplx}_f$ inherits a well-defined homotopy category $\Pi_n(\mathbf{DataCpl}_f)$ from $\mathbf{sSet}_f$, and any $\mathcal{X} \in \mathbf{DataCplx}_f$ admits pointed homotopy groups $\pi_n(\mathcal{X}, \tau_0)$

that characterize the weak equivalence. Moreover, the homotopy groups are isomorphic to the continuous homotopy groups of the topological realization of the underlying simplicial set.

Acknowledgments

We are very grateful to John Paschewitz and Tony Falcone for project guidance and technical direction, and to Greg Friedman, Justin Curry, Jose Perea, and Jonathan Mattingly for helpful discussions at various stages of the theoretical development.

References

[1] Samson Abramsky, *Relational databases and Bell's theorem*, In search of elegance in the theory and practice of computation, Lecture Notes in Comput. Sci., vol. 8000, Springer, Heidelberg, 2013, pp. 13–35, DOI 10.1007/978-3-642-41660-6_2. MR3124039

[2] Samson Abramsky, Rui Soares Barbosa, Kohei Kishida, Raymond Lal, and Shane Mansfield, *Contextuality, cohomology and paradox*, 24th EACSL Annual Conference on Computer Science Logic, LIPIcs. Leibniz Int. Proc. Inform., vol. 41, Schloss Dagstuhl. Leibniz-Zent. Inform., Wadern, 2015, pp. 211–228. MR3441764

[3] Torgny Lindvall, *Lectures on the coupling method*, Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics, John Wiley & Sons, Inc., New York, 1992. A Wiley-Interscience Publication. MR1180522

[4] Herbert Edelsbrunner and John L. Harer, *Computational topology*, American Mathematical Society, Providence, RI, 2010. An introduction. MR2572029

[5] P. J. Ehlers and T. Porter, *Joins for (augmented) simplicial sets*, J. Pure Appl. Algebra **145** (2000), no. 1, 37–44, DOI 10.1016/S0022-4049(98)00065-6. MR1732286

[6] Brendan Fong and David I. Spivak, *An invitation to applied category theory: Seven sketches in compositionality*, Cambridge University Press, Cambridge, 2019. MR3966447

[7] Greg Friedman, *Survey article: An elementary illustrated introduction to simplicial sets*, Rocky Mountain J. Math. **42** (2012), no. 2, 353–423, DOI 10.1216/RMJ-2012-42-2-353. MR2915498

[8] Eli Glasner, *Ergodic theory via joinings*, Mathematical Surveys and Monographs, vol. 101, American Mathematical Society, Providence, RI, 2003. MR1958753

[9] Paul G. Goerss and John F. Jardine, *Simplicial homotopy theory*, Modern Birkhäuser Classics, Birkhäuser Verlag, Basel, 2009. Reprint of the 1999 edition [MR1711612]. MR2840650

[10] Daniel M. Kan, *A combinatorial definition of homotopy groups*, Ann. of Math. (2) **67** (1958), 282–312, DOI 10.2307/1970006. MR111032

[11] Albert T. Lundell, *Obstruction theory of principal fibre bundles*, Trans. Amer. Math. Soc. **97** (1960), 161–192, DOI 10.2307/1993368. MR130694

[12] J. Peter May, *Simplicial objects in algebraic topology*, Chicago Lectures in Mathematics, University of Chicago Press, Chicago, IL, 1992. Reprint of the 1967 original. MR1206474

[13] Jason Morton, *Contextuality from missing and versioned data*, 1–21, `arXiv:1708.03264` 2017.

[14] Felix Naumann, Alexander Bilke, Jens Bleiholder, and Melanie Weis, *Data fusion in three steps: resolving inconsistencies at schema-, tuple-, and value-level*, Bulletin of the IEEE Computer Society Technical Committee on Data Engineering **29** (2006), no. 2, 21–31.

[15] Paul Olum, *Obstructions to extensions and homotopies*, Ann. of Math. (2) **52** (1950), 1–50, DOI 10.2307/1969510. MR36507

[16] Nina Otter, *Magnitude meets persistence. homology theories for filtered simplicial sets*, 1–21, `arXiv:1807.01540` 2018.

[17] Gabriel Peyré and Marco Cuturi, *Computational optimal transport*, `arXiv:1803.00567` 2018.

[18] Daniel G. Quillen, *Homotopical algebra*, Lecture Notes in Mathematics, vol. 43, Springer Berlin Heidelberg, Berlin, Heidelberg, 1967, doi:10.1007/BFb0097438.

[19] Egbert Rijke, *The join construction*, (2017), `arXiv:1701.07538`.

[20] Patrick Schultz, David I. Spivak, Christina Vasilakopoulou, and Ryan Wisnesky, *Algebraic databases*, Theory Appl. Categ. **32** (2017), Paper No. 16, 547–619. MR3641249

[21] Patrick Schultz and Ryan Wisnesky, *Algebraic data integration*, J. Funct. Programming **27** (2017), e24, 51, DOI 10.1017/S0956796817000168. MR3720789
[22] N. E. Steenrod, *Homology with local coefficients*, Ann. of Math. (2) **44** (1943), 610–627, DOI 10.2307/1969099. MR9114
[23] Norman Steenrod, *The Topology of Fibre Bundles*, Princeton Mathematical Series, vol. 14, Princeton University Press, Princeton, N. J., 1951. MR0039258
[24] Phillip B. Thurber, *Semi-localization of a one pointed Kan complex*, Pacific J. Math. **178** (1997), no. 1, 147–184, DOI 10.2140/pjm.1997.178.147. MR1447409

Department of Mathematics, Statistics, and Computer Science, University of Wisconsin-Stout, Menomonie, Wisconsin 54751; and Geometric Data Analytics, Inc., Durham, North Carolina 27707
Email address: smithabr@uwstout.edu

Department of Mathematics, Duke University, Durham, North Carolina 27708; and Geometric Data Analytics, Inc., Durham, North Carolina 27707
Email address: bendich@math.duke.edu

Department of Mathematics, Duke University, Durham, North Carolina 27707; and Geometric Data Analytics, Inc., Durham, North Carolina 27707
Email address: harer@math.duke.edu